

Causal Statistical Decision Theory|What are interventions?

David Johnston

August 29, 2022

Contents

1	Introduction	5
1.1	Making decisions with data	5
1.2	Assumptions in predictive and decision-making algorithms	6
1.3	Exploring alternative foundations	9
1.3.1	Other theories	11
1.4	Causally compatible variables	12
2	Technical Prerequisites	13
2.1	Conventions	14
2.2	Probability Theory	14
2.2.1	Standard Probability Theory	14
2.3	String Diagrams	19
2.3.1	Elements of string diagrams	19
2.3.2	Special maps	21
2.3.3	Commutative comonoid axioms	22
2.3.4	Manipulating String Diagrams	22
2.4	Probability Sets	24
2.4.1	Almost sure equality	26
2.4.2	Extended conditional independence	27
2.4.3	Examples	29
2.4.4	Maximal probability sets and valid conditionals	31
2.4.5	Existence of conditional probabilities	35
3	Models with choices and consequences	39
3.1	What is the point of causal inference?	40
3.1.1	Modelling decision problems	42
3.1.2	Formal definitions	43
3.2	Theories of decision making	44
3.2.1	von Neumann-Morgenstern utility	45
3.2.2	Savage decision theory	45
3.2.3	Jeffrey's decision theory	47
3.2.4	Causal decision theory	49
3.2.5	Statistical decision theory	49
3.3	Variables	57

3.3.1	Variables and measurement procedures	57
3.3.2	Measurement procedures	58
3.3.3	Observable variables	60
3.3.4	Model variables	60
3.3.5	Variable sequences and partial order	61
3.3.6	Decision procedures	61
3.4	Conclusion	62
3.5	Appendix: axiomatisation of decision theories	62
3.5.1	Savage axioms	62
3.5.2	Bolker axioms	64
4	Repeatable decision problems	65
4.0.1	Chapter outline	68
4.0.2	Key terminology	69
4.1	Previous work on causal symmetries	69
4.2	Response functions	71
4.3	Symmetries	72
4.3.1	Representation of IO contractible models	79
4.3.2	Data-independent inputs	91
4.4	Discussion	95
4.4.1	Simple symmetries vs strategic behaviour	95
4.4.2	Implications of IO contractibility	97
4.5	Data-dependent inputs	98
4.5.1	Combs	99
4.5.2	Representation of data-dependent inputs	102
4.6	IO contractible Markov kernels	103
4.6.1	Representation of IO contractible kernels	103
4.7	Discussion	111
5	Other causal modelling frameworks	113
5.1	What is a Causal Bayesian Network?	113
5.1.1	Definition of a Causal Bayesian Network	113
5.1.2	Unrolling a causal Bayesian network	116
5.1.3	Uncertainty in an unrolled causal Bayesian network	118
5.1.4	Probabilistic Graphical Models	119
5.1.5	Hidden confounders and precedents	121
5.2	What is a Potential Outcomes model?	129
5.3	Individual-level response functions	131
5.3.1	References to individual-level IO contractibility	132
5.3.2	Unique identifiers	133
5.3.3	Identification	134
5.3.4	Other examples	143
5.4	Conclusion	146
6	Discussion	147

Chapter 1

Introduction

Introduction stands alone

1.1 Making decisions with data

Beginning in the 1930s, a number of associations between cigarette smoking and lung cancer were established: on a population level, lung cancer rates rose rapidly alongside the prevalence of cigarette smoking. Lung cancer patients were far more likely to have a smoking history than demographically similar individuals without cancer and smokers were around 40 times as likely as demographically similar non-smokers to go on to develop lung cancer. In laboratory experiments, cells which were introduced to tobacco smoke developed *ciliastasis*, and mice exposed to cigarette smoke tars developed tumors (Proctor, 2012). Nevertheless, until the late 1950s, substantial controversy persisted over the question of whether the available data was sufficient to establish that smoking cigarettes *caused* lung cancer. Cigarette manufacturers famously argued against any possible connection (Oreskes and Conway, 2011) and Roland Fisher in particular argued that the available data was not enough to establish that smoking actually caused lung cancer (Fisher, 1958). Today, it is widely accepted that cigarettes do cause lung cancer, along with other serious conditions such as vascular disease and chronic respiratory disease (World Health Organisation, 2018; Wiblin, 2016).

The question of a causal link between smoking and cancer is a very important one to many different people. Individuals who enjoy smoking (or think they might) may wish to avoid smoking if cigarettes pose a severe health risk, so they are interested in knowing whether or not it is so. Additionally, some may desire reassurance that their habit is not too risky, whether or not this is true. Potential and actual investors in cigarette manufacturers may see health concerns as a barrier to adoption, and also may personally want to avoid supporting products that harm many people. Like smokers, such people might have some interest in knowing the truth of this question, and a separate interest in hearing that cigarettes are not too risky, whether or not this is true. Governments

and organisations with a responsibility for public health may see themselves as having responsibility to discourage smoking as much as possible if smoking is severely detrimental to health. The costs and benefits of poor decisions about smoking are large: 8 million annual deaths are attributed to cigarette-caused cancer and vascular disease in 2018 (World Health Organisation, 2018) while global cigarette sales were estimated at US\$711 billion in 2020 (Statista, 2020) (a figure which might be substantially larger if cigarettes were not widely believed to be harmful).

The question of whether or not cigarette smoking causes cancer illustrates two key facts about causal questions: First, having the right answers to causal questions can underpin decisions of tremendous importance to large numbers of people. Second, confusion over causal questions can persist even when a great deal of data and facts relevant to the question are agreed upon.

Causal conclusions are often justified on the basis of ad-hoc reasoning. For an arbitrarily chosen example, Krittanawong et al. (2020) state:

[...] the potential benefit of increased chocolate consumption, reducing coronary artery disease (CAD) risk is not known. We aimed to explore the association between chocolate consumption and CAD.

It is not clear whether Krittanawong et. al. mean that a negative association between chocolate consumption and CAD implies that deciding to eat more chocolate is likely to reduce coronary artery disease (which is suggested by the word “benefit”), or that an association may be relevant to the question and the reader should draw their own conclusions. Whether the implication is being suggested by Krittanawong et. al. or merely imputed by naïve readers, it is being drawn on an ad-hoc basis – no argument for the implication can be found in this paper. As Pearl (2009) has forcefully argued, additional assumptions are always required to answer causal questions from associational facts, and stating these assumptions explicitly allows those assumptions to be productively scrutinised.

1.2 Assumptions in predictive and decision-making algorithms

We can contrast the type of problem outlined above – where one is called on to make choices on the basis of given data – with prediction problems, where one is simply asked to provide a prediction of what is likely to happen in the future. We can also abuse terminology somewhat to include classification tasks under “prediction” – the relevant similarity being the fact that there is a single true answer (either a class or a future event) unknown to the predictor (or classifier), and the success of the prediction (or class assignment) is judged on the basis of its fit to the true answer.

Data driven prediction problems and data driven decision making problems have a lot in common. The outcomes some people are interested in predicting

1.2. ASSUMPTIONS IN PREDICTIVE AND DECISION-MAKING ALGORITHMS7

are often outcomes other people want to influence. A forecaster might want to predict the winner of the next election, while a party strategist is interested in maximising their party’s chance of victory. A product manager may be simultaneously interested in accurately inferring the sentiment expressed in reviews of their product, and in making product changes that increase the frequency that this sentiment is positive. Furthermore, data relevant to prediction is often relevant to decision making and vice-versa. Political parties often reason that electorates in which their predicted chance of victory is very low are not worth investing campaign resources in, and if a forecaster learns of evidence that one party had adopted a particularly effective election strategy they might want to revisit their prediction of the eventual election winner. The overlap is not perfect: comprehensive electorate level polls are probably more useful to the forecaster while small-scale controlled experiments are probably more useful to the strategist.

A key difference between prediction and influence problems is the “multiplicity of futures” that each problem asks us to consider. A forecaster wants to identify – loosely speaking – the single most likely outcome, while a strategist must consider multiple options and identify the likely outcomes associated with each of these. As a consequence of this difference, the forecaster receives more complete feedback about the quality of their forecast than the strategist. Unlike the forecaster, all but one of the options that the strategist considers are never realised, and the world never offers feedback on these alternative options.

This difference suggests that it might be easier to assess the reliability of a predictive algorithm than the reliability of a decision-making algorithm, and this is borne out in practice. Validating a predictive algorithm using data split into training and holdout sets is a ubiquitous in machine learning. For many data generating processes, appropriately conducted validation is widely considered to be a reliable indicator of an algorithm’s performance for sufficiently similar data generating processes. In contrast, the most well-known condition that is widely accepted to yield reliable decision making algorithms is that the data used to draw inferences comes from a well-conducted controlled experiment. Data that satisfies this is much rarer than data that standard machine learning validation approaches can be applied to. There are approaches to causal inference that don’t depend on experimental data, but they depend on other assumptions which are similarly applicable to a limited fraction of datasets. Alternatives to controlled experiments often come with the additional headache of being difficult to assess for a given dataset.

Some of the most far-reaching recent development in algorithmic decision making have involved only the elementary theory of randomised experiments. Operational advances that enable controlled experiments to be conducted at large scales have driven substantial changes in the operations of many online businesses (Kohavi and Thomke, 2017), and Abhijit Banerjee and Esther Duflo were recently awarded a Nobel prize in part for their pioneering role in the use of large numbers of randomised controlled trials (RCTs) to assess the effectiveness of different kinds of development interventions (Zhang, 2014). Some fields of science have also been significantly affected by “negative progress” in the science

of assessing experimental results. For example, in psychology, strong evidence has emerged that experimental findings from this field provide weaker evidence to a reader of the findings about the consequences of the reader’s actions than many had believed Open Science Collaboration (2015); Stroebe (2019). In a similar time frame, standards for what constitutes a “well-conducted” experiment have risen across many fields (Nosek et al., 2018; Liberati et al., 2009).

An individual who wants to use data to make better decisions can consider running a controlled experiment of their own. This may not be possible, and even if it is, there may be large amounts of apparently relevant data available that seems wasteful to ignore on the basis of its non-experimental provenance. This individual might therefore be motivated to make some additional assumptions which allow them to draw conclusions about how to act from non-experimental data.

Some examples of assumptions this person could consider are (a * indicates that causal conclusions may or may not follow, depending on the actual data at hand, and italics indicate technical terms that will be explained in more detail further into this thesis):

- **Conditional ignorability:** There is an input variable independent of the *potential outcomes* conditional on some covariates (Imbens and Rubin, 2015, Chap. 12), (Angrist and Pischke, 2014, Chaps 2, 3, 5)
- * **Potential outcomes proxy:** There is a variable closely correlated with the *potential outcomes* for each observation (Imbens and Rubin, 2015, Chap. 21)
- **Regression discontinuity:** *Potential outcomes* are continuous about some covariate cutoff, above which one potential outcome is always observed and below which the other is always observed (Hahn et al., 2001)
- * **Known causal structure:** The set of observed and unobserved variables have a known *causal structure* (Shpitser and Pearl, 2008; Richardson et al., 2017)
- * **Faithfulness:** The set of observed variables is *causally sufficient* and the *causal structure* is *faithful* to the conditional independence structure of the observed variables (Spirtes et al., 2000, Chap. 5)

These assumptions all invoke either “potential outcomes” or “causal structure”. Potential outcomes are, by definition, the union of a set of observed variables and a set of hypothetical claims. If one is inclined to accept that causal structures are ultimately underwritten by certain experimental procedures, then causal structures may be observed under some conditions (for example, if an experimental procedure is given for each *do()*-intervention, one can apply the method of (Eberhardt, 2008)) However, causal structures are often held to represent relationships that cannot be reasonably probed by an experiment ((Pearl, 2009, Chap. 11), Pearl (2018)), and in our view the general question of

whether causal structures are observable or a mixture of observable properties and hypothetical claims is unresolved.

Decision making must entertain some hypothetical statements. As has been pointed out above, a decision maker must consider multiple options and their likely consequences, and select from among the options. The proposition “the consequences of α are better than the consequences of α' ” is a hypothetical one: the decision maker ultimately chooses only one of α or α' , and usually cannot ever verify that the chosen option really is better than the alternative. However, the hypotheticals that the assumptions above ask us to entertain do not on the face of it seem necessary to compare different options like this. The hypothetical part of potential outcomes doesn’t refer to possible consequences of choices but to hypotheticals that have (in some sense) “already happened” by the time the data is reviewed. An identifiable causal structure may have many implications besides those that are necessary to make a choice in the decision problem at hand.

Assumptions with mixtures of hypothetical and real implications are difficult to evaluate. On the one hand, the fact that they have real implications means that one cannot simply accept anything, while on the other hand the fact that some implications are hypothetical means that it can be difficult to form a clear picture of how the world actually differs depending on whether the assumption is true or false.

In brief:

- Decision making algorithms require stronger assumptions than predictive algorithms so
- These assumptions are often particularly hard to evaluate

1.3 Exploring alternative foundations

The development of decision making algorithms is constrained by the assumptions we can make. Such problems are often addressed using theories of causal inference. Theories of causal inference allow us to state assumptions precisely, to understand them in various different ways so that we can better apply our informal means of evaluating them and, hopefully, to make sound choices about whether to accept an assumption or not. A number of such theories exist, and we can distinguish three traditions with wide adoption in different fields: causal Bayesian networks, potential outcomes and structural equation models. These are described in more detail in Chapter 5.

This thesis explores an alternative theory of algorithmic decision making. Our starting point is the observation in Section 1.2 above that decision making problems call on a decision maker to compare a number of different options on the basis of their consequences. The conclusion is that, to construct a decision making model, where in classical statistics we might consider a probability distribution \mathbb{P} defined on a sample space (Ω, \mathcal{F}) , we instead consider a function from the set

of options C to probability distributions on a sample space: $\mathbb{P} : C \rightarrow \Omega$. This extension of the basic statistical model is what underpins this entire thesis.

We call our approach a “decision theoretic approach”, and a number of authors have previously explored a similar approach to causal inference (Heckerman and Shachter, 1995; Dawid, 2000, 2002, 2012, 2020). One might ask us all: if we already have three ways to theorise about this problem, why do we need a fourth? To a large extent, for questions like these, the proof is in the pudding – do the alternative foundations offer insights that are hard to see from other points of view? However, we think that even before learning what alternatives can deliver, there are reasons to believe they are worth exploring.

A broad reason for paying special attention to theoretical foundations in causal inference is outlined in Section 1.2: causal inference typically requires assumptions that are strong and particularly hard to evaluate. Alternative theoretical foundations can yield alternative perspectives on old assumptions and suggest new assumptions that accomplish desired goals. For these reasons, causal inference is a field that merits particular attention to foundations. In fact, the reason we have several different causal inference traditions is because different researchers have, at different times, attacked the problem of how to formulate causal models in different ways. This project has been fruitful – it has facilitated the entire field of research into causal inference. In addition to the handful of identification results suggested above, some other insights enabled by these projects are:

- With potential outcomes, it is possible to offer a precise formal statement of what it is that controlled experiments should achieve – *strong ignorability* (Rubin, 2005)
- With potential outcomes we can precisely express the notion of “the effect of a choice on the people whose behaviour changed on account of that choice”, formalised as the *local average treatment effect* (Imbens and Angrist, 1994)
- There is a close correspondence between the intuitive idea of directed causal relationships between variables and the factorisation of joint probability distributions, formalised as the *causal Markov condition* (Pearl (2009, Chap. 1), Wright (1934))
 - Particular credit is due Pearl for pointing out how common it is for researchers to smuggle these “intuitive causal ideas” into discussions with no accompanying theory of causation - see for example Pearl (2009, pg. 96)
- The field of conditional-independence based *causal discovery* Spirtes et al. (2000)
- The field of causal discovery based on the *independence of cause and mechanism* Schölkopf (2022)
- Causal identification in semi-Markovian models Shpitser and Pearl (2008)

Here, items marked with a dot are (in our view) difficult to formulate using causal graphical models alone, while items marked with a dash are (again, in our view) difficult to understand from the potential outcomes perspective.

We have some specific reasons for wanting to go beyond these two frameworks. For the potential outcomes framework, the reason is quite simple: in order to make a decision, we must consider a map from some set of options C to distributions over consequences, and the potential outcomes framework does not provide such a function at the base level. Once we incorporate such a function, it is not clear that there is any longer a need for potential outcomes at the axiomatic level (though it may still be possible to define variables in particular problems that play a similar role).

The situation with causal graphical models – that typically do have a notion of intervention – requires some more explanation. Causal graphical models are characterised by sets of variables with directed relationships between them, and these directed relationships are underwritten by the idea of *interventions*. Our suspicion is that this kind of model describes a special case of the kind of model we investigate – which, recall, is defined by a function from a set of options C to a set of probability distributions over a fixed sample space – but is not appropriate to describe the general case. The issue centres on variables and interventions.

First, it is important to point out that while interventions are suggestively named, they are not defined as “things an experimenter actually goes and does”. Any identification of actions that can be taken with interventions is up to the analyst’s discretion for a particular problem.

the notion of intervention is incompatible with many sets of variables

1.3.1 Other theories of algorithmic decision making

Causal inference theories aren’t the only theories that address decision making algorithms. These questions are also addressed by the fields of reinforcement learning, optimal control and statistical decision theory (and, no doubt, others besides). One distinction we can draw between these fields and the field of causal inference is that a key difficulty in causal inference problems is just how to relate consequences of actions to observations. In reinforcement learning, an *environment* is typically assumed that represents the “ground truth” of consequences of actions, and the history of consequences can be used to infer which environment an agent is operating in (Barto, 1998). While optimal control is such a large field it’s inappropriate to make any sweeping generalisations, basic versions of control theory assume a *system model* is available that maps states and inputs to updated states and outputs (Ogata, 1995). Finally, in statistical decision theory the relevant notion of consequences of actions is given by the *state* and the *loss*, which like the environment in reinforcement learning, are basic elements of the problem (Wald, 1950).

In the causal graphical models tradition, the graphical model plays the role of relating observations to consequences, and potential outcomes (arguably) play

this role in the potential outcomes tradition (as is explained in more detail in Chapter 5). In contrast to reinforcement learning and optimal control, causal inference often deals with “one-shot” problems, where the given data is assumed to be fixed. In contrast to statistical decision theory, the space of consequences is not restricted to a scalar loss value, and in fact is typically assumed to be identical to the space of observations.

There is substantial overlap between these different methods for relating observations to consequences. For example, Lattimore (2017, Chap. 4) shows how the environment model in a reinforcement learning problem can be specified using a causal graphical model. In Chapter 4, we will discuss *hypotheses* in decision making models which are very similar to *states* in statistical decision theory, and in Chapter 3 we will show how a decision making model in conjunction with a utility function induces a statistical decision problem.

1.4 Causally compatible variables

Chapter 2

Technical Prerequisites

Our approach to causal inference is (like most other approaches) based on probability theory. Many results and conventions will be familiar to readers, and these are collected in Section 2.2.1.

Less likely to be familiar to readers is the string diagram notation we use to represent probabilistic functions. This is a notation created for reasoning about abstract Markov categories, and is somewhat different to existing graphical languages. The main difference is that in our notation wires represent variables and boxes (which are like nodes in directed acyclic graphs) represent probabilistic functions. Standard directed acyclic graphs annotate nodes with variable names and represent probabilistic functions implicitly. The advantage of explicitly representing probabilistic functions is that we can write equations involving graphics. It is introduced in Section 2.3.

We also extend the theory of probability to a theory of probability sets, which we introduce in Section 2.4. This section goes over some ground already trodden by Section 2.2.1; this structure was chosen so that people familiar with the Section 2.2.1 can skip to Section 2.4 for relevant generalisations to probability sets. Two key ideas introduced here are *uniform conditional probability*, similar but not identical to conditional probability, and *extended conditional independence* as introduced by Constantinou and Dawid (2017), similar but not identical to regular conditional independence.

We finally introduce the assumption of *validity*, which ensures that probability sets constructed by “assembling” collections of uniform conditionals are non-empty.

This is a reference chapter – a reader who is already quite familiar with probability theory may skip to Chapter 3. Where necessary, references back to theorems and definitions in this chapter are given. In Chapter 4, we will introduce one additional probabilistic primitive: *combs*, as we feel that additional context is helpful for understanding them.

2.1 Conventions

One of the unusual conventions in this thesis is the notation of uniform conditional probability. Given a set of probability distributions $\mathbb{P}_C := \{\mathbb{P}_\alpha | \alpha \in C\}$ on a common sample space (Ω, \mathcal{F}) with variables $\mathbf{X} : \Omega \rightarrow X$ and $\mathbf{Y} : \Omega \rightarrow Y$, $\mathbb{P}_C^{\mathbf{Y}|\mathbf{X}}$ represents a Markov kernel $X \rightarrow Y$ that satisfies the definition of the distribution of \mathbf{Y} given \mathbf{X} (Definition 2.2.16) for every $\alpha \in C$, while $\mathbb{P}_\alpha^{\mathbf{Y}|\mathbf{X}}$ is a conditional distribution only for α . There are two unusual feature: firstly, it is more common to write a conditional distribution $\mathbb{P}(\mathbf{Y}|\mathbf{X})$ and secondly, the subscript indicating the “domain of validity” of the conditional probability is unusual.

Because this thesis uses sets of probability measures rather than single probability measures, in general a conditional distribution may be valid only for some subset of the probability measures, and always including a subscript indicating which subset or element for which a conditional distribution is valid avoids any ambiguity about this. Avoiding notation of the form $\mathbb{P}(\mathbf{Y}|\mathbf{X})$ is an aesthetic preference; writing a conditional distribution like this suggests $\mathbb{P}(\mathbf{Y}|\mathbf{X})$ is the result of function composition between \mathbb{P} and some function denoted “ $\mathbf{Y}|\mathbf{X}$ ”. However, conditional probabilities are not given by composition of functions like this.

Name	notation	meaning
Iverson bracket	$\llbracket \cdot \rrbracket$	Function equal to 1 if \cdot is true, false otherwise
Identity function	idf_X	Identity function $X \rightarrow X$
Identity kernel	id_X	Kernel associated with the identity function $X \rightarrow X$

2.2 Probability Theory

2.2.1 Standard Probability Theory

σ -algebras

Definition 2.2.1 (Sigma algebra). Given a set A , a σ -algebra \mathcal{A} is a collection of subsets of A where

- $A \in \mathcal{A}$ and $\emptyset \in \mathcal{A}$
- $B \in \mathcal{A} \implies B^C \in \mathcal{A}$
- \mathcal{A} is closed under countable unions: For any countable collection $\{B_i | i \in \mathbb{N}\}$ of elements of \mathcal{A} , $\cup_{i \in \mathbb{N}} B_i \in \mathcal{A}$

Definition 2.2.2 (Measurable space). A measurable space (A, \mathcal{A}) is a set A along with a σ -algebra \mathcal{A} .

Definition 2.2.3 (Sigma algebra generated by a set). Given a set A and an arbitrary collection of subsets $U \subset \mathcal{P}(A)$, the σ -algebra generated by U , $\sigma(U)$, is the smallest σ -algebra containing U .

Common σ algebras For any A , $\{\emptyset, A\}$ is a σ -algebra. In particular, it is the only sigma algebra for any one element set $\{*\}$.

For countable A , the power set $\mathcal{P}(A)$ is known as the discrete σ -algebra.

Given A and a collection of subsets of $B \subset \mathcal{P}(A)$, $\sigma(B)$ is the smallest σ -algebra containing all the elements of B .

If A is a topological space with open sets T , $\mathcal{B}(\mathbb{R}) := \sigma(T)$ is the *Borel σ -algebra* on A .

If A is a separable, completely metrizable topological space, then $(A, \mathcal{B}(A))$ is a *standard measurable set*. All standard measurable sets are isomorphic to either $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ or $(C, \mathcal{P}(C))$ for denumerable C (Çinlar, 2011, Chap. 1).

Probability measures and Markov kernels

Definition 2.2.4 (Probability measure). Given a measurable space (E, \mathcal{E}) , a map $\mu : \mathcal{E} \rightarrow [0, 1]$ is a *probability measure* if

- $\mu(E) = 1$, $\mu(\emptyset) = 0$
- Given countable collection $\{A_i\} \subset \mathcal{E}$, $\mu(\cup_i A_i) = \sum_i \mu(A_i)$

Nxample 2.2.5 (Set of all probability measures). The set of all probability measures on (E, \mathcal{E}) is written $\Delta(E)$.

Definition 2.2.6 (Markov kernel). Given measurable spaces (E, \mathcal{E}) and (F, \mathcal{F}) , a *Markov kernel* or *stochastic function* is a map $\mathbb{M} : E \times \mathcal{F} \rightarrow [0, 1]$ such that

- The map $\mathbb{M}(A|\cdot) : x \mapsto \mathbb{M}(A|x)$ is \mathcal{E} -measurable for all $A \in \mathcal{F}$
- The map $\mathbb{M}(\cdot|x) : A \mapsto \mathbb{M}(A|x)$ is a probability measure on (F, \mathcal{F}) for all $x \in E$

Nxample 2.2.7 (Signature of a Markov kernel). Given measurable spaces (E, \mathcal{E}) and (F, \mathcal{F}) and $\mathbb{M} : E \times \mathcal{F} \rightarrow [0, 1]$, we write the signature of $\mathbb{M} : E \rightarrow F$, read “ \mathbb{M} maps from E to probability measures on F ”.

Definition 2.2.8 (Deterministic Markov kernel). A *deterministic* Markov kernel $\mathbb{A} : E \rightarrow \Delta(\mathcal{F})$ is a kernel such that $\mathbb{A}_x(B) \in \{0, 1\}$ for all $x \in E$, $B \in \mathcal{F}$.

Common probability measures and Markov kernels

Definition 2.2.9 (Dirac measure). The *Dirac measure* $\delta_x \in \Delta(X)$ is a probability measure such that $\delta_x(A) = \mathbb{I}[x \in A]$

Definition 2.2.10 (Markov kernel associated with a function). Given measurable $f : (X, \mathcal{X}) \rightarrow (Y, \mathcal{Y})$, $\mathbb{F}_f : X \rightarrow Y$ is the Markov kernel given by $x \mapsto \delta_{f(x)}$

Definition 2.2.11 (Markov kernel associated with a probability measure). Given (X, \mathcal{X}) , a one-element measurable space $(\{*\}, \{\{*\}, \emptyset\})$ and a probability measure $\mu \in \Delta(X)$, the associated Markov kernel $\mathbb{Q}_\mu : \{*\} \rightarrow X$ is the unique Markov kernel $* \mapsto \mu$

Lemma 2.2.12 (Products of functional kernels yield function composition). *Given measurable $f : X \rightarrow Y$ and $g : Y \rightarrow Z$, $\mathbb{F}_f \mathbb{F}_g = \mathbb{F}_{g \circ f}$.*

Proof.

$$\begin{aligned} (\mathbb{F}_f \mathbb{F}_g)_x(A) &= \int_X (\mathbb{F}_g)_y(A) d(\mathbb{F}_f)_x(y) \\ &= \int_X \delta_{g(y)}(A) d\delta_{f(x)}(y) \\ &= \delta_{g(f(x))}(A) \\ &= (\mathbb{F}_{g \circ f})_x(A) \end{aligned}$$

□

Variables, conditionals and marginals

Definition 2.2.13 (Variable). Given a measurable space (Ω, \mathcal{F}) and a measurable space of values (X, \mathcal{X}) , an *X-valued variable* is a measurable function $X : (\Omega, \mathcal{F}) \rightarrow (X, \mathcal{X})$.

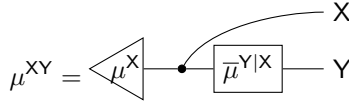
Definition 2.2.14 (Sequence of variables). Given a measurable space (Ω, \mathcal{F}) and two variables $X : (\Omega, \mathcal{F}) \rightarrow (X, \mathcal{X})$, $Y : (\Omega, \mathcal{F}) \rightarrow (Y, \mathcal{Y})$, $(X, Y) : \Omega \rightarrow X \times Y$ is the variable $\omega \mapsto (X(\omega), Y(\omega))$.

Definition 2.2.15 (Marginal distribution). Given a probability space $(\mu, \Omega, \mathcal{F})$ and a variable $X : \Omega \rightarrow (X, \mathcal{X})$, the *marginal distribution* of X with respect to μ , $\mu^X : \mathcal{X} \rightarrow [0, 1]$ by $\mu^X(A) := \mu(X^{-1}(A))$ for any $A \in \mathcal{X}$.

Definition 2.2.16 (Conditional distribution). Given a probability space $(\mu, \Omega, \mathcal{F})$ and variables $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$, the *conditional distribution* of Y given X is any Markov kernel $\mu^{Y|X} : X \rightarrow Y$ such that

$$\mu^{XY}(A \times B) = \int_A \mu^{Y|X}(B|x) d\mu^X(x) \quad \forall A \in \mathcal{X}, B \in \mathcal{Y}$$

$$\iff$$



Example 2.2.17 (Single-valued variable). We let $*$ stand for any single-valued variable $* : \Omega \rightarrow \{*\}$.

Markov kernel product notation

Three pairwise *product* operations involving Markov kernels can be defined: measure-kernel products, kernel-kernel products and kernel-function products. These are analogous to row vector-matrix products, matrix-matrix products and matrix-column vector products respectively.

Definition 2.2.18 (Measure-kernel product). Given $\mu \in \Delta(\mathcal{X})$ and $\mathbb{M} : X \rightarrow Y$, the *measure-kernel product* $\mu\mathbb{M} \in \Delta(Y)$ is given by

$$\mu\mathbb{M}(A) := \int_X \mathbb{M}(A|x)\mu(\mathrm{d}x)$$

for all $A \in \mathcal{Y}$.

Definition 2.2.19 (Kernel-kernel product). Given $\mathbb{M} : X \rightarrow Y$ and $\mathbb{N} : Y \rightarrow Z$, the *kernel-kernel product* $\mathbb{M}\mathbb{N} : X \rightarrow Z$ is given by

$$\mathbb{M}\mathbb{N}(A|x) := \int_Y \mathbb{N}(A|y)\mathbb{M}(\mathrm{d}y|x)$$

for all $A \in \mathcal{Z}$, $x \in X$.

Definition 2.2.20 (Kernel-function product). Given $\mathbb{M} : X \rightarrow Y$ and $f : Y \rightarrow Z$, the *kernel-function product* $\mathbb{M}f : X \rightarrow Z$ is given by

$$\mathbb{M}f(x) := \int_Y f(y)\mathbb{N}(\mathrm{d}y|x)$$

for all $x \in X$.

Definition 2.2.21 (Tensor product). Given $\mathbb{M} : X \rightarrow Y$ and $\mathbb{L} : W \rightarrow Z$, the tensor product $\mathbb{M} \otimes \mathbb{L} : X \times W \rightarrow Y \times Z$ is given by

$$(\mathbb{M} \otimes \mathbb{L})(A \times B|x, w) := \mathbb{M}(A|x)\mathbb{L}(B|w)$$

For all $x \in X$, $w \in W$, $A \in \mathcal{Y}$ and $B \in \mathcal{Z}$.

All products are associative (Çinlar, 2011, Chapter 1).

One application of the product notation is that marginal distributions can be alternatively defined in terms of a kernel product, as shown in Lemma 2.2.22.

Lemma 2.2.22 (Marginal distribution as a kernel product). *Given a probability space $(\mu, \Omega, \mathcal{F})$ and a variable $\mathsf{X} : \Omega \rightarrow (X, \mathcal{X})$, define $\mathbb{F}_{\mathsf{X}} : \Omega \rightarrow X$ by $\mathbb{F}_{\mathsf{X}}(A|\omega) = \delta_{\mathsf{X}(\omega)}(A)$, then*

$$\mu^{\mathsf{X}} = \mu\mathbb{F}_{\mathsf{X}}$$

Proof. Consider any $A \in \mathcal{X}$.

$$\begin{aligned} \mu\mathbb{F}_{\mathsf{X}}(A) &= \int_{\Omega} \delta_{\mathsf{X}(\omega)}(A) \mathrm{d}\mu(\omega) \\ &= \int_{\mathsf{X}^{-1}(A)} \mathrm{d}\mu(\omega) \\ &= \mu^{\mathsf{X}}(A) \end{aligned}$$

□

Semidirect product

Given a marginal μ^X and a conditional $\mu^{Y|X}$, the product of the two yields the marginal distribution of Y : $\mu^Y = \mu^X \mu^{Y|X}$. We define another product – the *semidirect* product \odot – as the product that yields the joint distribution of (X, Y) : $\mu^{XY} = \mu^X \odot \mu^{Y|X}$. The semidirect product is associative (Lemma 2.2.24)

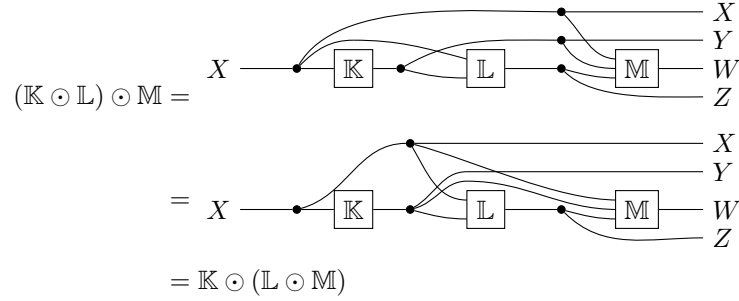
Definition 2.2.23 (Semidirect product). Given $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : Y \times X \rightarrow Z$, the semidirect product $\mathbb{K} \odot \mathbb{L} : X \rightarrow Y \times Z$ is given by

$$(\mathbb{K} \odot \mathbb{L})(A \times B|x) = \int_A \mathbb{L}(B|y, x) \mathbb{K}(dy|x) \quad \forall A \in \mathcal{Y}, B \in \mathcal{Z}$$

Lemma 2.2.24 (Semidirect product is associative). Given $\mathbb{K} : X \rightarrow Y$, $\mathbb{L} : Y \times X \rightarrow Z$ and $\mathbb{M} : Z \times Y \times X \rightarrow W$

$$(\mathbb{K} \odot \mathbb{L}) \odot \mathbb{M} = \mathbb{K} \odot (\mathbb{L} \odot \mathbb{M})$$

Proof.



□

The semidirect product can be used to define a notion of almost sure equality: two kernels $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : X \rightarrow Y$ are μ -almost surely equal if $\mu \odot \mathbb{K} = \mu \odot \mathbb{L}$. This is identical to the notion of almost sure equality in Cho and Jacobs (2019), who shows that under the assumption that (Y, \mathcal{Y}) is countably generated, $\mathbb{K} \stackrel{\mu}{\cong} \mathbb{L}$ if and only if $\mathbb{K} = \mathbb{L}$ μ -almost everywhere.

Definition 2.2.25 (Almost sure equality). Two Markov kernels $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : X \rightarrow Y$ are almost surely equal $\stackrel{\mathbb{P}_G}{\cong}$ with respect to a probability space (μ, X, \mathcal{X}) , written $\mathbb{K} \stackrel{\mu}{\cong} \mathbb{L}$ if

$$\mu \odot \mathbb{K} = \mu \odot \mathbb{L}$$

Theorem 2.2.26. Given (μ, X, \mathcal{X}) , $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : X \rightarrow Y$, $\mathbb{K} \stackrel{\mu}{\cong} \mathbb{L}$ if and only if, defining $U := \{x | \exists A \in \mathcal{Y} : \mathbb{K}(A|x) \neq \mathbb{L}(A|x)\}$, $\mu(U) = 0$.

Proof. Cho and Jacobs (2019) proposition 5.4. \square

We often want to talk about almost sure equality of two different versions \mathbb{K} and \mathbb{L} of a conditional distribution $\mathbb{P}^{Y|X}$ with respect to some ambient probability space $(\mathbb{P}, \Omega, \mathcal{F})$. This simply means \mathbb{K} and \mathbb{L} satisfy Definition 2.2.16 with respect to \mathbb{P} , X and Y , and they are almost surely equal with respect to the marginal \mathbb{P}^X . The relevant variables are usually obvious from the context and we leave them implicit and we will write $\mathbb{K} \stackrel{\mathbb{P}}{\cong} \mathbb{L}$. If the relevant marginal is ambiguous, we will instead write $\mathbb{K} \stackrel{\mathbb{P}^X}{\cong} \mathbb{L}$.

Definition 2.2.27 (Almost sure equality with respect to a pair of variables). Given $(\mathbb{P}, \Omega, \mathcal{F})$ and $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$, two Markov kernels $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : X \rightarrow Y$ are X -almost surely equal with respect to \mathbb{P} , written $\mathbb{K} \stackrel{\mathbb{P}}{\cong} \mathbb{L}$, if they are almost surely equal with respect to the marginal \mathbb{P}^X .

2.3 String Diagrams

We make use of string diagram notation for probabilistic reasoning. Graphical models are often employed in causal reasoning, and string diagrams are a kind of graphical notation for representing Markov kernels. The notation comes from the study of Markov categories, which are abstract categories that represent models of the flow of information. For our purposes, we don't use abstract Markov categories but instead focus on the concrete category of Markov kernels on standard measurable sets.

A coherence theorem exists for string diagrams and Markov categories. Applying planar deformation or any of the commutative comonoid axioms to a string diagram yields an equivalent string diagram. The coherence theorem establishes that any proof constructed using string diagrams in this manner corresponds to a proof in any Markov category (Selinger, 2011). More comprehensive introductions to Markov categories can be found in Fritz (2020); Cho and Jacobs (2019).

2.3.1 Elements of string diagrams

In the string, Markov kernels are drawn as boxes with input and output wires, and probability measures (which are Markov kernels with the domain $\{*\}$) are represented by triangles:

$$\begin{aligned} \mathbb{K} &:= \boxed{\mathbb{K}} \\ \mu &:= \triangleleft \mathbb{P} \end{aligned}$$

Given two Markov kernels $\mathbb{L} : X \rightarrow Y$ and $\mathbb{M} : Y \rightarrow Z$, the product $\mathbb{L}\mathbb{M}$ is represented by drawing them side by side and joining their wires:

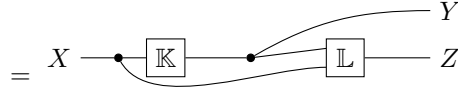
$$\mathbb{L}\mathbb{M} := X \begin{array}{|c|} \hline \mathbb{K} \\ \hline \mathbb{M} \\ \hline \end{array} Z$$

Given kernels $\mathbb{K} : W \rightarrow Y$ and $\mathbb{L} : X \rightarrow Z$, the tensor product $\mathbb{K} \otimes \mathbb{L} : W \times X \rightarrow Y \times Z$ is graphically represented by drawing kernels in parallel:

$$\mathbb{K} \otimes \mathbb{L} := \begin{array}{c} W \begin{array}{|c|} \hline \mathbb{K} \\ \hline \end{array} Y \\ X \begin{array}{|c|} \hline \mathbb{L} \\ \hline \end{array} Z \end{array}$$

Given $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : Y \times X \rightarrow Z$, the semidirect product is graphically represented by connecting \mathbb{K} and \mathbb{L} and keeping an extra copy

$$\mathbb{K} \odot \mathbb{L} := \text{copy}_X(\mathbb{K} \otimes \text{id}_X)(\text{copy}_Y \otimes \text{id}_X)(\text{id}_Y \otimes \mathbb{L})$$



A space X is identified with the identity kernel $\text{id}^X : X \rightarrow \Delta(\mathcal{X})$. A bare wire represents the identity kernel:

$$\text{Id}^X := X \text{ ————— } X$$

Product spaces $X \times Y$ are identified with tensor product of identity kernels $\text{id}^X \otimes \text{id}^Y$. These can be represented either by two parallel wires or by a single wire representing the identity on the product space $X \times Y$:

$$\begin{aligned} X \times Y \cong \text{Id}^X \otimes \text{Id}^Y &:= \begin{array}{c} X \text{ — } X \\ Y \text{ — } Y \end{array} \\ &= X \times Y \text{ — } X \times Y \end{aligned}$$

A kernel $\mathbb{L} : X \rightarrow \Delta(\mathcal{Y} \otimes \mathcal{Z})$ can be written using either two parallel output wires or a single output wire, appropriately labeled:

$$\begin{aligned} X \text{ — } \begin{array}{|c|} \hline \mathbb{L} \\ \hline \end{array} \begin{array}{c} Y \\ Z \end{array} \\ \equiv \\ X \text{ — } \begin{array}{|c|} \hline \mathbb{L} \\ \hline \end{array} \text{ — } Y \times Z \end{aligned}$$

We read diagrams from left to right (this is somewhat different to Fritz (2020); Cho and Jacobs (2019); Fong (2013) but in line with Selinger (2011)), and any diagram describes a set of nested products and tensor products of Markov kernels. There are a collection of special Markov kernels for which we can replace the generic “box” of a Markov kernel with a diagrammatic elements that are visually suggestive of what these kernels accomplish.

2.3.2 Special maps

Definition 2.3.1 (Identity map). The identity map $\text{id}_X : X \rightarrow X$ defined by $(\text{id}_X)(A|x) = \delta_x(A)$ for all $x \in X$, $A \in \mathcal{X}$, is represented by a bare line.

$$\text{id}_X := X \text{---} X$$

Definition 2.3.2 (Erase map). Given some 1-element set $\{*\}$, the erase map $\text{del}_X : X \rightarrow \{*\}$ is defined by $(\text{del}_X)(*|x) = 1$ for all $x \in X$. It “discards the input”. It looks like a lit fuse:

$$\text{del}_X := \text{---} * X$$

Definition 2.3.3 (Swap map). The swap map $\text{swap}_{X,Y} : X \times Y \rightarrow Y \times X$ is defined by $(\text{swap}_{X,Y})(A \times B|x, y) = \delta_x(B)\delta_y(A)$ for $(x, y) \in X \times Y$, $A \in \mathcal{X}$ and $B \in \mathcal{Y}$. It swaps two inputs and is represented by crossing wires:

$$\text{swap}_{X,Y} := \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array}$$

Definition 2.3.4 (Copy map). The copy map $\text{copy}_X : X \rightarrow X \times X$ is defined by $(\text{copy}_X)(A \times B|x) = \delta_x(A)\delta_x(B)$ for all $x \in X$, $A, B \in \mathcal{X}$. It makes two identical copies of the input, and is drawn as a fork:

$$\text{copy}_X := X \text{---} \begin{array}{c} \diagup \\ \diagdown \end{array} \begin{array}{c} X \\ X \end{array}$$

Definition 2.3.5 (n -fold copy map). The n -fold copy map $\text{copy}_X^n : X \rightarrow X^n$ is given by the recursive definition

$$\begin{aligned} \text{copy}_X^1 &= \text{copy}_X \\ \text{copy}_X^n &= \begin{array}{c} \text{---} \boxed{\text{copy}_X^{n-1}} \text{---} \\ \bullet \\ \text{---} \end{array} \quad n > 1 \end{aligned}$$

Plates In a string diagram, a plate that is annotated $i \in A$ means the tensor product of the $|A|$ elements that appear inside the plate. A wire crossing from outside a plate boundary to the inside of a plate indicates an $|A|$ -fold copy map, which we indicate by placing a dot on the plate boundary. For our purposes, we do not define anything that allows wires to cross from the inside of a plate to the outside; wires must terminate within the plate.

Thus, given $\mathbb{K}_i : X \rightarrow Y$ for $i \in A$,

$$\bigotimes_{i \in A} \mathbb{K}_i := \boxed{\begin{array}{c} \text{---} \boxed{\mathbb{K}_i} \text{---} \\ i \in A \end{array}} \text{copy}_X^{|A|} \left(\bigotimes_{i \in A} \mathbb{K}_i \right) := \text{---} \bullet \boxed{\begin{array}{c} \boxed{\mathbb{K}_i} \text{---} \\ i \in A \end{array}}$$

2.3.3 Commutative comonoid axioms

Diagrams in Markov categories satisfy the commutative comonoid axioms.

$$\begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \text{---} \\ \searrow \text{---} \end{array} = \text{---} \bullet \begin{array}{l} \nearrow \text{---} \\ \searrow \text{---} \end{array} \end{array} \quad (2.1)$$

$$\begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \text{---} \\ \searrow \text{---} \end{array} = \text{---} = \text{---} \bullet \begin{array}{l} \nearrow \text{---} \\ \searrow \text{---} \end{array} \end{array} \quad (2.2)$$

$$\begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \text{---} \\ \searrow \text{---} \end{array} = \text{---} \bullet \begin{array}{l} \nearrow \text{---} \\ \searrow \text{---} \end{array} \end{array}$$

as well as compatibility with the monoidal structure

$$\begin{array}{c} X \otimes Y \text{---} * \\ = \\ X \text{---} * \end{array} \quad \begin{array}{c} X \text{---} * \\ = \\ X \text{---} * \end{array}$$

$$\begin{array}{c} X \otimes Y \text{---} \bullet \begin{array}{l} \nearrow X \otimes Y \\ \searrow X \otimes Y \end{array} = \begin{array}{c} X \text{---} \bullet \begin{array}{l} \nearrow X \\ \searrow Y \end{array} \\ Y \text{---} \bullet \begin{array}{l} \nearrow X \\ \searrow Y \end{array} \end{array}$$

and the naturality of del , which means that

$$\begin{array}{c} \text{---} \boxed{f} \text{---} * \\ = \\ \text{---} * \end{array} \quad (2.3)$$

2.3.4 Manipulating String Diagrams

Planar deformations along with the applications of Equations (2.1) through to Equation (2.3) are almost the only rules we have for transforming one string diagram into an equivalent one. One further rule is given by Theorem 2.3.6.

Theorem 2.3.6 (Copy map commutes for deterministic kernels (Fong, 2013)).
For $\mathbb{K} : X \rightarrow Y$

$$\begin{array}{c} X \text{---} \boxed{\mathbb{K}} \bullet \begin{array}{l} \nearrow Y \\ \searrow Y \end{array} = \begin{array}{c} X \text{---} \bullet \begin{array}{l} \nearrow \boxed{\mathbb{K}} \text{---} Y \\ \searrow \boxed{\mathbb{K}} \text{---} Y \end{array} \end{array}$$

holds iff \mathbb{K} is deterministic.

Examples

String diagrams can always be converted into definitions involving integrals and tensor products. A number of shortcuts can help to make the translations efficiently.

For arbitrary $\mathbb{K} : X \times Y \rightarrow Z$, $\mathbb{L} : W \rightarrow Y$

$$\begin{aligned}
 & \text{Diagram: Two horizontal lines enter from the left. The bottom line passes through a box labeled \mathbb{L} . Both lines then enter a box labeled \mathbb{K} . A single line exits the right side of the \mathbb{K} box.} \\
 & = (\text{id}_X \otimes \mathbb{L})\mathbb{K} \\
 & [(\text{id}_X \otimes \mathbb{L})\mathbb{K}](A|x, w) = \int_Y \int_X \mathbb{K}(A|x', y') \mathbb{L}(dy'|w) \delta_x(dx') \\
 & = \int_Y \mathbb{K}(A|x, y') \mathbb{L}(dy'|w)
 \end{aligned}$$

That is, an identity map “passes its input directly to the next kernel”.

For arbitrary $\mathbb{K} : X \times Y \times Y \rightarrow Z$:

$$\begin{aligned}
 & \text{Diagram: Two horizontal lines enter from the left. A dot on the bottom line has two curved lines branching off to the right, entering a box labeled \mathbb{K} . The top line also enters the \mathbb{K} box. A single line exits the right side of the \mathbb{K} box.} \\
 & = (\text{id}_X \otimes \text{copy}_Y)\mathbb{K} \\
 & [(\text{id}_X \otimes \text{copy}_Y)\mathbb{K}](A|x, y) = \int_Y \int_Y \mathbb{K}(A|x, y', y'') \delta_y(dy') \delta_y(dy'') \\
 & = \mathbb{K}(A|x, y, y)
 \end{aligned}$$

That is, the copy map “passes along two copies of its input” to the next kernel in the product.

For arbitrary $\mathbb{K} : X \times Y \rightarrow Z$

$$\begin{aligned}
 & \text{Diagram: Two horizontal lines enter from the left, cross each other, and then enter a box labeled \mathbb{K} . A single line exits the right side of the \mathbb{K} box.} \\
 & = \text{swap}_{YX} \mathbb{K} \\
 & (\text{swap}_{YX} \mathbb{K})(A|y, x) = \int_{X \times Y} \mathbb{K}(A|x', y') \delta_y(dy') \delta_x(dx') \\
 & = \mathbb{K}(A|x, y)
 \end{aligned}$$

The swap map before a kernel switches the input arguments.

For arbitrary $\mathbb{K} : X \rightarrow Y \times Z$

$$\begin{aligned}
 & \text{Diagram: A single horizontal line enters from the left, enters a box labeled \mathbb{K} , and then splits into two lines that cross each other.} \\
 & = \mathbb{K} \text{swap}_{YZ} \\
 & (\mathbb{K} \text{swap}_{YZ})(A \times B|x) = \int_{Y \times Z} \delta_y(B) \delta_z(A) \mathbb{K}(dy \times dz|x) \\
 & = \int_{B \times A} \mathbb{K}(dy \times dz|x) \\
 & = \mathbb{K}(B \times A|x)
 \end{aligned}$$

Given $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : Y \rightarrow Z$:

$$\begin{aligned}
 (\mathbb{K} \odot \mathbb{L})(\text{id}_Y \otimes \text{del}_Z) &= \begin{array}{c} \text{Diagram 1: } X \text{ -- } \boxed{\mathbb{K}} \text{ -- } \bullet \text{ -- } \begin{array}{l} \text{--- } Y \\ \text{--- } \boxed{\mathbb{L}} \text{ -- } * \end{array} \end{array} \\
 &= \begin{array}{c} \text{Diagram 2: } X \text{ -- } \boxed{\mathbb{K}} \text{ -- } \bullet \text{ -- } \begin{array}{l} \text{--- } Y \\ \text{--- } * \end{array} \end{array} \quad \text{by Eq. (2.3)} \\
 &= \begin{array}{c} \text{Diagram 3: } X \text{ -- } \boxed{\mathbb{K}} \text{ -- } Y \end{array} \quad \text{by Eq. (2.2)}
 \end{aligned}$$

Thus the action of the del map is to marginalise over the deleted wire. With integrals, we can write

$$\begin{aligned}
 (\mathbb{K} \odot \mathbb{L})(\text{id}_Y \otimes \text{del}_Z)(A \times \{*\}|x) &= \int_Y \int_{\{*\}} \delta_y(A) \delta_*(\{*\}) \mathbb{L}(\text{d}z|y) \mathbb{K}(\text{d}y|x) \\
 &= \int_A \mathbb{K}(\text{d}y|x) \\
 &= \mathbb{K}(A|x)
 \end{aligned}$$

2.4 Probability Sets

A probability set is a set of probability measures. This section establishes a number of useful properties of conditional probability with respect to probability sets. Unlike conditional probability with respect to a probability space, conditional probabilities don't always exist for probability sets. Where they do, however, they are almost surely unique and we can marginalise and disintegrate them to obtain other conditional probabilities with respect to the same probability set.

Definition 2.4.1 (Probability set). A probability set \mathbb{P}_C on (Ω, \mathcal{F}) is a collection of probability measures on (Ω, \mathcal{F}) . In other words it is a subset of $\mathcal{P}(\Delta(\Omega))$, where \mathcal{P} indicates the power set.

Given a probability set \mathbb{P}_C , we define marginal and conditional probabilities as probability measures and Markov kernels that satisfy Definitions 2.2.15 and 2.2.16 respectively for *all* base measures in \mathbb{P}_C . There are generally multiple Markov kernels that satisfy the properties of a conditional probability with respect to a probability set, and this definition ensures that marginal and conditional probabilities are “almost surely” unique (Definition 2.4.7) with respect to probability sets.

Definition 2.4.2 (Marginal probability with respect to a probability set). Given a sample space (Ω, \mathcal{F}) , a variable $X : \Omega \rightarrow X$ and a probability set \mathbb{P}_C , the marginal distribution $\mathbb{P}_C^X = \mathbb{P}_\alpha^X$ for any $\mathbb{P}_\alpha \in \mathbb{P}_C$ if a distribution satisfying this condition exists. Otherwise, it is undefined.

Definition 2.4.3 (Uniform conditional distribution). Given a sample space (Ω, \mathcal{F}) , variables $X : \Omega \rightarrow X$ and $Y : \Omega \rightarrow Y$ and a probability set \mathbb{P}_C , a uniform conditional distribution $\mathbb{P}_C^{Y|X}$ is any Markov kernel $X \rightarrow Y$ such that $\mathbb{P}_C^{Y|X}$ is a $Y|X$ conditional probability of \mathbb{P}_α for all $\mathbb{P}_\alpha \in \mathbb{P}_C$. If no such Markov kernel exists, $\mathbb{P}_C^{Y|X}$ is undefined.

Given a conditional distribution $\mu^{ZY|X}$ we can define a higher order conditional $\mu^{Z|(Y|X)}$, which is a version of $\mu^{Z|XY}$. This is useful because uniform conditionals don't always exist, but we can use higher order conditionals to show that if a probability set \mathbb{P}_C has a uniform conditional $\mathbb{P}_C^{ZY|X}$ then it also has a uniform conditional $\mathbb{P}_C^{Z|XY}$ (Theorems 2.4.31 and 2.4.33). Given $\mu^{XY|Z}$ and $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$ standard measurable, it has recently been proven that a higher order conditional $\mu^{Z|(Y|X)}$ exists Bogachev and Malofeev (2020), Theorem 3.5.

Definition 2.4.4 (Higher order conditionals). Given a probability space $(\mu, \Omega, \mathcal{F})$ and variables $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$ and $Z : \Omega \rightarrow Z$, a higher order conditional $\mu^{Z|(Y|X)} : X \times Y \rightarrow Z$ is any Markov kernel such that, for some $\mu^{Y|X}$,

$$\mu^{ZY|X}(B \times C|x) = \int_B \mu^{Z|(Y|X)}(C|x, y) \mu^{Y|X}(dy|x)$$

$$\iff$$

$\mu^{ZY|X} =$

Definition 2.4.5 (Uniform higher order conditional). Given a sample space (Ω, \mathcal{F}) , variables $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$ and $Z : \Omega \rightarrow Z$ and a probability set \mathbb{P}_C , if $\mathbb{P}_C^{ZY|X}$ exists then a uniform higher order conditional $\mathbb{P}_C^{Z|(Y|X)}$ is any Markov kernel $X \times Y \rightarrow Z$ that is a higher order conditional of some version of $\mathbb{P}_C^{ZY|X}$. If no $\mathbb{P}_C^{ZY|X}$ exists, $\mathbb{P}_C^{Z|(Y|X)}$ is undefined.

Definition 2.4.6 (Almost sure equality). Two Markov kernels $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : X \rightarrow Y$ are \mathbb{P}_C, X, Y -almost surely equal if for all $A \in \mathcal{X}$, $B \in \mathcal{Y}$, $\alpha \in C$

$$\int_A \mathbb{K}(B|x) \mathbb{P}_\alpha^\mathbb{X}(dx) = \int_A \mathbb{L}(B|x) \mathbb{P}_\alpha^\mathbb{X}(dx)$$

we write this as $\mathbb{K} \stackrel{\mathbb{P}_C}{\cong} \mathbb{L}$, as the variables X and Y are clear from the context.

Equivalently, \mathbb{K} and \mathbb{L} are almost surely equal if the set $C : \{x | \exists B \in \mathcal{Y} : \mathbb{K}(B|x) \neq \mathbb{L}(B|x)\}$ has measure 0 with respect to $\mathbb{P}_\alpha^\mathbb{X}$ for all $\alpha \in C$.

2.4.1 Almost sure equality

Two Markov kernels are almost surely equal with respect to a probability set \mathbb{P}_C if the semidirect product \odot of all marginal probabilities of \mathbb{P}_α^X with each Markov kernel is identical.

Definition 2.4.7 (Almost sure equality). Two Markov kernels $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : X \rightarrow Y$ are almost surely equal $\stackrel{\mathbb{P}_C}{\cong}$ with respect to a probability set \mathbb{P}_C and variable $X : \Omega \rightarrow X$ if for all $\mathbb{P}_\alpha \in \mathbb{P}_C$,

$$\mathbb{P}_\alpha^X \odot \mathbb{K} = \mathbb{P}_\alpha^X \odot \mathbb{L}$$

Lemma 2.4.8 (Uniform conditional distributions are almost surely equal). *If $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : X \rightarrow Y$ are both versions of $\mathbb{P}_C^{Y|X}$ then $\mathbb{K} \stackrel{\mathbb{P}_C}{\cong} \mathbb{L}$*

Proof. For all $\mathbb{P}_\alpha \in \mathbb{P}_C$

$$\begin{aligned} \mathbb{P}_\alpha^X \odot \mathbb{K} &= \mathbb{P}_\alpha^{XY} \\ &= \mathbb{P}_\alpha^X \odot \mathbb{L} \end{aligned}$$

□

Lemma 2.4.9 (Substitution of almost surely equal Markov kernels). *Given \mathbb{P}_C , if $\mathbb{K} : X \times Y \rightarrow Z$ and $\mathbb{L} : X \times Y \rightarrow Z$ are almost surely equal $\mathbb{K} \stackrel{\mathbb{P}_C}{\cong} \mathbb{L}$, then for any $\mathbb{P}_\alpha \in \mathbb{P}_C$*

$$\mathbb{P}_\alpha^{Y|X} \odot \mathbb{K} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_\alpha^{Y|X} \odot \mathbb{L}$$

Proof. For any $\mathbb{P}_\alpha \in \mathbb{P}_C$

$$\begin{aligned} \mathbb{P}_\alpha^{XY} \odot \mathbb{K} &\stackrel{\mathbb{P}_C}{\cong} (\mathbb{P}_\alpha^X \odot \mathbb{P}_C^{Y|X}) \odot \mathbb{K} \\ &\stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_\alpha^X \odot (\mathbb{P}_C^{Y|X} \odot \mathbb{K}) \\ &\stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_\alpha^X \odot (\mathbb{P}_C^{Y|X} \odot \mathbb{L}) \end{aligned}$$

□

Theorem 2.4.10 (Semidirect product of uniform conditional distributions is a joint uniform conditional distribution). *Given a probability set \mathbb{P}_C on (Ω, \mathcal{F}) , variables $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$ and uniform conditional distributions $\mathbb{P}_C^{Y|X}$ and $\mathbb{P}_C^{Z|XY}$, then $\mathbb{P}_C^{YZ|X}$ exists and is equal to*

$$\mathbb{P}_C^{YZ|X} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y|X} \odot \mathbb{P}_C^{Z|XY}$$

Proof. By definition, for any $\mathbb{P}_\alpha \in \mathbb{P}_C$

$$\begin{aligned}\mathbb{P}_\alpha^{XYZ} &= \mathbb{P}_\alpha^X \odot \mathbb{P}_\alpha^{YZ|X} \\ &= \mathbb{P}_\alpha^X \odot (\mathbb{P}_\alpha^{Y|X} \odot \mathbb{P}_\alpha^{Z|YX}) \\ &= \mathbb{P}_\alpha^X \odot (\mathbb{P}_C^{Y|X} \odot \mathbb{P}_C^{Z|YX})\end{aligned}$$

□

2.4.2 Extended conditional independence

Just like we defined uniform conditional probability as a version of “conditional probability” appropriate for probability sets, we need some version of “conditional independence” for probability sets. One such has already been given in some detail: it is the idea of *extended conditional independence* defined in Constantinou and Dawid (2017).

We will first define regular conditional independence. We define it in terms of a having a conditional that “ignores one of its inputs”, which, provided conditional probabilities exists, is equivalent to other common definitions (Theorem 2.4.12).

Definition 2.4.11 (Conditional independence). For a *probability model* \mathbb{P}_α and variables A, B, Z , we say B is conditionally independent of A given C , written $B \perp\!\!\!\perp_{\mathbb{P}_\alpha} A|C$, if

$$\begin{aligned}\mathbb{P}^{Y|WX} &\stackrel{\mathbb{P}}{\cong} \begin{array}{c} W \text{ --- } \boxed{\mathbb{K}} \text{ --- } Y \\ X \text{ --- } * \end{array} \\ \iff \mathbb{P}^{Y|WX}(A|w, x) &\stackrel{\mathbb{P}}{\cong} \mathbb{K}(A|w) \quad \forall A \in \mathcal{Y}\end{aligned}$$

Conditional independence can equivalently be stated in terms of the existence of a conditional probability that “ignores” one of its inputs.

Theorem 2.4.12. *Given standard measurable (Ω, \mathcal{F}) , a probability model \mathbb{P} and variables $W : \Omega \rightarrow W$, $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$, $Y \perp\!\!\!\perp_{\mathbb{P}} X|W$ if and only if there exists some version of $\mathbb{P}^{Y|WX}$ and $\mathbb{K} : W \rightarrow Y$ such that*

$$\begin{aligned}\mathbb{P}^{Y|WX} &\stackrel{\mathbb{P}}{\cong} \begin{array}{c} W \text{ --- } \boxed{\mathbb{K}} \text{ --- } Y \\ X \text{ --- } * \end{array} \\ \iff \mathbb{P}^{Y|WX}(A|w, x) &\stackrel{\mathbb{P}}{\cong} \mathbb{K}(A|w) \quad \forall A \in \mathcal{Y}\end{aligned}$$

Proof. See Cho and Jacobs (2019). □

Extended conditional independence as introduced by Constantinou and Dawid (2017) is defined in terms of “nonstochastic variables” on the choice set C . A nonstochastic variable is essentially a variable defined on C rather than on the sample space Ω

Definition 2.4.13 (Nonstochastic variable). Given a sample space (Ω, \mathcal{F}) , a choice set (C, \mathcal{C}) , a codomain (X, \mathcal{X}) and a probability set \mathbb{P}_C , a nonstochastic variable is a measurable function $\phi : C \rightarrow X$.

In particular, we want to consider *complementary* nonstochastic variable - that is, pairs of nonstochastic variables ϕ and ξ such that the sequence (ϕ, ξ) is invertible. For example, if $\phi := \text{idf}_C$, then

Definition 2.4.14 (Complementary nonstochastic variables). A pair of nonstochastic variables ϕ and ξ are complementary if (ϕ, ξ) is invertible.

Nxample 2.4.15. The letters ϕ and ξ are used to represent complementary nonstochastic variables.

Unlike Constantinou and Dawid (2017), we limit ourselves to a definition of extended conditional independence where regular uniform conditional probabilities exist. Our definition is otherwise identical.

Definition 2.4.16 (Extended conditional independence). Given a probability set \mathbb{P}_C , variables X, Y and Z and complementary nonstochastic variables ϕ and ξ , the extended conditional independence $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e X \phi | Z \xi$ holds if for each $a \in \xi(C)$, $\mathbb{P}_{\xi^{-1}(a)}^{Y|XZ}$ and $\mathbb{P}_{\xi^{-1}(a)}^{Y|X}$ exist and

$$\begin{array}{ccc} & & \begin{array}{c} Z \text{ --- } \boxed{\mathbb{P}_C^{Y|Z}} \text{ --- } Y \\ X \text{ --- } * \end{array} \\ \mathbb{P}_{\xi^{-1}(a)}^{Y|XZ} & \stackrel{\mathbb{P}_{\xi^{-1}(a)}}{\cong} & \\ \iff & & \\ \mathbb{P}_{\xi^{-1}(a)}^{Y|XZ}(A|x, z) & \stackrel{\mathbb{P}_{\xi^{-1}(a)}}{\cong} & \mathbb{P}_{\xi^{-1}(a)}^{Y|Z}(A|z) \quad \forall A \in \mathcal{Y}, (x, z) \in X \times Z \end{array}$$

Very often, we consider a particular kind of extended conditional independence that does not explicitly make use of nonstochastic variables. We call this *uniform conditional independence*.

Definition 2.4.17 (Uniform conditional independence). Given a probability set \mathbb{P}_C and variables X, Y and Z , the uniform conditional independence $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e XC | Z$ holds if $\mathbb{P}_C^{Y|XZ}$ and $\mathbb{P}_C^{Y|X}$ exist and

$$\begin{array}{ccc} & & \begin{array}{c} Z \text{ --- } \boxed{\mathbb{P}_C^{Y|Z}} \text{ --- } Y \\ X \text{ --- } * \end{array} \\ \mathbb{P}_C^{Y|XZ} & \stackrel{\mathbb{P}_C}{\cong} & \\ \iff & & \\ \mathbb{P}_C^{Y|XZ}(A|x, z) & \stackrel{\mathbb{P}_C}{\cong} & \mathbb{P}_C^{Y|Z}(A|z) \quad \forall A \in \mathcal{Y}, (x, z) \in X \times Z \end{array}$$

For countable sets C (which, recall, is an assumption we generally accept), as shown by Constantinou and Dawid (2017) we can reason with collections of extended conditional independence statements as if they were regular conditional independence statements, with the provision that a complementary pair of nonstochastic variables must appear either side of the “|” symbol.

1. Symmetry: $X \perp\!\!\!\perp_{\mathbb{P}_C}^e Y\phi|Z\xi$ iff $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e X\phi|Z\xi$
2. $X \perp\!\!\!\perp_{\mathbb{P}_C}^e YC|YC$
3. Decomposition: $X \perp\!\!\!\perp_{\mathbb{P}_C}^e Y\phi|W\xi$ and $Z \preceq Y$ implies $X \perp\!\!\!\perp_{\mathbb{P}_C}^e Z\phi|W\xi$
4. Weak union:
 - (a) $X \perp\!\!\!\perp_{\mathbb{P}_C}^e Y\phi|W\xi$ and $Z \preceq Y$ implies $X \perp\!\!\!\perp_{\mathbb{P}_C}^e Y\phi|(Z, W)\xi$
 - (b) $X \perp\!\!\!\perp_{\mathbb{P}_C}^e Y\phi|W\xi$ and $\lambda \preceq \phi$ implies $X \perp\!\!\!\perp_{\mathbb{P}_C}^e Y\phi|(Z, W)(\xi, \lambda)$
5. Contraction: $X \perp\!\!\!\perp_{\mathbb{P}_C}^e Z\phi|W\xi$ and $X \perp\!\!\!\perp_{\mathbb{P}_C}^e Y\phi|(Z, W)\xi$ implies $X \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y, Z)\phi|W\xi$

The following forms of these properties are often used here:

1. Symmetry: $X \perp\!\!\!\perp_{\mathbb{P}_C}^e YC|Z$ iff $Y \perp\!\!\!\perp_{\mathbb{P}}^e XC|Z$
2. Decomposition: $X \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y, Z)C|W$ implies $X \perp\!\!\!\perp_{\mathbb{P}}^e YC|W$ and $X \perp\!\!\!\perp_{\mathbb{P}}^e ZC|W$
3. Weak union: $X \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y, Z)C|W$ implies $X \perp\!\!\!\perp_{\mathbb{P}_C}^e YC|(Z, W)$
4. Contraction: $X \perp\!\!\!\perp_{\mathbb{P}_C}^e ZC|W$ and $X \perp\!\!\!\perp_{\mathbb{P}_C}^e YC|(Z, W)$ implies $X \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y, Z)C|W$

2.4.3 Examples

Example 2.4.18 (Choice variable). Suppose we have a decision procedure $\mathcal{S}_C := \{\mathcal{S}_\alpha | \alpha \in C\}$ that consists of a measurement procedure for each element of a denumerable set of choices C . Each measurement procedure \mathcal{S}_α is modeled by a probability distribution \mathbb{P}_α on a shared sample space (Ω, \mathcal{F}) such that we have an observable “choice” variable $(D, D \circ \mathcal{S}_\alpha)$ where $D \circ \mathcal{S}_\alpha$ always yields α .

Furthermore, Define $Y : \Omega \rightarrow \Omega$ as the identity function. Then, by supposition, for each $\alpha \in A$, \mathbb{P}_α^{YC} exists and for $A \in \mathcal{Y}$, $B \in \mathcal{C}$:

$$\mathbb{P}_\alpha^{YC}(A \times B) = \mathbb{P}_\alpha(A)\delta_\alpha(B)$$

This implies, for all $\alpha \in C$

$$\mathbb{P}_\alpha^{Y|D} = \mathbb{P}_\alpha^Y$$

Thus $\mathbb{P}_C^{Y|D}$ exists and

$$\mathbb{P}_C^{Y|D}(A|\alpha) = \mathbb{P}_\alpha^Y(A) \quad \forall A \in \mathcal{Y}, \alpha \in C$$

Because only deterministic marginals \mathbb{P}_α^D are available, for every $\alpha \in C$ we have $Y \perp\!\!\!\perp_{\mathbb{P}_\alpha}^e D$. This reflects the fact that *after we have selected a choice* α the value of C provides no further information about the distribution of Y , because D is deterministic given any α . It does not reflect the fact that “choosing different values of C has no effect on Y ”.

Theorem 2.4.19 (Uniform conditional independence representation). *Given a probability set \mathbb{P}_C with a uniform conditional probability $\mathbb{P}_C^{XY|Z}$,*

$$\begin{array}{c} \mathbb{P}_C^{XY|Z} \stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{Y|Z}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{X|Z}} \text{ --- } X \end{cases} \end{array} \\ \iff \\ \mathbb{P}_C^{XY|Z}(A \times B|z) \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{X|Z}(A|z) \mathbb{P}_C^{Y|Z}(B|z) \quad \forall A \in \mathcal{X}, B \in \mathcal{Y}, z \in Z \end{array}$$

if and only if $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e XC|Z$

Proof. If: By Theorem 2.4.33

$$\begin{array}{c} \mathbb{P}_C^{XY|Z} = \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{Y|ZX}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{X|Z}} \text{ --- } X \end{cases} \end{array} \\ \stackrel{\mathbb{P}}{\parallel} \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{Y|Z}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{X|Z}} \text{ --- } X \end{cases} \end{array} \\ = \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{Y|Z}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{X|Z}} \text{ --- } X \end{cases} \end{array} \end{array}$$

Only if: Suppose

$$\mathbb{P}_C^{XY|Z} \stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{Y|Z}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{X|Z}} \text{ --- } X \end{cases} \end{array}$$

and suppose for some $\alpha \in C$, $A \times C \in \mathcal{X} \otimes \mathcal{Z}$, $B \in \mathcal{Y}$ $\mathbb{P}_\alpha^{XZ}(A \times C) > 0$ and

$$\mathbb{P}_C^{Y|XZ}(B|x, z) > \mathbb{P}_C^{Y|Z}(B|z) \quad \forall (x, z) \in A \times C \quad (2.4)$$

then

$$\begin{aligned} \mathbb{P}_\alpha^{XYZZ}(A \times B \times C) &= \int_{A \times C} \mathbb{P}_C^{Y|XZ}(B|x, z) \mathbb{P}_C^{X|Z}(dx|z) \mathbb{P}_\alpha^Z(dz) \\ &> \int_{A \times C} \mathbb{P}_C^{Y|X}(B|z) \mathbb{P}_C^{X|Z}(dx|z) \mathbb{P}_\alpha^Z(dz) \\ &= \int_C \mathbb{P}_C^{XY|X}(A \times B|z) \mathbb{P}_\alpha^Z(dz) \\ &= \mathbb{P}_\alpha^{XYZZ}(A \times B \times C) \end{aligned}$$

a contradiction. An analogous argument follows if we replace “>” with “<” in Eq. (2.4). \square

2.4.4 Maximal probability sets and valid conditionals

So far, we have been implicitly supposing that we first set up a probability set and from that set we may sometimes derive uniform conditional probabilities, extended conditional independences and so forth. However, sometimes we want to work backwards: start with a collection of uniform conditional probabilities, and work with the probability set implicitly defined by this collection. For example, when we have a Causal Bayesian Network, the collection of operations of the form “do($X = x$)” specify a probability set by a collection of uniform conditional probabilities on variables other than X , along with marginal probabilities of X . Specifically:

$$\mathbb{P}_{X=x}^{Y|\text{Pa}(Y)} = \begin{cases} \mathbb{P}_{\text{obs}}^{Y|\text{Pa}(Y)} & Y \text{ is a causal variable and not equal to } X \\ \delta_x & Y = X \end{cases}$$

The qualification “ Y is a causal variable” is usually not an explicit condition for causal Bayesian networks, but it is an important one. For example, $2X$ is not equal to X , but we cannot define a causal Bayesian network where both X and $2X$ are causal variables, see Example 2.4.27.

When working backwards like this, we can run into a couple of problems: we may end up with a probability set where some probabilities are non-unique, or we might inadvertently define an empty probability set. *Validity* is a condition that can ensure that we at least avoid the second problem.

Thus, if we start with a probability set, we know how to check if certain uniform conditional probabilities exist or not. However, there is a particular line of reasoning that comes up most often in the graphical models tradition of causal inference where we start with collections of conditional probabilities and assemble them into probability models as needed. A simple example of this is the causal Bayesian network given by the graph $X \longrightarrow Y$ and some observational probability distribution $\mathbb{P}^{XY} \in \Delta(X \times Y)$. Using the standard notion of “hard interventions on X ”, this model induces a probability set which we could informally describe as the set $\mathbb{P}_{\square} := \{\mathbb{P}_a^{XY} | a \in X \cup \{o\}\}$ where o is a special element corresponding to the observational setting. The graph $X \longrightarrow Y$ implies the existence of the uniform conditional probability $\mathbb{P}_{\square}^{Y|X}$ under the nominated set of interventions, while the usual rules of hard interventions imply that $\mathbb{P}_a^X = \delta_a$ for $a \in X$.

Reasoning “backwards” like this – from uniform conditionals and marginals back to probability sets – must be done with care. The probability set associated with a collection of conditionals and marginals may be empty or nonunique. Uniqueness may not always be required, but an empty probability set is clearly not a useful model.

Consider, for example, $\Omega = \{0, 1\}$ with $X = (Z, Z)$ for $Z := \text{id}_{\Omega}$ and any measure $\kappa \in \Delta(\{0, 1\}^2)$ such that $\kappa(\{1\} \times \{0\}) > 0$. Note that $X^{-1}(\{1\} \times \{0\}) =$

$Z^{-1}(\{1\}) \cap Z^{-1}(\{0\}) = \emptyset$. Thus for any probability measure $\mu \in \Delta(\{0, 1\})$, $\mu^{\mathbf{X}}(\{1\} \times \{0\}) = \mu(\emptyset) = 0$ and so κ cannot be the marginal distribution of \mathbf{X} for any base measure at all.

We introduce the notion of *valid distributions* and *valid conditionals*. The key result here is: probability sets defined by collections of recursive valid conditionals and distributions are nonempty. While we suspect this condition is often satisfied by causal models in practice, we offer one example in the literature where it apparently is not. The problem of whether a probability set is valid is analogous to the problem of whether a probability distribution satisfying a collection of constraints exists discussed in Vorobev (1962). As that work shows, there are many questions of this nature that can be asked and that are not addressed by the criterion of validity.

There is also a connection between the notion of validity and the notion of *unique solvability* in Bongers et al. (2016). We ask “when can a set of conditional probabilities together with equations be jointly satisfied by a probability model?” while Bongers et. al. ask when a set of equations can be jointly satisfied by a probability model.

Definition 2.4.20 (Valid distribution). Given (Ω, \mathcal{F}) and a variable $\mathbf{X} : \Omega \rightarrow X$, an \mathbf{X} -valid probability distribution is any probability measure $\mathbb{K} \in \Delta(X)$ such that $\mathbf{X}^{-1}(A) = \emptyset \implies \mathbb{K}(A) = 0$ for all $A \in \mathcal{X}$.

Definition 2.4.21 (Valid conditional). Given (Ω, \mathcal{F}) , $\mathbf{X} : \Omega \rightarrow X$, $\mathbf{Y} : \Omega \rightarrow Y$ a $\mathbf{Y}|\mathbf{X}$ -valid conditional probability is a Markov kernel $\mathbb{L} : X \rightarrow Y$ that assigns probability 0 to impossible events, unless the argument itself corresponds to an impossible event:

$$\forall B \in \mathcal{Y}, x \in X : (\mathbf{X}, \mathbf{Y}) \bowtie \{x\} \times B = \emptyset \implies (\mathbb{L}(B|x) = 0) \vee (\mathbf{X} \bowtie \{x\} = \emptyset)$$

When a probability distribution is interpreted as a Markov kernel, both of these definitions agree.

Theorem 2.4.22 (Equivalence of validity definitions). *Given $\mathbf{X} : \Omega \rightarrow X$, with Ω and X standard measurable, a probability measure $\mathbb{P}^{\mathbf{X}} \in \Delta(X)$ is valid if and only if the conditional $\mathbb{P}^{\mathbf{X}|\ast} := \ast \mapsto \mathbb{P}^{\mathbf{X}}$ is valid.*

Proof. $\ast \bowtie \ast = \Omega$ necessarily. Thus validity of $\mathbb{P}^{\mathbf{X}|\ast}$ means

$$\forall A \in \mathcal{X} : \mathbf{X} \bowtie A = \emptyset \implies \mathbb{P}^{\mathbf{X}|\ast}(A|\ast) = 0$$

But $\mathbb{P}^{\mathbf{X}|\ast}(A|\ast) = \mathbb{P}^{\mathbf{X}}(A)$ by definition, so this is equivalent to

$$\forall A \in \mathcal{X} : \mathbf{X} \bowtie A = \emptyset \implies \mathbb{P}^{\mathbf{X}}(A) = 0$$

□

Conditionals can be used to define *maximal probability sets*, which is the set of all probability distributions with those conditionals.

Definition 2.4.23 (Maximal probability set). Given (Ω, \mathcal{F}) , $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$ and a $Y|X$ -valid conditional probability $\mathbb{L} : X \rightarrow Y$ the maximal probability set \mathbb{P}_C associated with \mathbb{L} is the probability set such that for all $\mathbb{P}_\alpha \in \mathbb{P}_C$, \mathbb{L} is a version of $\mathbb{P}_\alpha^{Y|X}$.

Theorem 2.4.24 shows that the semidirect product of any pair of valid conditional probabilities is itself a valid conditional. Suppose we have some collection of $X_i|X_{[i-1]}$ -valid conditionals $\{\mathbb{P}_i^{X_i|X_{[i-1]}} | i \in [n]\}$; then recursively taking the semidirect product $\mathbb{M} := \mathbb{P}_1^{X_1} \odot (\mathbb{P}_2^{X_2|X_1} \odot \dots)$ yields a $X_{[n]}$ valid distribution. Furthermore, the maximal probability set associated with \mathbb{M} is nonempty.

Collections of recursive conditional probabilities often arise in causal modelling – in particular, they are the foundation of the structural equation modelling approach Richardson and Robins (2013); Pearl (2009).

Note that validity is not a necessary condition for a conditional to define a non-empty probability set. Given some $\mathbb{K} : X \rightarrow Y$, \mathbb{K} might be an invalid conditional on if every value of X is considered, but it might be valid on some subset of X . A marginal of X that assigns measure 0 to the subset of X where \mathbb{K} is invalid can still define a valid distribution when combined with \mathbb{K} . On the other hand, if \mathbb{K} is required to combine with arbitrary valid marginals of X , then the validity of \mathbb{K} is necessary (Theorem 2.4.26).

Theorem 2.4.24 (Semidirect product of valid conditional distributions is valid). *Given (Ω, \mathcal{F}) , $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$, $Z : \Omega \rightarrow Z$ (all spaces standard measurable) and any valid candidate conditional $\mathbb{P}^{Y|X}$ and $\mathbb{Q}^{Z|YX}$, $\mathbb{P}^{Y|X} \odot \mathbb{Q}^{Z|YX}$ is also a valid candidate conditional.*

Proof. Let $\mathbb{R}^{YZ|X} := \mathbb{P}^{Y|X} \odot \mathbb{Q}^{Z|YX}$.

We only need to check validity for each $x \in X(\Omega)$, as it is automatically satisfied for other values of X .

For all $x \in X(\Omega)$, $B \in \mathcal{Y}$ such that $X \bowtie \{x\} \cap Y \bowtie B = \emptyset$, $\mathbb{P}^{Y|X}(B|x) = 0$ by validity. Thus for arbitrary $C \in \mathcal{Z}$

$$\begin{aligned} \mathbb{R}^{YZ|X}(B \times C|x) &= \int_B \mathbb{Q}^{Z|YX}(C|y, x) \mathbb{P}^{Y|X}(dy|x) \\ &\leq \mathbb{P}^{Y|X}(B|x) \\ &= 0 \end{aligned}$$

For all $\{x\} \times B$ such that $X \bowtie \{x\} \cap Y \bowtie B \neq \emptyset$ and $C \in \mathcal{Z}$ such that $(X, Y, Z) \bowtie \{x\} \times B \times C = \emptyset$, $\mathbb{Q}^{Z|YX}(C|y, x) = 0$ for all $y \in B$ by validity. Thus:

$$\begin{aligned} \mathbb{R}^{YZ|X}(B \times C|x) &= \int_B \mathbb{Q}^{Z|YX}(C|y, x) \mathbb{P}^{Y|X}(dy|x) \\ &= 0 \end{aligned}$$

□

Corollary 2.4.25 (Valid conditionals are validly extendable to valid distributions). *Given Ω , $U : \Omega \rightarrow U$, $W : \Omega \rightarrow W$ and a valid conditional $\mathbb{T}^{W|U}$, then for any valid conditional \mathbb{V}^U , $\mathbb{V}^U \odot \mathbb{T}^{W|U}$ is a valid probability.*

Proof. Applying Lemma 2.4.24 choosing $X = *$, $Y = U$, $Z = W$ and $\mathbb{P}^{Y|X} = \mathbb{V}^{U|*}$ and $\mathbb{Q}^{Z|YX} = \mathbb{T}^{W|U*}$ we have $\mathbb{R}^{WU|*} := \mathbb{V}^{U|*} \odot \mathbb{T}^{W|U*}$ is a valid conditional probability. Then $\mathbb{R}^{WU} \cong \mathbb{R}^{WU|*}$ is valid by Theorem 2.4.22. \square

Theorem 2.4.26 (Validity of conditional probabilities). *Suppose we have Ω , $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$, with Ω , X , Y discrete. A conditional $\mathbb{T}^{Y|X}$ is valid if and only if for all valid distributions \mathbb{V}^X , $\mathbb{V}^X \odot \mathbb{T}^{Y|X}$ is also a valid distribution.*

Proof. If: this follows directly from Corollary 2.4.25.

Only if: suppose $\mathbb{T}^{Y|X}$ is invalid. Then there is some $x \in X$, $y \in Y$ such that $X \bowtie (x) \neq \emptyset$, $(X, Y) \bowtie (x, y) = \emptyset$ and $\mathbb{T}^{Y|X}(y|x) > 0$. Choose \mathbb{V}^X such that $\mathbb{V}^X(\{x\}) = 1$; this is possible due to standard measurability and valid due to $X^{-1}(x) \neq \emptyset$. Then

$$\begin{aligned} (\mathbb{V}^X \odot \mathbb{T}^{Y|X})(x, y) &= \mathbb{T}^{Y|X}(y|x) \mathbb{V}^X(x) \\ &= \mathbb{T}^{Y|X}(y|x) \\ &> 0 \end{aligned}$$

Hence $\mathbb{V}^X \odot \mathbb{T}^{Y|X}$ is invalid. \square

Example 2.4.27. Body mass index is defined as a person's weight divided by the square of their height. Suppose we have a measurement process $\mathcal{S} = (W, \mathcal{H})$ and $\mathcal{B} = \frac{W}{\mathcal{H}^2}$ - i.e. we figure out someone's body mass index first by measuring both their height and weight, and then passing the result through a function that divides the second by the square of the first. Thus, given the random variables W, H modelling W, \mathcal{H} , \mathcal{B} is the function given by $B = \frac{W}{H^2}$.

With this background, suppose we postulate a decision model in which body mass index can be directly controlled by a variable C , while height and weight are not. Specifically, we have a probability set \mathbb{P}_{\square} with

$$\mathbb{P}_{\square}^{B|WHC} = \begin{array}{c} H \text{ --- } * \\ C \text{ ----- } B \\ W \text{ --- } * \end{array} \quad (2.5)$$

Then pick some $w, h, x \in \mathbb{R}$ such that $\frac{w}{h^2} \neq x$ and $(W, H) \bowtie (w, h) \neq \emptyset$ (which is to say, our measurement procedure could potentially yield (w, h) for a person's height and weight). We have $\mathbb{P}_{\square}^{B|WHC}(\{x\}|w, h, x) = 1$, but

$$\begin{aligned} (B, W, H) \bowtie \{(x, w, h)\} &= \{\omega | (W, H)(\omega) = (w, h), B(\omega) = \frac{w}{h^2}\} \\ &= \emptyset \end{aligned}$$

so $\mathbb{P}_{\square}^{B|WHC}$ is invalid. Thus there is some valid μ^{WHC} such that the probability set $\mathbb{P}_{\square}^{B|WHC} = \mu^{WHC} \odot \mathbb{P}_{\square}^{Y|X}$ is empty.

Validity rules out conditional probabilities like (2.5). We conjecture that in many cases this condition is implicitly taken into account – it is obviously silly to posit a model in which body mass index can be controlled independently of height and weight. We note, however, that presuming the authors intended their model to be interpreted according to the usual semantics of causal Bayesian networks, the invalid conditional probability (2.5) would be used to evaluate the causal effect of body mass index in the causal diagram found in Shahar (2009).

2.4.5 Existence of conditional probabilities

Lemma 2.4.28 (Conditional pushforward). *Suppose we have a sample space (Ω, \mathcal{F}) , variables $X : \Omega \rightarrow X$ and $Y : \Omega \rightarrow Y$, $Z : \Omega \rightarrow Z$ and a probability set \mathbb{P}_C with conditional $\mathbb{P}_C^{X|Y}$ such that $Z = f \circ Y$ for some $f : Y \rightarrow Z$. Then there exists a conditional probability $\mathbb{P}_C^{Z|X} = \mathbb{P}_C^{Y|X} \mathbb{F}_f$.*

Proof. Note that $(X, Z) = (\text{id}_X \otimes f) \circ (X, Y)$. Thus, by Lemma 2.2.22, for any $\mathbb{P}_\alpha \in \mathbb{P}_C$

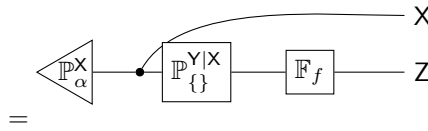
$$\mathbb{P}_\alpha^{XZ} = \mathbb{P}_\alpha^{XY} \mathbb{F}_{\text{id}_X \otimes f}$$

Note also that for all $A \in \mathcal{X}$, $B \in \mathcal{Z}$, $x \in X$, $y \in Y$:

$$\begin{aligned} \mathbb{F}_{\text{id}_X \otimes f}(A \times B|x, y) &= \delta_x(A) \delta_{f(y)}(B) \\ &= \mathbb{F}_{\text{id}_X}(A|x) \otimes \mathbb{F}_f(B|y) \\ \implies \mathbb{F}_{\text{id}_X \otimes f} &= \mathbb{F}_{\text{id}_X} \otimes \mathbb{F}_f \end{aligned}$$

Thus

$$\mathbb{P}_\alpha^{XZ} = (\mathbb{P}_\alpha^X \odot \mathbb{P}_C^{Y|X}) \mathbb{F}_{\text{id}_X} \otimes \mathbb{F}_f$$



Which implies $\mathbb{P}_C^{Y|X} \mathbb{F}_f$ is a version of $\mathbb{P}_\alpha^{Z|X}$. Because this holds for all α , it is therefore also a version of $\mathbb{P}_C^{Z|X}$. \square

The following theorem is a standard result in many probability texts. In this work, the measurable spaces considered will all be standard measurable and so Theorem 2.4.29 always applies. We will simply assume that conditional probabilities exist, and avoid referencing this theorem every time.

Theorem 2.4.29 (Existence of regular conditionals). *Suppose we have a sample space (Ω, \mathcal{F}) , variables $X : \Omega \rightarrow X$ and $Y : \Omega \rightarrow Y$ with Y standard measurable and a probability model \mathbb{P}_α on (Ω, \mathcal{F}) . Then there exists a conditional $\mathbb{P}_\alpha^{Y|X}$.*

Proof. Çinlar (2011, Theorem 2.18) □

The following theorem was proved by Bogachev and Malofeev (2020).

Theorem 2.4.30. *Given a Borel measurable map $m : X \rightarrow Y \times Z$ let $\Pi_Y : Y \times Z \rightarrow Y$ be the projection onto Y . Then there exists a Borel measurable map $n : X \times Y \rightarrow Y \times Z$ such that*

$$n(\Pi_Y^{-1}(y)|x, y) = 1 \quad (2.6)$$

$$m(Y^{-1}(A) \cap B|x) = \int_A n(B|x, y) m\mathbb{F}_{\Pi_Y}(dy|x) \quad \forall A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z} \quad (2.7)$$

Proof. Bogachev and Malofeev (2020, Theorem 3.5) □

The following corollary implies that, given a uniform conditional, higher order conditionals can generically be found for probability sets.

Corollary 2.4.31 (Existence of higher order conditionals with respect to probability sets). *Take a sample space (Ω, \mathcal{F}) , variables $X : \Omega \rightarrow X$ and $Y : \Omega \rightarrow Y$, $Z : \Omega \rightarrow Z$ and a probability set \mathbb{P}_C with uniform conditional distribution $\mathbb{P}_C^{YZ|X}$, and Y and Z standard measurable. Then there exists a higher order uniform conditional $\mathbb{P}_C^{Z|(Y|X)}$.*

Proof. Take $\mathbb{P}_C^{YZ|X}$ to be the Borel measurable map m from Theorem 2.4.30, and note that $\Pi_Y \circ (Y, Z) = Y$. Then equation (2.7) implies for all $A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z}$ there is some n such that

$$\begin{aligned} \mathbb{P}_C^{YZ|X}(Y^{-1}(A) \cap B|x) &= \int_A n(B|x, y) \mathbb{P}_C^{YZ|X} \mathbb{F}_{\Pi_Y}(dy|x) \\ &= \int_A n(B|x, y) \mathbb{P}_C^{Y|X}(dy|x) \end{aligned} \quad (2.8)$$

where Equation (2.8) follows from Lemma 2.4.28.

Then, for any $\mathbb{P}_\alpha \in \mathbb{P}_C$

$$\mathbb{P}_C^{YZ|X}(Y^{-1}(A) \cap B|x) = \int_A n(B|x, y) \mathbb{P}_\alpha^{Y|X}(dy|x)$$

which implies n is a version of $\mathbb{P}_C^{YZ|(Y|X)}$. By Lemma 2.4.28, $n\mathbb{F}_{\Pi_Y}$ is a version of $\mathbb{P}_C^{Z|(Y|X)}$. □

We might be motivated to ask whether the higher order conditionals in Theorem 2.4.31 can be chosen to be valid. Despite Lemma 2.4.32 showing that the existence of proper conditional probabilities implies the existence of valid ones, we cannot make use of this in the above theorem because Equation (2.6) makes n proper with respect to the “wrong” sample space $(Y \times Z, \mathcal{Y} \otimes \mathcal{Z})$ while what we would need is a proper conditional probability with respect to (Ω, \mathcal{F}) .

We can choose higher order conditionals to be valid in the case of discrete sets, and whether we can choose them to be valid in more general measurable spaces is an open question.

Lemma 2.4.32. *Given a probability space $(\mu, \Omega, \mathcal{F})$ and variables $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$, if there is a regular proper conditional probability $\mu^{|X} : X \rightarrow \Omega$ then there is a valid conditional distribution $\mu^{Y|X}$.*

Proof. Take $\mathbb{K} = \mu^{|X} \mathbb{F}_Y$. We will show that \mathbb{K} is valid, and a version of $\mu^{Y|X}$.

Defining $O := \text{id}_\Omega$ (the identity function $\Omega \rightarrow \Omega$), $\mu^{|X}$ is a version of $\mu^{O|X}$. Note also that $Y = Y \circ O$. Thus by Lemma 2.4.28, \mathbb{K} is a version of $\mu^{Y|X}$.

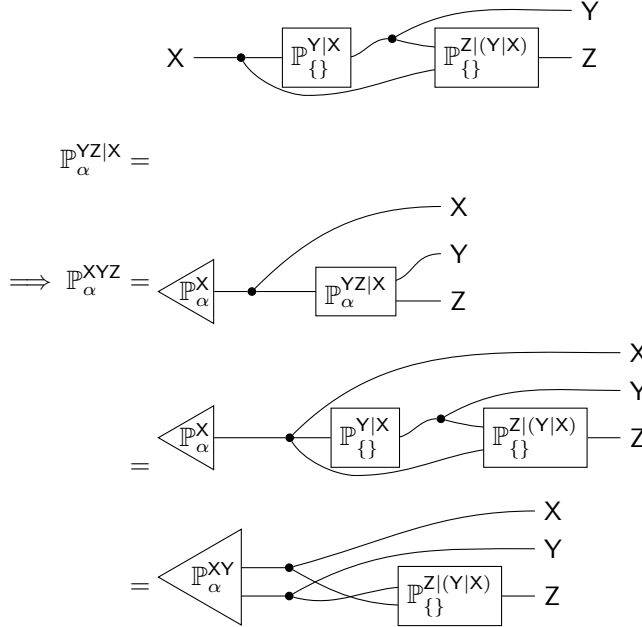
It remains to be shown that \mathbb{K} is valid. Consider some $x \in X$, $A \in \mathcal{Y}$ such that $X^{-1}(\{x\}) \cap Y^{-1}(A) = \emptyset$. Then by the assumption $\mu^{|X}$ is proper

$$\begin{aligned} \mathbb{K}(Y \bowtie A | x) &= \delta_x(Y^{-1}(A)) \\ &= 0 \end{aligned}$$

Thus \mathbb{K} is valid. □

Theorem 2.4.33 (Higher order conditionals). *Suppose we have a sample space (Ω, \mathcal{F}) , variables $X : \Omega \rightarrow X$ and $Y : \Omega \rightarrow Y$, $Z : \Omega \rightarrow Z$ and a probability set \mathbb{P}_C with conditional $\mathbb{P}_C^{YZ|X}$. Then $\mathbb{P}_C^{Z|(Y|X)}$ is a version of $\mathbb{P}_C^{Z|YX}$.*

Proof. For arbitrary $\mathbb{P}_\alpha \in \mathbb{P}_C$



Thus $\mathbb{P}_C^{Z|(Y|X)}$ is a version of $\mathbb{P}_\alpha^{Z|YX}$ for all α and hence also a version of $\mathbb{P}_C^{Z|YX}$. □

Theorem 2.4.34. *Given probability gap model \mathbb{P}_C , X, Y, Z such that $\mathbb{P}_C^{Z|YX}$ exists, $\mathbb{P}_C^{Z|Y}$ exists iff $Z \perp\!\!\!\perp_{\mathbb{P}_C} X|Y$.*

Proof. If: If $Z \perp\!\!\!\perp_{\mathbb{P}_C} X|Y$ then by Theorem 2.4.12, for each $\mathbb{P}_\alpha \in \mathbb{P}_C$ there exists $\mathbb{P}_\alpha^{Z|Y}$ such that

$$\mathbb{P}_\alpha^{Y|WX} = \begin{array}{ccc} W & \xrightarrow{\quad \boxed{\mathbb{K}} \quad} & Y \\ X & \xrightarrow{\quad * \quad} & \end{array}$$

□

Theorem 2.4.35 (Valid higher order conditionals). *Suppose we have a sample space (Ω, \mathcal{F}) , variables $X : \Omega \rightarrow X$ and $Y : \Omega \rightarrow Y$, $Z : \Omega \rightarrow Z$ and a probability set \mathbb{P}_C with regular conditional $\mathbb{P}_C^{YZ|X}$, Y discrete and Z standard measurable. Then there exists a valid regular $\mathbb{P}_C^{Z|XY}$.*

Proof. By Theorem 2.4.31, we have a higher order conditional $\mathbb{P}_C^{Z|(Y|X)}$ which, by Theorem 2.4.33 is also a version of $\mathbb{P}_C^{Z|XY}$.

We will show that there is a Markov kernel \mathbb{Q} almost surely equal to $\mathbb{P}_C^{Z|XY}$ which is also valid. For all $x, y \in X \times Y$, $A \in \mathcal{Z}$ such that $(X, Y, Z) \bowtie \{(x, y)\} \times A = \emptyset$, let $\mathbb{Q}(A|x, y) = \mathbb{P}_C^{Z|XY}(A|x, y)$.

By validity of $\mathbb{P}_C^{YZ|X}$, $x \in X(\Omega)$ and $(X, Y, Z) \bowtie \{(x, y)\} \times A = \emptyset$ implies $\mathbb{P}_C^{YZ|X}(\{y\} \times A|x) = 0$. Thus we need to show

$$\forall A \in \mathcal{Z}, x \in X, y \in Y : \mathbb{P}_C^{YZ|X}(\{y\} \times A|x) = 0 \implies (\mathbb{Q}(A|x, y) = 0) \vee ((X, Y) \bowtie \{(x, y)\} = \emptyset)$$

For all x, y such that $\mathbb{P}_\square^{Y|X}(\{y\}|x)$ is positive, we have $\mathbb{P}^{YZ|X}(\{y\} \times A|x) = 0 \implies \mathbb{P}_\square^{Z|XY}(A|x, y) = 0 =: \mathbb{Q}(A|x, y)$.

Furthermore, where $\mathbb{P}_\square^{Y|X}(\{y\}|x) = 0$, we either have $(X, Y, Z) \bowtie \{(x, y)\} \times A = \emptyset$ or can choose some $\omega \in (X, Y, Z) \bowtie \{(x, y)\} \times A$ and let $\mathbb{Q}(Z(\omega)|x, y) = 1$. This is an arbitrary choice, and may differ from the original $\mathbb{P}_C^{Z|XY}$. However, because Y is discrete the union of all points y where $\mathbb{P}_\square^{Y|X}(\{y\}|x) = 0$ is a measure zero set, and so \mathbb{Q} differs from $\mathbb{P}_\square^{Y|X}$ on a measure zero set. □

Chapter 3

Models with choices and consequences

Probability sets, introduced in Chapter 2, will be used to model *decision problems*, which are problems that involve choices and consequences. In such problems, three things are given: a set of options (one of which must be chosen), a set of consequences and a means of judging which consequences are more desirable than others. Such a problem requires an understanding of how each choice corresponds to consequences, as far as this is able to be understood. The fundamental type of problem studied in this thesis is how to map choices to consequences.

In practice, causal inference is concerned with a wider variety of problems than this. A great deal of empirical causal analysis is concerned with problems a step removed from this: the purpose is to advise other decision makers on a course of action rather than to recommend an action directly. Nevertheless, a great deal of causal analysis is ultimately motivated by problems involving a choice among options, even if the analysis only addresses such problems indirectly. Section 3.1 briefly reviews the attitude of prominent theorists of causal inference towards decision problems. Subsequently, it presents the basic definition of a decision problem, and two different kinds of models that can be used to represent the relationship between choices and consequences.

In our approach, a decision maker facing a decision problem must select one choice from a set of candidates. We assume that they use a *model* to help them make this selection, which associates with each choice a probability distribution over a set of possible consequences, and a *utility* that enables them to compare the desirability of different consequences. The reasons we consider this setup are several: firstly, the idea that we make decisions by comparing the consequences of different choices is intuitively appealing, and probability is a well-understood and widely used theory for representing uncertain prospects. Furthermore, many theories of decision making that aim for more rigorous foundations for formal decision making arrive at models that map choices to distributions over consequences, sometimes along with some additional structure.

Chapter 5 shows how many causal inference frameworks also induce functions from some underlying set to probability distributions, even though they are not at face value theories of decision making, and the underlying sets are not necessarily identified with possible choices.

We aren't trying to claim that this is the only possible way to formalise decision making. This chapter examines some key decision theories that justify their modelling choices by suggesting axioms for rational theories of decision under uncertainty. However, despite the various attempts at axiomatisation, the nature of theories of “rational choice” is contested – there is no clear standard among the theories surveyed here, or developed elsewhere. This work is not trying to resolve this dispute, yet modelling choices must still be made. Section 3.2 provides an overview of four major decision theories along with their axiomatisations (where applicable). These are *Savage decision theory*, *Jeffrey decision theory* (or evidential decision theory), Lewis' *causal decision theory* and *statistical decision theory*.

Section 3.2 describes in particular detail the connections between *statistical decision theory* (Wald, 1950) and probability set models of decision problems. We are able to demonstrate a close connection between probability set models of decision problems and the classical statistical notion of *risk* of a decision rule, even though causal considerations are often not central to classical statistics. Secondly, the kind of probability set model – which we call a *see-do model* – shows up again in Chapter 4 where we consider the question of when a probability set model supports a certain notion of “the causal effect of a variable”, and again in Chapter 5 where we consider the kinds of probability set models induced by other causal reasoning frameworks.

The formal definition of a variable in a probabilistic model is well known (Definition 2.2.13). However, in practice the definitions of variables often includes informal content that enables the interpretation of a probabilistic model. In the field of causal models, one is likely to come across many different “kinds” of variables: for example, observed variables, unobserved variables, counterfactual variables and causal variables all play important roles in various causal inference frameworks. However, there is no formal distinction between these different kinds of variables – Definition 2.2.13 applies to them all. Section 3.3 is an attempt to clarify an understanding of the informal role of variables as “pointing to the parts of the world that the model is about”. In comparison to the wide variety of variable types encountered in the causal literature, it offers a very limited theory of the informal semantics of variables. In short, observed variables correspond to a measurement procedure (in a sense that will be made precise), and unobserved variables do not.

3.1 What is the point of causal inference?

Pearl and Mackenzie (2018) argue that causal reasoning frameworks should be understood by the kinds of questions that they may be able to answer. They classify causal questions into three types, which they claim form a hierarchy or

a “ladder”. That is: questions of type m are also questions of type n for $m < n$. The question types are (Bareinboim et al., 2020):

1. *Associational*: “questions about relationships and predictions”; formally defined as queries that can be answered by a single probability distribution
2. *Interventional*: “questions about the consequences of interventions”; formally defined as queries that can be answered by a causal Bayesian network (CBN)
3. *Counterfactual*: “questions concerning imagining alternate worlds”; formally defined as queries that can be answered by a structural causal model (SCM)

Models that address decision problems are concerned primarily with consequences of choices, which seems to place them at the second level of this ladder. Given that this thesis is concerned with foundational questions in causal inference and that counterfactual questions are, according to this ladder, a more general kind of causal question, one might ask why this thesis only focuses on questions at level 2.

There are a few reasons for focusing on level 2 questions as the primary motivation for a theory of causal inference. First, decision problems are a particularly important subset of causal inference problems. Within the causal inference literature, “interventional” questions and interpretations are much more prominent than strictly counterfactual questions. For example, Rubin (2005) points out that causal inference often informs a decision maker by providing “scientific knowledge”, but does not make recommendations by itself. (Imbens and Rubin, 2015) introduces causal inference as the study of “outcomes of manipulations” and (Spirtes et al., 2000) highlights the universal relevance of understanding how to control certain outcomes, while further arguing that clarifying commonsense ideas of causation is also an important aim of causal inference. Hernán and Robins (2020) present causal knowledge as critical for assessing the consequences of actions. Second, sometimes we want to justify a technical choice by appealing to features of the problem the theory is supposed to solve, and this is much easier for me to do with decision problems – for which I have strong intuitions – than strictly counterfactual questions, where my intuitions are generally much less clear. Third, as discussed in Chapter 1 and will be further discussed in Chapter 5, a key feature of causal models is the fact that they come with a set of “possibilities under consideration”. These possibilities might be interventions or counterfactual proposals, and they may be explicit or implicit. If the set of possibilities is difficult to ascertain, then the causal model becomes difficult to understand. In decision problems specifically, the set of possible choices is a natural candidate for this set of “possibilities under consideration” – as we understand them, given a decision maker facing a decision problem, there is a set of possible choices that the decision maker may ultimately select. This same set is the set of things that the decision maker wants to evaluate with regard to their likely consequences, which is to say it is

the set of possibilities under consideration. If we do not suppose that the causal problem is ultimately embedded in a decision problem, it is not at all obvious to us where else the “set of possibilities” could come from.

Speculatively, counterfactual queries may also be able to be interpreted as decision problems with fanciful options. Consider an informal decision problem and a counterfactual query addressing similar material:

- Decision problem: I want my headache to go away. If I take Aspirin, will it do so?
- Counterfactual query: I wish I didn’t have headache. If I had taken the Aspirin, would I still have it?

If I haven’t taken aspirin, then there’s nothing I can actually choose to do to make it so that I had. However, if I imagine that I did have some option available that accomplished this, then the structure of the two questions seems rather similar. Both ask: if I take the option, what will the consequence be? Of course, it’s hard to say what makes a correct answer to the second question, but this is a feature of counterfactual questions in general.

3.1.1 Modelling decision problems

People who need to make a decisions might (and often do) make them with no mathematical reasoning at all. However, this work is concerned with making decisions assisted by mathematical reasoning. In order to reason mathematically about a decision to be made, we assume that somehow, we have access to two sets:

1. There is a set of choices C that need to be compared
2. There is a set of consequences Ω along with a utility function $u : \Omega \rightarrow \mathbb{R}$, which measures the goodness of each $\omega \in \Omega$

Given some means of relating between C and Ω , the order on Ω induced by u will induce some order on C . There are a great number of different ways that of relating elements of C to elements of Ω . For example, a binary relation between the two sets will, given a total order on Ω , induce a preorder on C . However, in this work the assumption is made that the relevant kinds of relations are either

- A Markov kernel $C \rightarrow \Omega$
- A Markov kernel $C \times H \rightarrow \Omega$ for some set of hypotheses H

That is, for each choice $c \in C$ we have either a probability distribution in $\Delta(\Omega)$ or a set of probability distributions indexed by $h \in H$. Sections 3.2.5 and 3.2.5 discuss each choice in more detail. Where it is needed, we also assume that a utility function $u : \Omega \rightarrow \mathbb{R}$ is available and that choices are evaluated using the principle of expected utility.

Usually, someone confronted with a decision problem will not know for certain the consequences that arise from any given choice, and yet they may have

some views about which consequences are more likely than others. Probability has a long and successful history of representing uncertain knowledge of this type. There are many works that aim to show that any method for representing uncertain knowledge that adheres to certain principles must be a probability distribution de Finetti ([1937] 1992); Horvitz et al. (1986), along with criticism of these principles Halpern (1999). A notable alternative to representing uncertainty with a single probability distribution represents uncertainty with a set of probability distributions, which is a type of *vague probability* model (Walley, 1991).

More relevant to the question of modelling decision problems are a number of works that establish conditions under which “desirability” or “preference” relations over sets of choices or propositions must be represented by a probability distribution along with a utility function. These works are surveyed in Section 3.2. Ultimately, however, the question of whether probability is the right choice to represent uncertain knowledge in decision models is not a key focus of this work. It is a conventional choice, and one that is accepted here.

3.1.2 Formal definitions

As mentioned above, we suppose that we are given a few basic ingredients: a set of choices C equipped with an algebra \mathcal{C} , a set of consequences Ω with an algebra of events \mathcal{F} and a utility function $u : \Omega \rightarrow \mathbb{R}$. We call these ingredients a “decision problem”.

Definition 3.1.1 (Decision problem). A decision problem is a triple (C, Ω, u) consisting of a measurable set (C, \mathcal{C}) of choices, (Ω, \mathcal{F}) consequences and a utility function $u : \Omega \rightarrow \mathbb{R}$.

Our task is to find a *model* that relates choice C to consequences Ω . We assume two forms of model – a *choices only model* associates each choice with a unique probability distribution, and a *choices and hypotheses model* associates each choice with a set of probability distributions. A model consisting of a single probability distribution is also sometimes called a “Bayesian” model, so we could call choices only models “Bayesian” and choices and hypotheses models “non-Bayesian”.

Definition 3.1.2 (Choices only model). Given a decision problem (C, Ω, u) , a *choices only model* is a function $C \rightarrow \Omega$.

Definition 3.1.3 (Choices and hypotheses model). Given a decision problem (C, Ω, u) , a model with *choices and hypotheses* is a function $C \times H \rightarrow \Omega$ for some hypothesis set H .

By convention, we use \mathbb{P} with the subscript \cdot to denote a model, subscripts \mathbb{P}_α where α is an element of the domain to refer to the model evaluated at α , and the subscript $\mathbb{P}_{C \times H}$ to refer to the image of the model.

Nxample 3.1.4 (Model). Given a decision problem (C, Ω, u) and a model $\mathbb{P} : C \times H \rightarrow \Omega$, $\mathbb{P}_\alpha := \mathbb{P}(\alpha)$.

Nxample 3.1.5 (Image of a model). Given a decision problem (C, Ω, u) and a model $\mathbb{P} : C \times H \rightarrow \Omega$, the $\mathbb{P}_{C \times H} := \{\mathbb{P}_\alpha | \alpha \in C \times H\}$ is the image of the model.

3.2 Theories of decision making

The question of how decision problems ought to be represented has received substantial attention. We survey a number of key theories from this literature, and point out connections with our scheme:

- Every theory surveyed proposes that choices are evaluated by way of a probabilistic map from choices to consequences, along with some measure of the desirability of consequences
- Most theories have some analogue of hypotheses (Definition 3.1.3, see also Chapter 4)
- Most theories have some notion of a “prior” over hypotheses, which induces a choice only model (Definition 3.1.2)

Statistical Decision Theory (SDT), introduced by Wald (1950), further proves a *complete class theorem*, which shows that, under some conditions, choices that are admissible (Definition 3.2.23) are also optimal with respect to some prior over hypotheses. That is, any admissible decision under a choices and hypotheses model can be rationalised as a decision under a choices only model with some prior (though, importantly, this *doesn't* establish that proposing a prior is always the appropriate way to go about making a decision). We show that SDT corresponds to a particular class of models we call conditionally independent see-do models (Definition 3.2.15) combined with the principle of expected utility maximisation, and that the complete class theorem to a broader class of see-do models.

The following discussion will often make reference to *complete preference relations*. A complete preference relation is a relation \succ, \prec, \sim on a set A such that for any a, b, c in A we have:

- Exactly one of $a \succ b$, $a \prec b$, $a \sim b$ holds
- $(a \succ b) \iff (b \prec a)$
- $a \succ b$ and $b \succ c$ implies $a \succ c$

In short, it is a total order without antisymmetry (a and b can be equally preferred even if they are not in fact equal).

This definition is meant to correspond to the common sense idea of having preferences over some set of things, where \succ can be read as “strictly better than”, \prec read as “strictly worse than” and \sim read as “as good as”. Given any two things from the set, I can say which one I prefer, or if I prefer neither (and all of these are mutually exclusive). If I prefer a to a' then I think a' is worse than a . Furthermore, if I prefer a to a' and a' to a'' then I prefer a to a'' .

Define $a \preceq b$ to mean $a \prec b$ or $a \sim b$.

3.2.1 von Neumann-Morgenstern utility

Von Neumann and Morgenstern (1944) (henceforth abbreviated to vNM) proved that when the *vNM axioms* hold (not defined here; see the original reference or Steele and Stefánsson (2020)), an agent’s preferences between “lotteries” (probability distributions in $\Delta(\Omega)$ for some (Ω, \mathcal{F})) can be represented as the comparison of the expected value under each lottery of a utility function u unique up to affine transformation. That is, for lotteries \mathbb{P}_α and $\mathbb{P}_{\alpha'}$, there exists some $u : \Omega \rightarrow \mathbb{R}$ unique up to affine transformation such that $\mathbb{E}_{\mathbb{P}_\alpha}[u] > \mathbb{E}_{\mathbb{P}_{\alpha'}}[u]$ if and only if $\mathbb{P}_\alpha \succ \mathbb{P}_{\alpha'}$.

In vNM theory, the set of lotteries is the set of all probability measures on (Ω, \mathcal{F}) . Thus von Neumann-Morgenstern theorem gives conditions under which preferences *over distributions of consequences* can be represented using expected utility. If a decision problem were given such that the set of available choices was in 1-to-1 correspondence with the set of probability distributions in $\Delta(\Omega)$, then the vNM theory provides conditions on the preference relation such that, if these conditions are satisfied, the preference relation can be represented by some utility function on the set of consequences. Typically, the set of choices is not in 1-to-1 correspondence with probability distributions in $\Delta(\Omega)$. Indeed, the starting point of this work is that the relation between choices and consequences is not always obvious, and this situation might be improved by a better understanding of models that relate the two.

3.2.2 Savage decision theory

Savage’s decision theory distinguishes *acts* C , *consequences* Ω and *states* (S, \mathcal{S}) (Savage, 1954). In our framework, acts are similar to choices, consequences to consequences and states are similar to hypotheses. Unlike vNM theory, the mapping from acts to consequences is not assumed to be given at the outset. Instead, each act is assumed to induce a known mapping from each state to an element of the set of consequences. His theorem conditions under which, given such a map from acts and states to consequences, a preference relation over acts can be represented by a “prior” over states and a utility function $u : \Omega \rightarrow \mathbb{R}$ in combination with the principle of expected utility. As Theorem 3.2.3 shows, the prior over states induces a probabilistic map from choices to consequences that, in combination with the utility, is sufficient to evaluate the desirability of the choices.

We have said that acts are similar to choices and states are similar to hypotheses in our framework – but there are differences. We’ve taken the set of choices to be the set of all the things that the decision maker might choose once they’ve finished considering their problem, *and* the way they make this selection is to compare each choice on the basis of the consequences it is expected to bring about. In Savage’s theory, like ours, the decision maker has a preference relation over the set of acts. Unlike our theory, however, the set of acts is precisely the set of all functions from states to consequences. That is, in contrast to our “choices and hypotheses models” (Definition 3.1.3) the map from states and

acts to consequences is deterministic where we take the map from choices and hypotheses to consequences to be stochastic, and the set of acts is assumed to contain every function from states to consequences.

This could be considered a requirement of extendability: given a choices and hypotheses model $\mathbb{P} : C \times H \rightarrow \Omega$, we might consider the model a Savage decision model if the set of choices can be extended to the convex closure of the set of all deterministic functions $H \rightarrow \Omega$ such that the Savage axioms (Appendix 3.5.1) hold. The reason why Savage's theory has such a rich set of choices is the need to go from a preference relation over choices to a preference relation over consequences. All a decision maker actually needs is the ordering on the choices that they're actually considering, and this might be compatible with many orderings of consequences. We don't know if there are cases of decision problems where this extendability requirement introduces difficulties.

Definition 3.2.1 (Elements of a Savage decision problem). A *Savage decision problem* features a measurable set of states (S, \mathcal{S}) , a set of consequences (Ω, \mathcal{F}) and a set of acts C such that $|C| = \Omega^S$ and a measurable evaluation function $T : S \times C \rightarrow \Omega$ such that for any $f : S \rightarrow \Omega$ there exists $c \in C$ such that $T(\cdot, c) = f$.

Theorem 3.2.2 is Savage's representation theorem. The Savage axioms aren't investigated in detail in this work, but for the reader's convenience they're given in Appendix 3.5.1.

Theorem 3.2.2. *Given any Savage decision problem (S, Ω, C, T) with a preference relation (\prec, \sim) on C that satisfies the Savage axioms, there exists a unique probability distribution $\mu \in \Delta(S)$ and a utility $u : \Omega \rightarrow \mathbb{R}$ unique up to affine transformation such that*

$$\alpha \preceq \alpha' \iff \int_S u(T(s, \alpha)) \mu(ds) \leq \int_S u(T(s, \alpha')) \mu(ds) \quad \forall \alpha, \alpha' \in C$$

Proof. Savage (1954) □

Savage's setup implies the existence of a unique probabilistic function $C \rightarrow \Delta(\Omega)$ representing the “probabilistic consequences” of each choice.

Theorem 3.2.3. *Given any Savage decision problem (S, Ω, C, T) with a preference relation (\prec, \sim) on C that satisfies the Savage axioms, and a σ -algebra \mathcal{F} on Ω such that T is measurable, there is a probabilistic function $\mathbb{P} : C \rightarrow \Delta(\Omega)$ and a utility $u : \Omega \rightarrow \mathbb{R}$ unique up to affine transformation such that*

$$\alpha \preceq \alpha' \iff \int_{\Omega} u(f) \mathbb{P}_{\alpha}(df) \leq \int_{\Omega} u(f) \mathbb{P}_{\alpha'}(df) \quad \forall \alpha, \alpha' \in C$$

Proof. Define $\mathbb{P} : C \rightarrow \Delta(\Omega)$ by

$$\mathbb{P}_{\alpha}(A) := \mu(T_{\alpha}^{-1}(A)) \quad \forall A \in \mathcal{F}$$

where $T_\alpha : S \rightarrow F$ is the function $s \mapsto T(s, \alpha)$. \mathbb{P}_α is the pushforward of T_α under μ .

Then

$$\begin{aligned} \int_{\Omega} u(f) \mathbb{P}_\alpha(df) &= \int_S u \circ T_\alpha(s) \mu(ds) \\ &= \int_S u(T(s, \alpha)) \mu(ds) \end{aligned}$$

□

3.2.3 Jeffrey's decision theory

Jeffrey's decision theory is an alternative to Savage's that starts from a different set of assumptions. One of the key differences is in what is assumed at the outset: where Savage assumes a set of states S , acts C and consequences Ω , Jeffrey's theory only considers a single space \mathcal{F} , which is a complete atomless boolean algebra. Elements of \mathcal{F} are said to be propositions. We note that \mathcal{F} cannot be understood as the set of events with respect to a finite measurement procedure (Section 3.3). The collection of finite propositions regarding the results of some finite measurement procedure followed by flipping an infinite number of coins could perhaps be represented by a complete atomless boolean algebra. The theory is set out in Jeffrey (1965), and the key representation theorem proved in Bolker (1966).

Recall that our fundamental problem is relating a set C of things we can choose to a set F of things we can compare. Jeffrey's theory uses a different strategy to accomplish this than Savage's; where Savage identifies a set of acts C with all functions $S \rightarrow F$ and proposes axioms that constrain a preference relation on C , Jeffrey assumes that choices are elements of the algebra \mathcal{F} , accompanied by other propositions that do not correspond precisely to choices. Jeffrey's axioms pertain to a preference relation on \mathcal{F} , and preferences over choices are given by the restriction of the preference relation to C . In common with Savage's theory, the preference relation is assumed to be available over a much richer set than the set of choices actually under consideration.

Complete atomless boolean algebras are somewhat different to standard measurable σ -algebras. The σ -algebra $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a complete Boolean algebra when identifying \wedge with \cap , \vee with \cup , 0 with \emptyset and 1 with \mathbb{R} , but it has atoms: any singleton $\{x\}$ has only the subsets \emptyset and $\{x\}$. An example of a complete atomless boolean algebra can be constructed from the set of Lebesgue measurable sets on $[0, 1]$ with any two sets that differ by a set of measure zero identified Bolker (1967).

Definition 3.2.4 (Complete atomless boolean algebra). A complete atomless boolean algebra \mathcal{F} is a tuple $(A, \wedge, \vee, \cdot, 0, 1)$ such that, for all $a, b, c \in A$:

- $(a \vee b) \vee c = a \vee (b \vee c)$ and $(a \wedge b) \wedge c = a \wedge (b \wedge c)$
- $a \vee b = b \vee a$ and $a \wedge b = b \wedge a$

- $a \vee (a \wedge b) = a$ and $a \wedge (a \vee b) = a$
- $a \vee 0 = a$ and $a \wedge 1 = a$
- $a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$ and $a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$
- $a \vee \neg a = 1$ and $a \wedge \neg a = 0$

say $a \leq b$ exactly when $a \vee b = b$. A boolean algebra is atomless if for any b there is some $a \neq 0$ such that $a \leq b$. A boolean algebra is complete if for every $B \subset A$, there is some c such that c is an upper bound of B and for all upper bounds c' of B , such that $c \leq c'$.

The Bolker axioms are also not analysed deeply in this work, but for the reader's convenience they can be found in Appendix 3.5.2).

Theorem 3.2.5. *Suppose there is a complete atomless Boolean algebra \mathcal{F} with a preference relation \preceq . If \preceq satisfies the Bolker axioms then there exists a desirability function $\text{des} : \mathcal{F} \rightarrow \mathbb{R}$ and a probability distribution $\mu \in \Delta(\mathcal{F})$ such that for $A, B \in \mathcal{F}$ and finite partition $D_1, \dots, D_n \in \mathcal{F}$:*

$$(A \preceq B) \iff \sum_i^n \text{des}(D_i) \mu(D_i|A) \leq \sum_i^n \text{des}(D_i) \mu(D_i|B) \quad (3.1)$$

where $\mu(D_i|A) := \frac{\mu(A \cap D_i)}{\mu(A)}$ for $\mu(A) > 0$, undefined otherwise.

Proof. Bolker (1966)) □

As mentioned, in Jeffrey's theory the *choices* under consideration C are assumed to be some subset of \mathcal{F} . Thus we can deduce from a Jeffrey model a function $C \rightarrow \Delta(\mathcal{F})$ that “represents the consequences of choices” in the sense of Theorem 3.2.6.

Theorem 3.2.6. *Suppose there is a complete atomless Boolean algebra \mathcal{F} with a preference relation \preceq that satisfies the Bolker axioms, and a set of choices C over which a preference relation is sought with $\mu(\alpha) > 0$ for all $\alpha \in C$. Then there is a function $\mathbb{P} : C \rightarrow \Delta(\mathcal{F})$ such that for any $\alpha, \alpha' \in C$ and finite partition $D_1, \dots, D_n \in \mathcal{F}$:*

$$\alpha \preceq \alpha' \iff \sum_i^n \text{des}(D_i) \mathbb{P}_\alpha(D_i) \leq \sum_i^n \text{des}(D_i) \mathbb{P}_{\alpha'}(D_i) \quad (3.2)$$

Where μ and des are as in Theorem 3.2.5

Proof. Define \mathbb{P} by $\alpha \mapsto \mu(\cdot|\alpha)$. Then Equation (3.2) follows from Equation (3.1). □

3.2.4 Causal decision theory

Causal decision theory was developed after both Jeffrey's and Savage's theory. A number of authors Lewis (1981); Skyrms (1982) felt that Jeffrey's theory erred by treating the consequences of a choice as an "ordinary conditional probability". Lewis (1981) suggested that causal decision theory can be used to evaluate choices when we are given a set Ω of consequences over which preferences are known, a set C of choices and a set H of dependency hypotheses (the letters have been changed to match usage in this work; in the original the consequences were called S , the choices A and the dependency hypotheses H). Choices are then evaluated according to the causal decision rule. We have taken the liberty to state Lewis' rule in the language of the present work.

Definition 3.2.7 (Causal decision rule). Given a set C of choices, sample space (Ω, \mathcal{F}) , variables $H : \Omega \rightarrow H$ (the *dependency hypothesis*) and $S : \Omega \rightarrow S$ (the *consequence*) and a utility $u : \Omega \rightarrow \mathbb{R}$, the *causal utility* of a choice $\alpha \in C$ is given by

$$U(\alpha) := \int_S \int_H u(s) \mathbb{P}_\alpha^{S|H}(ds|h) \mathbb{P}_C^H(dh) \quad (3.3)$$

For some probabilistic function $\mathbb{P} : C \rightarrow \Delta(\Omega)$.

The reasons why Lewis wanted to introduce dependency hypothesis and modify Jeffrey's rule to Equation (3.3) are controversial and do not come up in this work. However, causal decision theory is still relevant to this work in two ways: firstly, once again is a probabilistic function $\mathbb{P} : C \rightarrow \Delta(\Omega)$. Secondly, causal decision theory introduces the notion of the dependency hypothesis H . The dependency hypothesis is similar to the state in Savage's theory, however Lewis does not require a deterministic map from dependency hypotheses to consequences, nor does he require a choice to correspond to every possible function from dependency hypotheses to states.

Dependency hypotheses are quite an important idea in causal reasoning. Together Lewis' decision rule connect the theory of probability sets with *statistical decision theory*, as Section 3.2.5 will show. Chapter 4 goes into considerable detail concerning the question of when probability sets support certain types of dependency hypothesis. While they are typically not explicitly represented in common frameworks for causal inference, Chapter 5 discusses how dependency hypotheses are often implicit in these approaches, and shows how they can be made explicit.

3.2.5 Statistical decision theory

Statistical decision theory (SDT), created by Wald (1950), predates all of the decision theories discussed above. Savage's theory appears to have developed in part to explain some features of SDT Savage (1951), and Jeffrey's theory and subsequent causal decision theories were in turn influenced by Savage's decision theory. While the later decision theories were concerned with articulating why

their theory fit the role of a theory for rational decision under uncertainty, Wald focused much more on the mathematical formalism and solutions to statistical problems. Statistical decision theory introduced many fundamental ideas that have since entered the “water supply” of machine learning theory, such as *decision rules* and *risk* as a measure of the quality of a decision rule.

In contrast to the later decision theories, SDT has no explicit representation of the “consequences” of a decision. Rather, it is assumed that a loss function is given that maps decisions and hypotheses directly to a loss, which is a kind of desirability score similar to a utility (although it is minimised rather than maximised). The following definitions are all standard to SDT.

Definition 3.2.8 (Statistical decision problem). A statistical decision problem (SDP) is a tuple (X, H, D, l, \mathbb{P}) where (X, \mathcal{X}) is a set of outcomes, (H, \mathcal{H}) is a set of hypotheses, (D, \mathcal{D}) is a set of decisions, $l : D \times H \rightarrow \mathbb{R}$ is a loss function and $\mathbb{P} : H \rightarrow \mathcal{X}$ is a Markov kernel from hypotheses to outcomes.

Statistical decision theory is concerned with the selection of *decision rules*, rather than the selection of decisions directly. A decision rule maps observations to decisions, and may be deterministic or stochastic.

Definition 3.2.9 (Decision rule). Given a statistical decision problem (X, H, D, l, \mathbb{P}) , a decision rule is a Markov kernel $\mathbb{D}_\alpha : \Omega \rightarrow D$.

Because decision rules in SDT play the role of what we call *choices*, we denote the set of all available decision rules by C . A further feature of SDT that is unlike the later decision theories is that SDT does not offer a single rule for assessing the desirability of any choice in C . Instead, it offers a definition of the risk, which assesses the desirability of a choice *relative to a particular hypothesis*. The risk function completely characterises the problem of choosing a decision function. Two different rules are for turning this “intermediate assessment” into a final assessment of the available choices - Bayes optimality and minimax optimality. Bayes optimality requires a prior over hypotheses, while minimax optimality does not.

Definition 3.2.10 (SDP Risk). Given a statistical decision problem (X, H, D, l, \mathbb{P}) and decision functions C , the *risk* functional $R : C \times H \rightarrow \mathbb{R}$ is defined by

$$R(\mathbb{D}_\alpha, h) := \int_X \int_D l(d, h) \mathbb{D}_\alpha(\mathrm{d}d|f) \mathbb{P}_h(\mathrm{d}f)$$

It is possible to find risk functions in problems that aren’t SDPs. The definitions of Bayes and Minimax optimality still apply to risk functions obtained on other manners. Thus Bayes optimality and minimax optimality are defined in terms of risk functions in general, not SDP risk functions.

Definition 3.2.11 (Bayes risk). Given decision functions C , hypotheses (H, \mathcal{H}) , risk $R : C \times H \rightarrow \mathbb{R}$ and prior $\mu \in \Delta(H)$, the μ -Bayes risk is

$$R_\mu(\mathbb{D}_\alpha) := \int_H R(\mathbb{D}_\alpha, h) \mu(\mathrm{d}h)$$

Definition 3.2.12 (Bayes optimal). Given decision functions C , hypotheses (H, \mathcal{H}) , risk $R : C \times H \rightarrow \mathbb{R}$ and prior $\mu \in \Delta(H)$, $\alpha \in C$ is μ -Bayes optimal if

$$R_\mu(\mathbb{D}_\alpha) = \inf_{\alpha' \in C} R_\mu(\mathbb{D}_{\alpha'})$$

Definition 3.2.13 (Minimax optimal). Given decision functions C , hypotheses (H, \mathcal{H}) , risk $R : C \times H \rightarrow \mathbb{R}$, a *minimax decision function* is any decision function \mathbb{D}_α satisfying

$$\sup_{h \in H} R(\mathbb{D}_\alpha, h) = \inf_{\alpha' \in C} \sup_{h \in H} R(\mathbb{D}_{\alpha'}, h)$$

From consequences to statistical decision problems

In this section, we relate our new work to the standard formulation of SDT presented above.

Statistical decision theory ignores the notion of general consequences of choices; the only “consequence” in the theory is the loss incurred by a particular decision under a particular hypothesis. The kinds of probability set models studied here probabilistically map decisions to consequences, and the set of consequences is understood to have a utility function to allow for assessment of the desirability of different choices via the principle of expected utility. Not every probability set model induces a statistical decision problem in this manner. A family of models that does are what we call *conditionally independent see-do models*. These models feature observations (the “see” part) along with decisions and consequences (the “do” part), and the observations come “before” the decisions (hence see-do). Examples of this type of model will be encountered again in Chapters 4 and 5. Furthermore, there is a hypothesis such that consequences are assumed to be independent of observations conditional on the decision and the hypothesis. This is why they are qualified as “conditionally independent” see-do models.

Definition 3.2.14 (See-do model). A probability set model of a statistical decision problem, or a *see-do model* for short, is a tuple $(\mathbb{P}_{C \times H}, \mathbf{X}, \mathbf{Y}, \mathbf{D})$ where $\mathbb{P}_{C \times H}$ is a probability set indexed by elements of $C \times H$ on (Ω, \mathcal{F}) , $\mathbf{X} : \Omega \rightarrow X$ are the observations, $\mathbf{Y} : \Omega \rightarrow Y$ are the consequences and $\mathbf{D} : \Omega \rightarrow D$ are the decisions. $\mathbb{P}_{C \times H}$ must observe the following conditional independences:

$$\begin{aligned} \mathbf{X} &\perp\!\!\!\perp_{\mathbb{P}_{C \times H}}^e \mathbf{C} | \mathbf{H} \\ \mathbf{D} &\perp\!\!\!\perp_{\mathbb{P}_{C \times H}}^e \mathbf{H} | \mathbf{C} \end{aligned}$$

where $\mathbf{C} : C \times H \rightarrow C$ and $\mathbf{H} : C \times H \rightarrow H$ are the respective projections (refer to Definition 2.4.16 for the definition of extended conditional independence $\perp\!\!\!\perp_{\mathbb{P}_{C \times H}}^e$).

Definition 3.2.15 (Conditionally independent see-do model). A conditionally independent see-do model is a see do model $(\mathbb{P}_{C \times H}, \mathbf{X}, \mathbf{Y}, \mathbf{D})$ where the following additional conditional independence holds:

$$\mathbf{Y} \perp\!\!\!\perp_{\mathbb{P}_{C \times H}}^e (\mathbf{X}, \mathbf{C}) | (\mathbf{D}, \mathbf{H})$$

We assume that a utility function is available depending on the consequence Y only, and identify the loss with the negative expected utility, conditional on a particular decision and hypothesis.

Definition 3.2.16 (Induced loss). Given a see-do model $(\mathbb{P}_{C \times H}, X, Y, D)$ and a utility $u : Y \rightarrow \mathbb{R}$, the induced loss $l : D \times H \rightarrow \mathbb{R}$ is defined as

$$l(d, h) := - \int_Y u(y) \mathbb{P}_{C \times \{h\}}^{Y|D}(dy|d)$$

where the uniform conditional $\mathbb{P}_{C \times \{h\}}^{Y|D}$'s existence is guaranteed by $Y \perp\!\!\!\perp_{\mathbb{P}_{C \times H}}^e (X, C)|(D, H)$.

A see-do model induces a set of decision functions: for each $\alpha \in C$, there is an associated probability distribution $\mathbb{P}_{\alpha}^{D|X}$. Using the above definition of loss, the expected loss of a decision function in a conditionally independent see-do model induces a risk function identical to the SDP risk.

Theorem 3.2.17 (Induced SDP risk). *Given a conditionally independent see-do model $(\mathbb{P}_{C \times H}, X, Y, D)$ along with a utility $u : Y \rightarrow \mathbb{R}$, the expected utility for each choice $\alpha \in C$ and hypothesis $h \in H$ is equal to the negative SDP risk of the associated decision rule $\mathbb{P}_{\alpha}^{D|X}$ and hypothesis h .*

$$\mathbb{P}_{\alpha, h}^Y u = -R(\mathbb{P}_{\{\alpha\} \times H}^{D|X}, h)$$

Proof. The expected utility given α and h is

$$\begin{aligned} \int_Y u(y) \mathbb{P}_{\alpha, h}^Y(dy) &= \int_Y \int_D \int_X u(y) \mathbb{P}_{\alpha, h}^{Y|DX}(dy|d, x) \mathbb{P}_{\alpha, h}^{D|X}(dd|x) \mathbb{P}_{\alpha, h}^X(dx) \\ &= \int_X \int_D \int_Y u(y) \mathbb{P}_{\alpha, h}^{Y|D}(dy|d) \mathbb{P}_{\alpha, h}^{D|X}(dd|x) \mathbb{P}_{\alpha, h}^X(dx) \quad (3.4) \\ &= \int_X \int_D \int_Y u(y) \mathbb{P}_{C \times \{h\}}^{Y|D}(dy|d) \mathbb{P}_{\{\alpha\} \times H}^{D|X}(dd|x) \mathbb{P}_{C \times \{h\}}^X(dx) \\ &= - \int_D \int_X l(d, h) \mathbb{P}_{\{\alpha\} \times H}^{D|X}(dd|x) \mathbb{P}_{C \times \{h\}}^X(dx) \\ &= -R(\mathbb{P}_{\{\alpha\} \times H}^{D|X}, h) \end{aligned}$$

where Equation (3.4) follows from $Y \perp\!\!\!\perp_{\mathbb{P}_{C \times H}}^e (X, C)|(D, H)$, the uniform conditional $\mathbb{P}_{\{\alpha\} \times H}^{D|X}$ exists due to $D \perp\!\!\!\perp_{\mathbb{P}_{C \times H}}^e H|C$ and the uniform conditional $\mathbb{P}_{C \times \{h\}}^X$ exists due to $X \perp\!\!\!\perp_{\mathbb{P}_{C \times H}}^e C|H$. \square

Theorem 3.2.17 does *not* hold for general see-do models. General see-do models allow for the utility to depend on X even after conditioning on D and H , while the form of the loss function in SDT forces no direct dependence on observations. The generic “see-do risk” (Definition 3.2.18) provides a notion of risk for the more general case, while Theorem 3.2.17 shows it reduces to SDP risk in the case of conditionally independent see-do models with a utility that depends only on the consequences Y .

Definition 3.2.18 (See-do risk). Given a see-do model $(\mathbb{P}_{C \times H}, \mathbf{X}, \mathbf{Y}, \mathbf{D})$ along with a utility $u : X \times Y \rightarrow \mathbb{R}$, the *see-do risk* $R : C \times H \rightarrow \mathbb{R}$ is given by

$$R(\alpha, h) := -\mathbb{P}_{\alpha, h}^{\mathbf{X}\mathbf{Y}} u \quad \forall \alpha \in C, h \in H$$

Section 3.1.1 noted that two types of probability set model are considered: probability sets \mathbb{P}_C indexed by choices alone, and probability sets $\mathbb{P}_{C \times H}$ jointly indexed by choices and hypotheses. See-do models are an instance of the second kind, jointly indexed by choices and hypotheses. Bayesian see-do models are of the former type, indexed by choices alone. A see-do model $(\mathbb{P}_{C \times H}, \mathbf{X}, \mathbf{Y}, \mathbf{D})$ and a prior over hypotheses $\mu \in \Delta(H)$ can be combined to form a Bayesian see-do model, and under the right conditions the risk of the Bayesian model reduces to the Bayes risk of the original see-do model.

Definition 3.2.19 (Bayesian see-do model). A Bayesian see-do model is a tuple $(\mathbb{P}_C, \mathbf{X}, \mathbf{Y}, \mathbf{D}, \mathbf{H})$ where \mathbb{P}_C is a probability set on (Ω, \mathcal{F}) , $\mathbf{X} : \Omega \rightarrow X$ are the observations, $\mathbf{Y} : \Omega \rightarrow Y$ are the consequences, $\mathbf{D} : \Omega \rightarrow D$ are the decisions and $\mathbf{H} : \Omega \rightarrow H$ is the hypothesis. \mathbb{P}_C must observe the following conditional independences:

$$\begin{aligned} \mathbf{X} &\perp\!\!\!\perp_{\mathbb{P}_C}^e \mathbf{C} | \mathbf{H} \\ \mathbf{D} &\perp\!\!\!\perp_{\mathbb{P}_C}^e \mathbf{H} | \mathbf{C} \\ \mathbf{H} &\perp\!\!\!\perp_{\mathbb{P}_C}^e \mathbf{C} \end{aligned}$$

Definition 3.2.20 (Induced Bayesian see-do model). Given a see-do model $(\mathbb{P}_{C \times H}, \mathbf{X}, \mathbf{Y}, \mathbf{D}, \mathbf{H})$ on (Ω, \mathcal{F}) and a prior $\mu \in \Delta(H)$, the induced Bayesian see-do model \mathbb{P}_C on $(\Omega \times H, \mathcal{F} \otimes \mathcal{H})$ is

$$\mathbb{P}_C(A) = \int_{\mathbf{H}^{-1}(A)} \mathbb{P}_{C \times \{h\}}(\Pi_\Omega^{-1}(A)) \mu(dh) \quad \forall A \in \mathcal{F} \otimes \mathcal{H}$$

Where $\Pi_\Omega : \Omega \times H \rightarrow \Omega$ is the projection onto Ω .

Theorem 3.2.21 (Induced SDP Bayes risk). *Given a conditionally independent see-do model $(\mathbb{P}_C, \mathbf{X}, \mathbf{Y}, \mathbf{D}, \mathbf{H})$ along with a utility $u : Y \rightarrow \mathbb{R}$ and a prior $\mu \in \Delta(H)$, the expected utility for each choice $\alpha \in C$ under the induced Bayesian see-do model is equal to the negative μ -Bayes risk of that decision rule.*

Proof. First, note that $h \mapsto \mathbb{P}_{C \times \{h\}}^{\mathbf{Y}|\mathbf{XD}}$ is a version of $\mathbb{P}_C^{\mathbf{Y}|\mathbf{XD}}$ and hence $\mathbf{Y} \perp\!\!\!\perp_{\mathbb{P}_C}^e (\mathbf{X}, \mathbf{C}) | (\mathbf{H}, \mathbf{D})$, a property it inherits from the underlying see-do model.

Also, note that $\mathbb{P}_C^{\mathbf{H}} = \mu$, by construction.

The expected utility of $\alpha \in C$ is

$$\begin{aligned}
\mathbb{P}_\alpha^Y u &= \int_Y u(y) \mathbb{P}_\alpha^Y(dy) \\
&= \int_Y \int_D \int_X \int_H u(y) \mathbb{P}_\alpha^{Y|D \times H}(dy|d, x, h) \mathbb{P}_\alpha^{D|X \times H}(dd|x, h) \mathbb{P}_\alpha^{X|H}(dx|h) \mathbb{P}_\alpha^H(dh) \\
&= \int_X \int_D \int_Y \int_H u(y) \mathbb{P}_\alpha^{Y|DH}(dy|d, h) \mathbb{P}_\alpha^{D|X}(dd|x) \mathbb{P}_\alpha^{X|H}(dx|h) \mathbb{P}_\alpha^H(dh) \\
&= \int_X \int_D \int_Y \int_H u(y) \mathbb{P}_C^{Y|DH}(dy|d, h) \mathbb{P}_\alpha^{D|X}(dd|x) \mathbb{P}_C^{X|H}(dx|h) \mu(dh) \\
&= - \int_D \int_X \int_H l(d, h) \mathbb{P}_\alpha^{D|X}(dd|x) \mathbb{P}_C^{X|H}(dx|h) \mu(dh) \\
&= - \int_H R(\mathbb{P}_\alpha^{D|X}, h) \mu(dh) \\
&= -R_\mu(\mathbb{P}_\alpha^{D|X})
\end{aligned}$$

□

Complete class theorem

The *complete class theorem* is a key theorem of classical SDT that establishes, under certain conditions, any *admissible* decision rule (Definition 3.2.23) for a see-do model $\mathbb{P}_{C \times H}$ with a utility u must minimise the Bayes risk for a Bayesian model constructed from $\mathbb{P}_{C \times H}$ and some prior over hypotheses $\mu \in \Delta(H)$. This can be interpreted in a similar way to the decision theoretic representation discussed above: if you accept that the relevant assumptions apply to the decision problem at hand, then there is a Bayesian see-do model along with u that captures the important features of this problem. The assumptions are that a see-do model $\mathbb{P}_{C \times H}$ with a utility u that satisfies the relevant conditions is available, and that the principle used to evaluate decision rules should yield an admissible decision rule (though it may also be desired to satisfy other properties as well).

If more is required of the decision rule than merely admissibility, then the complete class theorem does not prove that it is easy to find any Bayesian model that will yield rules satisfying these requirements. It also does not prove that a Bayesian approach is helpful for finding a “correct” decision rule according to some vague notion of “correct”.

We have shown in Theorem 3.2.17 that conditionally independent see-do models induce statistical decision problems. However, the complete class theorem itself (Theorem 3.2.24) depends only on the risk function induced by a decision making model. In particular, the complete class theorem can also apply to general see-do models, without the assumption of conditional independence, which we show in Example 3.2.31 and 3.2.32.

Definition 3.2.22 (Risk function). Given a set of choice C and a set of hypotheses H , a risk function is a map $R : H \times C \rightarrow \mathbb{R}$.

If the second set H were, instead of hypotheses about nature, a set of options available to a second player playing a game, then a “risk function” defines a two-player zero-sum game Ferguson (1967).

Definition 3.2.23 (Admissible choice). Given a risk function $R : C \times H \rightarrow \mathbb{R}$, a choice $\alpha \in C$ dominates a choice $\alpha' \in C$ if for all $h \in H$, $R(\alpha, h) \leq R(\alpha', h)$ and for at least on h^* , $R(\alpha, h) < R(\alpha', h)$. An *admissible choice* is a choice $\alpha \in C$ such that there is no $\alpha' \in C$ dominating α .

Definition 3.2.24 (Complete class). A *complete class* is any $B \subset C$ such that, for any $\alpha' \notin B$ there is some $\alpha \in B$ that dominates α' . A *minimal complete class* is a complete class B such that no proper subset of B is complete

Theorem 3.2.25. *If a minimal complete class $B \subset C$ exists then B is the set consisting of all the admissible decision rules.*

Proof. See Ferguson (1967, Theorem 2.1) □

Definition 3.2.26 (Risk set). Given a finite set of hypotheses H , a set of choices C and a risk function $R : C \times H \rightarrow \mathbb{R}$, the risk set is the subset of $\mathbb{R}^{|H|}$ given by

$$S := \{(R(\alpha, h))_{h \in H} | \alpha \in C\}$$

Theorem 3.2.27 (Complete class theorem). *Given a risk function $R : C \times H \rightarrow \mathbb{R}$, if the risk set S is convex, bounded from below and closed downwards, and H is finite, then the set of Bayes optimal choices is a minimal complete class.*

Proof. See Ferguson (1967, Theorem 2.10.2) □

Two examples of the application of the complete class theorem will be presented (Examples 3.2.31 and 3.2.32). In order to explain them, we need a few lemmas.

Lemma 3.2.28. *Given H and C both finite and a risk function $R : C \times H \rightarrow \mathbb{R}$ and an associated probability set \mathbb{P}_C on (Ω, \mathcal{F}) , Ω finite, if the function*

$$\mathbb{P}_{\alpha, h}^{\mathcal{D}|X} \mapsto R(\alpha, h)$$

is linear and

$$Q := ((\mathbb{P}_{\alpha, h}^{\mathcal{D}|X})_{h \in H})_{\alpha \in C}$$

is convex closed, then the risk set S is convex closed.

Proof. By linearity of

$$\mathbb{P}_{\alpha, h}^{\mathcal{D}|X} \mapsto R(\alpha, h)$$

we also have linearity of

$$(\mathbb{P}_{\alpha, h}^{\mathcal{D}|X})_{h \in H} \mapsto (R(\alpha, h))_{h \in H}$$

Furthermore, Q is bounded when viewed as an element of $\mathbb{R}^{\Omega \times H \times C}$, and so S is the linear image of a compact convex set, and is therefore also compact convex. □

Lemma 3.2.29. *For a see-do model $(\mathbb{P}_{C \times H}, \mathbf{X}, \mathbf{Y}, \mathbf{D}, \mathbf{H})$ with utility $u : X \times Y \rightarrow \mathbb{R}$, the map*

$$\mathbb{P}_{\alpha, h}^{\mathbf{D}|\mathbf{X}} \mapsto R(\alpha, h)$$

is linear.

Proof. By definition,

$$\begin{aligned} R(\alpha, h) &= -\mathbb{P}_{\alpha, h}^{\mathbf{X}\mathbf{Y}} u \\ &= -\mathbb{P}_{C \times \{h\}}^{\mathbf{X}} \odot \mathbb{P}_{\alpha \times h}^{\mathbf{D}|\mathbf{X}} \odot \mathbb{P}_{C \times \{h\}}^{\mathbf{Y}|\mathbf{D}\mathbf{X}} u \end{aligned}$$

Which is a composition of kernel products involving $\mathbb{P}_{\alpha \times H}^{\mathbf{D}|\mathbf{X}}$, and kernel products are linear, hence this function is linear. \square

The preceding theorem does *not* hold for a utility defined on Ω rather than on $X \times Y$. In this case we have instead

$$-\mathbb{P}_{C \times \{h\}}^{\mathbf{X}} \odot \mathbb{P}_{\alpha \times h}^{\mathbf{D}|\mathbf{X}} \odot \mathbb{P}_{\alpha, h}^{\Omega|\mathbf{D}\mathbf{X}} u$$

where α appears twice on the right hand side, rendering the map nonlinear.

Lemma 3.2.30. *For finite X and D , the set of all Markov kernels $X \rightarrow D$ is convex closed.*

Proof. From Blackwell (1979), the set of all Markov kernels $X \rightarrow D$ is the convex hull of the set of all deterministic Markov kernels $X \rightarrow D$. There are a finite number of deterministic Markov kernels, and so the convex hull of this set is closed. \square

Example 3.2.31. Suppose we have a conditionally independent see-do model $(\mathbb{P}_C, \mathbf{X}, \mathbf{Y}, \mathbf{D}, \mathbf{H})$ along with a bounded utility $u : Y \rightarrow \mathbb{R}$ where H, D, X and Y are all finite, and $\{\mathbb{P}_{\alpha}^{\mathbf{D}|\mathbf{X}} | \alpha \in C\}$ is the set of all Markov kernels $X \rightarrow D$. Then the risk set is convex and closed downwards, and so the set of Bayes optimal choices is exactly the set of admissible choices.

The boundedness of the risk set S follows from the boundedness of the utility u ; if u is bounded above by k , then S is bounded below in every dimension by $-k$.

The fact that S is convex and closed follows from Lemmas 3.2.28, 3.2.29 and 3.2.30.

Example 3.2.32. As before, but suppose we have the see-do model is not conditionally independent. Because none of the lemmas 3.2.28, 3.2.29 and 3.2.30 made use of the conditional independence assumption, the risk set is still convex and closed downwards and so the set of Bayes optimal choices is also exactly the set of admissible choices.

3.3 Variables

In probability theory, it is standard to assume the existence of a probability space $(\mu, \Omega, \mathcal{F})$ and to define *random variables* as measurable functions from (Ω, \mathcal{F}) to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. However, variables aren't *just* functions – they're also typically understood to correspond to some measured aspect of the real world. For example, Pearl (2009) offers the following two, purportedly equivalent, definitions of variables:

By a *variable* we will mean an attribute, measurement or inquiry that may take on one of several possible outcomes, or values, from a specified domain. If we have beliefs (i.e., probabilities) attached to the possible values that a variable may attain, we will call that variable a random variable.

This is a minor generalization of the textbook definition, according to which a random variable is a mapping from the sample space (e.g., the set of elementary events) to the real line. In our definition, the mapping is from the sample space to any set of objects called “values,” which may or may not be ordered.

However, these are actually two different things. The first is a *measurement*, which is something we can do in the real world that produces as a result an element of a mathematical set. The second is a *function*, a purely mathematical object with a domain and a codomain and a mapping from the former into the latter. Measurement procedures play the extremely important role of “pointing to the parts of the world” that the model addresses.

The general scheme considered in this work is to assume that there is a collection of “complete measurement procedure” S_α , one for each choice $\alpha \in C$. S_α is considered to be the procedure that measures all quantities of interest, and any subprocedure corresponding to a particular quantity of interest reconstructed from the result of S by applying a function to its result. The function X that, when applied to the result of S , yields the result of a measurement subprocedure \mathcal{X} is the *variable* associated with the measurement procedure \mathcal{X} . In this way, a variable X – which is by itself just a mathematical function – is associated with a measurement procedure in the real world.

3.3.1 Variables and measurement procedures

Consider Newton's second law in the form $F = MA$. This model relates “variables” F , M and A . As Feynman (1979) noted, in order to understand this law, some pre-existing understanding of force, mass and acceleration is required. In order to offer a numerical value for the net force on a given object is, even the most knowledgeable physicist will have to go and do a measurement, which involves interacting with the object in some manner that cannot be completely mathematically specified, and which will return a numerical value that will be taken to be the net force.

In order to make sense of the equation $F = MA$, it must be understood relative to some measurement procedure S that simultaneously measures the force on an object, its mass and its acceleration, which can be recovered by the functions F , M and A respectively. The equation then says that, whatever result s this procedure yields, $F(s) = M(s)A(s)$ will hold.

A measurement procedure S is akin to Menger (2003)’s notion of variables as “consistent classes of quantities” that consist of pairing between real-world objects and quantities of some type. S itself is not a well-defined mathematical thing. At the same time, the set of values it may yield *is* a well-defined mathematical set. No actual procedure can be guaranteed to return elements of a mathematical set known in advance – anything can fail – but we assume that we can study procedures reliable enough that we don’t lose much by ignoring this possibility.

Note that, because S is not a purely mathematical thing, we cannot perform mathematical reasoning with S directly. It is much more practical to relegate S to the background, and reason in terms of the functions F , M and A . However, even if we don’t talk about it much, S remains an important element of the law.

3.3.2 Measurement procedures

Definition 3.3.1 (Measurement procedure). A *measurement procedure* \mathcal{B} is a procedure that involves interacting with the real world somehow and delivering an element of a mathematical set X as a result. A procedure \mathcal{B} is said to takes values in a set B .

We adopt the convention that the procedure name \mathcal{B} and the set of values B share the same letter.

Definition 3.3.2 (Values yielded by procedures). $\mathcal{B} \bowtie x$ is the proposition that the the procedure \mathcal{B} will yield the value $x \in X$. $\mathcal{B} \bowtie A$ for $A \subset X$ is the proposition $\bigvee_{x \in A} \mathcal{B} \bowtie x$.

Definition 3.3.3 (Equivalence of procedures). Two procedures \mathcal{B} and \mathcal{C} are equal if they both take values in X and $\mathcal{B} \bowtie x \iff \mathcal{C} \bowtie x$ for all $x \in X$.

If two involve different measurement actions in the real world but necessarily yield the same result, we say they are equivalent.

It is worth noting that this notion of equivalence identifies procedures with different real-world actions. For example, “measure the force” and “measure everything, then discard everything but the force” are often different – in particular, it might be possible to measure the force only before one has measured everything else. Thus the result yielded by the first procedure could be available before the result of the second. However, if the first is carried out in the course of carrying out the second, they both yield the same result in the end and so we treat them as equivalent.

Measurement procedures are like functions without well-defined domains. Just like we can compose functions with other functions to create new functions, we can compose measurement procedures with functions to produce new measurement procedures.

Definition 3.3.4 (Composition of functions with procedures). Given a procedure \mathcal{B} that takes values in some set B , and a function $f : B \rightarrow C$, define the “composition” $f \circ \mathcal{B}$ to be any procedure \mathcal{C} that yields $f(x)$ whenever \mathcal{B} yields x . We can construct such a procedure by describing the steps: first, do \mathcal{B} and secondly, apply f to the value yielded by \mathcal{B} .

For example, \mathcal{MA} is the composition of $h : (x, y) \mapsto xy$ with the procedure $(\mathcal{M}, \mathcal{A})$ that yields the mass and acceleration of the same object. Measurement procedure composition is associative:

$$\begin{aligned} (g \circ f) \circ \mathcal{B} \text{ yields } x &\iff \mathcal{B} \text{ yields } (g \circ f)^{-1}(x) \\ &\iff \mathcal{B} \text{ yields } f^{-1}(g^{-1}(x)) \\ &\iff f \circ \mathcal{B} \text{ yields } g^{-1}(x) \\ &\iff g \circ (f \circ \mathcal{B}) \text{ yields } x \end{aligned}$$

One might wonder whether there is also some kind of “tensor product” operation that takes a standalone \mathcal{M} and a standalone \mathcal{A} and returns a procedure $(\mathcal{M}, \mathcal{A})$. Unlike function composition, this would be an operation that acts on two procedures rather than a procedure and a function. Thus this “append” combines real-world operations somehow, which might introduce additional requirements (we can’t just measure mass and acceleration; we need to measure the mass and acceleration of the same object at the same time), and may be under-specified. For example, measuring a subatomic particle’s position and momentum can be done separately, but if we wish to combine the two procedures then we can get different results depending on the order in which we combine them.

Our approach here is to suppose that there is some complete measurement procedure \mathcal{S} to be modeled, which takes values in the observable sample space (Ψ, \mathcal{E}) and for all measurement procedures of interest there is some f such that the procedure is equivalent to $f \circ \mathcal{S}$ for some f . In this manner, we assume that any problems that arise from a need to combine real world actions have already been solved in the course of defining \mathcal{S} .

Given that measurement processes are in practice finite precision and with finite range, Ψ will generally be a finite set. We can therefore equip Ψ with the collection of measurable sets given by the power set $\mathcal{E} := \mathcal{P}(\Psi)$, and (Ψ, \mathcal{E}) is a standard measurable space. \mathcal{E} stands for a complete collection of logical propositions we can generate that depend on the results yielded by the measurement procedure \mathcal{S} .

One could also consider measurement procedures to produce results in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ (i.e. the reals with the Borel sigma-algebra) or a set isomorphic to it. This choice is often made in practice, and following standard practice we also often consider variables to take values in sets isomorphic to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. However, for measurement in particular this seems to be a choice of convenience rather than necessity – for any measurement with finite precision and range, it is possible to specify a finite set of possible results.

3.3.3 Observable variables

Our *complete* procedure \mathcal{S} represents a large collection of subprocedures of interest, each of which can be obtained by composition of some function with \mathcal{S} . We call the pair consisting of a subprocedure of interest \mathcal{X} along with the variable X used to obtain it from \mathcal{S} an *observable variable*.

Definition 3.3.5 (Observable variable). Given a measurement procedure \mathcal{S} taking values in (Ψ, \mathcal{E}) , an observable variable is a pair $(X \circ \mathcal{S}, X)$ where $X : (\Psi, \mathcal{E}) \rightarrow (X, \mathcal{X})$ is a measurable function and $\mathcal{X} := X \circ \mathcal{S}$ is the measurement procedure induced by X and \mathcal{S} .

For the model $F = MA$, for example, suppose we have a complete measurement procedure \mathcal{S} that yields a triple (force, mass, acceleration) taking values in the sets X, Y, Z respectively. Then we can define the “force” variable (\mathcal{F}, F) where $\mathcal{F} := F \circ \mathcal{S}$ and $F : X \times Y \times Z \rightarrow X$ is the projection function onto X .

A measurement procedure yields a particular value when it is completed. We will call a proposition of the form “ \mathcal{X} yields x ” an *observation*. Note that \mathcal{X} need not be a complete procedure here. Given the complete procedure \mathcal{S} , a variable $X : \Psi \rightarrow X$ and the corresponding procedure $\mathcal{X} = X \circ \mathcal{S}$, the proposition “ \mathcal{X} yields x ” is equivalent to the proposition “ \mathcal{S} yields a value in $X^{-1}(x)$ ”. Because of this, we define the *event* $X \bowtie x$ to be the set $X^{-1}(x)$.

Definition 3.3.6 (Event). Given the complete procedure \mathcal{S} taking values in Ψ and an observable variable $(X \circ \mathcal{S}, X)$ for $X : \Psi \rightarrow X$, the *event* $X \bowtie x$ is the set $X^{-1}(x)$ for any $x \in X$.

If we are given an observation “ \mathcal{X} yields x ”, then the corresponding event $X \bowtie x$ is *compatible with this observation*.

It is common to use the symbol $=$ instead of \bowtie to stand for “yields”, but we want to avoid this because $Y = y$ already has a meaning, namely that Y is a constant function everywhere equal to y .

An *impossible event* is the empty set. If $X \bowtie x = \emptyset$ this means that we have identified no possible outcomes of the measurement process \mathcal{S} compatible with the observation “ \mathcal{X} yields x ”.

3.3.4 Model variables

Observable variables are special in the sense that they are tied to a particular measurement procedure \mathcal{S} . However, the measurement procedure \mathcal{S} does not enter into our mathematical reasoning; it guides our construction of a mathematical model, but once this is done mathematical reasoning proceeds entirely with mathematical objects like sets and functions, with no further reference to the measurement procedure.

A *model variable* is simply a measurable function with domain (Ψ, \mathcal{E}) .

Model variables do not have to be derived from observable variables. We may instead choose a sample space for our model (Ω, \mathcal{F}) that does not correspond to the possible values that \mathcal{S} might yield. In that case, we require a surjective

model variable $S : \Omega \rightarrow \Psi$ called the complete observable variable, and every observable variable $(X' \circ S, X')$ is associated with the model variable $X := X' \circ S$.

An *unobserved variable* is a variable whose set of possible values is not constrained by the results of the measurement procedure.

Definition 3.3.7 (Unobserved variable). Given a sample space (Ω, \mathcal{F}) and a complete observable variable $S : \Omega \rightarrow \Psi$, a model variable $Y : \Omega \rightarrow Y$ is *unobserved* if $Y(S \bowtie s) = Y$ for all $s \in \Psi$.

3.3.5 Variable sequences and partial order

Given $Y : \Omega \rightarrow X$, we can define a sequence of variables: $(X, Y) := \omega \mapsto (X(\omega), Y(\omega))$. (X, Y) has the property that $(X, Y) \bowtie (x, y) = X \bowtie x \cap Y \bowtie y$, which supports the interpretation of (X, Y) as the values yielded by X and Y together.

Define the partial order on variables \preceq where $X \preceq Y$ can be read “ X is completely determined by Y ”.

Definition 3.3.8 (Variables determined by another variable). Given a sample space (Ω, \mathcal{F}) and variables $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$, $X \preceq Y$ if there is some $f : Y \rightarrow X$ such that $X = f \circ Y$.

Clearly, $X \preceq (X, Y)$ for any X and Y .

3.3.6 Decision procedures

The kind of problem we want to solve requires us to compare the consequences of different choices from a set of possibilities C . We take the *consequences of* $\alpha \in C$ to refer to the values obtained by some measurement procedure \mathcal{S}_α associated with the choice α .

As we have said, what exactly a “measurement procedure” is is a bit vague – it’s “what we actually do to get the numbers we associate with variables”. It seems we could describe the above in terms of a single measurement procedure \mathcal{S} , which involves:

1. Choose α
2. Proceed according to \mathcal{S}_α

However, \mathcal{S} is problematic to model. The model is often part of the process of choosing α , and so a model of \mathcal{S} that involves the step “choose α ” will be self-referential. Because of this, we don’t try to model \mathcal{S} , and whether this changes anything is an open question.

Definition 3.3.9 (Decision procedure). A decision procedure is a collection $\{\mathcal{S}_\alpha\}_{\alpha \in C}$ of measurement procedures.

Like measurement procedures, a decision procedure $\{\mathcal{S}_\alpha\}_{\alpha \in A}$ isn’t a well-defined mathematical object; it’s not really a “set”, because the contents are real-world actions.

3.4 Conclusion

We define “decision making models” as maps from a set of choices C to distributions over a set of consequences Ω . We suppose that decision making models are accompanied by a utility function that rates the desirability of each consequence, though we do not often explicitly consider the utility function. This general scheme is common to many theories of decision making.

We distinguish decision making models with choices only from decision making models with choices and consequences. The former are “Bayesian” models, with the consequences of each choice given by a unique probability distribution, while the latter are “non-Bayesian”. Bayesian models with an expected utility induce a complete order on the choices – each choice is either better, worse or just the same as another choice. On the other hand, non-Bayesian models induce a partial order, with admissible choices being better than inadmissible choices, but pairs of admissible choices are not known to be indifferent. The complete class theorem shows that any rule for selecting from the admissible choices can be rationalised as a rule for selecting Bayes-optimal choices with respect to *some* prior, and we show that this theorem applies to see-do models equipped with a utility function if the set of hypotheses is finite and the induced risk function is convex and downward closed.

We introduce variables and measurement procedures as our understanding of how models correspond to “real world decision problems”. Measurement procedures are typically in the background, and we don’t explicitly discuss them. However, when we talk about “observed variables”, we mean that there is a measurement procedure in the background, and an observed variable is a partial result of this procedure.

3.5 Appendix: axiomatisation of decision theories

3.5.1 Savage axioms

Careful analysis of Savage’s theorem is outside the scope of this work, but for the reader’s convenience we will reproduce the axioms from Savage (1954) with a small amount of commentary. Keep in mind that Savage’s theorem establishes that the following are sufficient for representation with a probability set, not necessary, and furthermore the probability set representation of preferences satisfying these axioms is unique.

Given acts C , states (S, \mathcal{S}) and consequences F and a map $T : S \times C \rightarrow F$, let all greek letters α, β etc. be elements of C . Savage’s axioms are:

P1: There is a complete preference relation \preceq on C

D1: $\alpha \preceq \beta$ given $B \in \mathcal{S}$ if and only if $\alpha' \preceq \beta'$ for every α' and β' such that $T(\alpha, s) = T(\alpha', s)$ for $s \in B$ and $T(\alpha', r) = T(\beta', r)$ for $r \notin B$, and $\beta' \preceq \alpha'$ either for every such pair or for none.

P2: For every α, β and $B \in \mathcal{S}$, $\alpha \preceq \beta$ given B or $\beta \preceq \alpha$ given B

D2: for $q, q' \in F$, $q \preceq q'$ if and only if $\alpha \preceq \alpha'$ where $T(\alpha, s) = q$ and $T(\alpha', s) = q'$ for all $s \in S$

D2: $B \in \mathcal{S}$ is null if and only if $\alpha \preceq \beta$ given B for every $\alpha, \beta \in C$

P3: If $T(\alpha, s) = q$ and $T(\alpha', s) = q'$ for every $s \in B$, $B \in \mathcal{S}$ non-null, then $\alpha \preceq \alpha'$ given B if and only if $q \preceq q'$

D4: For $A, B \in \mathcal{S}$, $A \leq B$ if and only if $\alpha_A \preceq \alpha_B$ or $q \preceq q'$ for all $\alpha_A, \alpha_B \in C$, $q, q' \in F$ such that $T(\alpha_A, s) = q$ for $s \in A$, $T(\alpha_A, s') = q'$ for $s' \notin A$, $T(\alpha_B, s) = q$ for $s \in B$, $T(\alpha_B, s') = q'$ for $s' \notin B$. Read \leq as “is less probable than”

P4: For every $A, B \in \mathcal{S}$, $A \leq B$ or $B \leq A$

P5: For some α, β , $\alpha \prec \beta$

P6: Suppose $\alpha \not\preceq \beta$. Then for every γ there is a finite partition of S such that if α' agrees with α and β' agrees with β except on some element B of the partition, α' and β' being equal to γ on B , then $\alpha \not\preceq \beta'$ and $\alpha' \not\preceq \beta$

D5: $\alpha \preceq q$ for $q \in F$ given B if and only if $\alpha \preceq \beta$ given B where $T(\beta, s) = q$ for all $s \in S$

P7: If $\alpha \preceq T(\beta, s)$ given B for every $s \in B$, then $\alpha \preceq \beta$ given B

P7': The proposition given by inverting every expression in D5 and P7

D1 formalises the idea of one act α being not preferred to another β given the knowledge that the true state lies in the set B (in short: “given B ” or “conditional on B ”). P2 is sometimes called the “sure thing principle”, as it implies the following: for any α, β if α is better than β on some states and no worse on any other, then $\alpha \succ \beta$. In Savage’s model, the “likelihood” that of any state cannot depend on the act chosen.

D4 + P4 defines the “probability preorder” \leq on (S, \mathcal{S}) and assumes it is complete.

P5 is the requirement that the preference relation is non-trivial; not everything is equally desirable. This doesn’t seem like it should be a practical requirement to me; we might hope that a model can distinguish between some of our options, but that doesn’t mean we should assume it can. Savage claims that this requirement is “innocuous” because any exception must be trivial, but I’m not sure I agree.

P6 is a requirement of continuity; for any $\alpha \preceq \beta$, we can divide S finely enough to squeeze a “small slice” of any third outcome γ into the gap between the two.

P7 in combination with the other axioms forces preferences to be bounded.

3.5.2 Bolker axioms

$\underline{\mathcal{F}}$ a complete, atomless Boolean algebra with the impossible proposition removed.

A1: \preceq is a complete preference relation

B2: $\underline{\mathcal{F}}$ is a complete, atomless Boolean algebra with the impossible proposition removed

C3: For $A, B \in \underline{\mathcal{F}}$, if $A \cap B = \emptyset$, then

a) If $A \succ B$ then $A \succ A \cup B \succ B$

b) If $A \sim B$ then $A \sim A \cup B \sim B$

D4: Given $A \cap B = \emptyset$ and $A \sim B$, if $A \cup G \sim B \cup G$ for some G where $A \cap G = B \cap G = \emptyset$ and $G \not\sim A$, then $A \cup G \sim B \cup G$ for every such G

D1: The supremum (infimum) of a subset $W \subset \underline{\mathcal{F}}$ is a set G (D) such that for all $A \in W$, $G \subset A$ ($A \subset D$), and for any E that also has this property, $G \subset E$ ($E \subset D$)

E5: Given $W := \{W_i\}_{i \in M \subset \mathbb{N}}$ with $i < j \implies W_j \subset W_i$ and $W \subset \underline{\mathcal{F}}$ with supremum G (infimum D), whenever $A \prec G \prec B$ ($A \prec D \prec B$) then there exists some $k \in M$ such that $i \geq k$ ($i \leq k$) implies $A \prec W_i \prec B$.

Like Savage's theory, A1 requires the preference relation to be complete.

A3 is the assumption that the desirability of disjunctions of events lies between the desirability of each event; it is sometimes called "averaging". It notably rules out the following: if $A \succ B$ we cannot have $A \cup B \sim A$. In the Jeffrey-Bolker theory, propositions all have positive probabilities.

A4 allows a probability order to be defined on $\underline{\mathcal{F}}$. The conditions $A \cap B = \emptyset$, $A \sim B$, $A \cup G \sim B \cup G$ for some G where $A \cap G = B \cap G = \emptyset$ and $G \not\sim A$ can be seen as a test for A and B being "equally probable". A4 requires that if A and B are rated as equally probable by one such test, then they are rated as equally probable by all such tests.

A5 is an axiom of continuity.

Chapter 4

Models of repeatable decision problems

Also mention that response functions are formally identical to “causal mechanisms” in the sense of Pearl’s stable mechanisms and Janzing’s independent mechanisms

Chapter 2 introduced probability sets as generic tools for causal modelling, while Chapter 3 examined how probability set models can be used in decision problems and Section 3.2.5 in particular introduced *see-do* models, which featured four variables representing observations, consequences, choices and hypotheses. A decision maker wants to pick choices that promote desirable consequences, and in this chapter, we investigate how they can use observations to inform their views of which choices go with which consequences.

A distinguishing feature of decision making is the need to compare the consequences of multiple different options (see Section 1.2). In the setup we consider in this work, a decision maker is interested in making a single choice. After making their choice, they can observe the consequences of the option they chose, but they can only speculate about the consequences of all of the other options they had available. Thus, instead of observing an entire (stochastic) response function that maps choices to consequences, which is what they used to make the decision, they only observe the function’s output for the particular choice they made. For example, if the decision maker is a person considering taking a medicine to help with a headache they can either take the medicine, in which case they never find out what would have happened if they didn’t take it, or they could avoid the medicine and never learn the consequences of taking it.

A simple case to consider where the decision maker can learn a function mapping their choices to consequences is when they face a choice that is, in the appropriate sense, repeated. Specifically, we suppose that there is a fixed response function that maps “inputs” to “outputs”¹ and the decision maker has

¹a note on terminology: in this work, for practical reasons we say that the decision maker

multiple opportunities to pick an input and observe the outputs. In this case, the decision maker can learn the entire function by trying each possible input a number of times. For example, if the person previously discussed frequently has headaches they could sometimes take the medicine and sometimes avoid it (these are the different inputs), and compare the subsequent course of each headache (the outputs).

However, it's not entirely clear what this assumption – that the decision maker's choice determines inputs to repeated copies of a fixed response function – actually means. Our headache-prone individual cannot examine the source code of the universe and find that some fixed function is called every time they have a headache and take (or avoid) medication for it. This assumption of a repeated response function plays a similar role to the assumption of independent and identically distributed (IID) data in traditional probabilistic inference. Just as the IID assumption underpins a great deal of theory and methods for probabilistic inference, the assumption of repeated response functions underpins many common models of causal inference, as will be discussed in Chapter 5.

In comparison with the IID assumption, the assumption of repeated response functions is less often appropriate. Consider our headache-prone individual once more. For argument's sake, they might reasonably assume that their daily history of medication and subsequent headaches can be modeled by an IID sequence of pairs of variables (X_i, Y_i) where X_i represents whether they took medication on day i and Y_i the severity of their subsequent headache. The distribution of this sequence \mathbb{P} will also induce a conditional distribution $\mathbb{P}^{Y_1|X_1}$ which is a response map from medication to outcomes. However, this is *not* the stochastic map that this individual should use to help them make a decision today. Supposing that in the past they only took the medication when they had a headache, then most of the days on which they took no medication were days on which they had no headache to begin with; in this case, the conditional distribution $\mathbb{P}^{Y_1|X_1}$ may well indicate that they are much more likely to have headaches on days where they took the medication, even if the medication is in reality quite effective at relieving their pain.

This story describes a standard case of confounding – this person's experience of headaches after taking medication is confounded by whether or not they had a headache before taking it. Confounding is ubiquitous in situations in which people are trying to use data to inform choices, and is one of the major reasons for the aphorism “correlation does not imply causation”.

In summary, the assumption of repeated response functions is often a critical assumption for causal inference (like the assumption of IID variables in classical statistics), it is not clear exactly what justifies this assumption and it is often inappropriate. This chapter aims to clarify the justification for this assumption. In this chapter, we present results that facilitate an alternative interpretation of this assumption, which (to some extent) addresses the second point – that it's not clear exactly what justifies the assumption of repeated response functions

can make one *choice* that determines the value of many *inputs*, and observes many *outputs* that together determine the overall *consequences*.

– though rather than offering new practical justifications for the assumption of repeated response functions, the perspective we offer mainly reinforces the widely held view that this assumption is mostly inappropriate².

What we show is analogous to a well-known result of Bruno De Finetti for conditionally independent and identically distributed sequences of observations. De Finetti considered models where observations were given by repetitions of identical but unknown probability distributions, where we consider input-output pairs given by repetitions of identical but unknown stochastic functions. De Finetti shows that this structural assumption was equivalent to an assumption that the measurement procedure obeyed a certain symmetry. In particular, the assumption of conditionally independent and identically distributed sequences was appropriate precisely when the measurement procedure in question was, for the purposes of modelling, identical to any measurement procedure that proceeded in the same fashion but permuted the indices to which each observation was assigned³.

In this chapter, we examine symmetries of this sort. The key equivalence we show can be roughly stated in the following form: given a model of a sequence of input-output pairs, these pairs can be related by repeated response functions if and only if the distribution of finite sequences of outputs conditioned on a corresponding sequence of inputs *and* an infinite history of other input-output pairs is unchanged under arbitrary permutations.

The motivation of deriving this result was, in part, to consider alternative justifications for the assumption of repeated response functions. However, this result is, in our view, mostly negative. Our result implies, for example:

- Suppose we have sequence of input-output pairs from a well-conducted experiment and a similar sequence from passive observation, and want to predict a held-out experimental output; the assumption of repeatable response functions implies that the experimental and observational data are interchangeable for this purpose
- Suppose we have sequence of input-output pairs from a well-conducted experiment, and are interested in predicting the consequences of our own plans under consideration; the assumption of repeatable response functions implies that this problem is essentially the same as predicting held-out experimental outputs

²It’s very easy to find statements like “this assumption seems unreasonably strong, but we have to make it if we want to work anything out” – see, for example, Saarela et al. (2020, pg. 11), Hernán and Robins (2006, pg. 579) or Pearl (2009, pg. 40). One can also find many criticisms of inappropriate use of this assumption, see for example Muller (2015) or Berk (2010).

³Note that this result does not apply in a non-Bayesian setting where we use sets of probability distributions rather than a single probability distribution to model observations. In this setting, the symmetry over measurement procedures described here does not imply the structural results of independent and identically distributed variables, see Walley (1991, pg. 463). We consider the “Bayesian” setting here of a single stochastic function because it is simpler.

In many situations, we expect that both of these implications are not acceptable. In fact, little domain expertise seems to be required to recognise that the different problems discussed are *not* essentially the same. Whether the experiment is testing medical treatments, educational interventions or software modifications – in all of these circumstances, one doesn’t need deep domain knowledge to know that data generated in different contexts is usually not interchangeable.

This is not to say that the assumption of repeated response functions is never acceptable, but that the required symmetry places some strict limits on the cases when it is. In this chapter we use the example of a “multi-armed bandit”, where the assumption is justified by the fact that the experiment is repeatedly interacting with a machine that is known to implement a fixed input-output function. A/B testing, where a developer randomly chooses which version of a page is served to users for some time, and deterministically picks the best page thereafter plausibly satisfies the second symmetry above – serving page version “B” because you’ve decided it’s the best and serving page version “B” because you’re continuing the experiment do seem like two situations that call for the same predictive model (at least, if there is good reason to neglect potential interactions between versions that load at different times).

Thus, the major practical conclusion we draw from these results is that the assumption of repeated response functions is usually too strong. We do want to use data to help make decisions, so we’re motivated to find assumptions that allow us to do this that are weaker than that of repeated response functions. In Chapter 5 we present two existing solutions to this problem, as well as introducing the assumptions of *precedented responses* and *individual-level response functions*, each inspired by one of the existing solutions.

This chapter also has a preliminary investigation into repeated response functions in the case of data-dependent models, where inputs are allowed to depend arbitrarily on any of the previous inputs and outputs. This is a generalisation of the standard causal inference setting where actions taken and consequences experienced after the data is reviewed do not appear in the model. In this setting, we consider *probability combs* which are a kind of generalised conditional probability introduced by Chiribella et al. (2008) and applied to causal models by Jacobs et al. (2019). We show that data-dependent models with repeated stochastic functions feature probability comb symmetries.

4.0.1 Chapter outline

This chapter introduces sequences of *conditionally independent and identical response functions*, a precise term for what we refer to above as “repeated response functions”. The key theorem in this chapter, Theorem 4.3.22, relates the assumption of conditionally independent and identical response functions to a kind of symmetry which we call *IO contractibility*. A model with data-independent actions features conditionally independent and identical response functions if and only if it is IO contractible. Theorem 4.6.8 introduces a more general notion of IO contractibility and relaxes the data-independent assumption,

but comes with some different side conditions.

Section 4.1 surveys previous work, particularly related to symmetries of causal models. Section 4.2 defines and explains the idea of conditionally independent and identical response functions. Section 4.3 defines IO contractibility, as well as setting out key definitions, lemmas and the proof of Theorem 4.3.22. Section ?? presents a collection of examples that illustrate various features of models that are (or are not) IO contractible. Section 4.5 extends the work from Section 4.3 to models where inputs can be data-dependent. The extension is dense and retreads a lot of ground already covered in a slightly different way, but Section 4.5.1 introduces the notion of a comb, which is an extension of a conditional probability, that has applications in areas of causal inference beyond what is covered in this chapter, and this subsection stands on its own. Finally, some concluding remarks are in Section 4.7.

4.0.2 Key terminology

C C inputs outputs $\mathbb{P}_C^{Y|D}$ exchange commutative local IO contractible H

4.1 Previous work on causal symmetries

de Finetti ([1937] 1992) introduced two key ideas to probability modelling: first, he established an equivalence between exchangeable sequences and conditionally independent and identically distributed sequences, and secondly he proposed that we can deduce symmetries of probability models from informal idea that measurement procedures differing only by label permutations are essentially identical. De Finetti’s technical result has been extended in many ways, including to finite sequences Kerns and Székely (2006); Diaconis and Freedman (1980) and for partially exchangeable arrays Aldous (1981). A comprehensive overview of results is presented in Kallenberg (2005b). A result from classical statistics that is particularly similar to the result presented in this chapter is the notion of “partial exchangeability” from Diaconis (1988).

The application of similar ideas to causal models has received some attention, though comparatively little in comparison. Lindley and Novick (1981) discuss models consisting of a sequence of exchangeable observations along with “one more observation”, a structure that is similar to the models with observations and consequences discussed in section 4.4.1. Lindley discusses the application of this model to questions of causation, but does not explore this deeply due to the perceived difficulty of finding a satisfactory definition of causation. Rubin (2005)’s overview of causal inference with potential outcomes along with the text Imbens and Rubin (2015) make use of the assumption of exchangeable potential outcomes to prove several identification results, though because potential outcomes are not observable it’s not immediately obvious how to deduce the exchangeability of potential outcomes from a perceived symmetry of measurement procedures. Saarela et al. (2020), using structural causal models, proposes *conditional exchangeability*, defined using structural causal model as

the exchangeability of the non-intervened causal parents of a target variable under intervention on some of its parents. Sareela et. al. suggest that this could be interpreted as a symmetry of an experiment involving administering treatments to patients with respect to exchanging the patients in the experiment. In fact, many authors have posited causal notions of exchangeability that involve swapping people or experimental units involved in an experiment: Hernán and Robins (2006); Hernán (2012); Greenland and Robins (1986); Banerjee et al. (2017); Dawid (2020) all discuss assumptions of this type. Section 5.3 discusses the notion of “swapping individuals” in more detail.

A stronger symmetry assumption than commutativity of exchange, which is comparable to the symmetries discussed above, is the assumption of *IO contractibility* (Definition 4.3.2), which adds the assumption of *locality*. This additional assumption appears to have similarities to the stable unit treatment distribution assumption (SUTDA) in Dawid (2020), and the stable unit treatment value assumption (SUTVA) in (Rubin, 2005): “(”SUTVA) comprises two sub-assumptions. 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.

There are two subtle caveats to existing causal treatments of symmetries. First, the kind of symmetry originally considered by De Finetti and by subsequent work in classical statistics involved probability models that were unchanged under permutations of a sequence of random variables (or under some other transformation). By contrast, the causal treatments usually consider models where the “true” interventional distributions (for example, Saarela et al. (2020)) or the “true” conditional distributions of potential outcomes (for example, Hernán and Robins (2006)) are unchanged under some transformation. As we clarify in this chapter, this amounts to the claim that the *in the limit of infinite conditioning data*, the distribution of an output conditional on an input is unchanged by the transformations in question. Other features of the model might be substantially changed by the transformation.

The second subtle point is the nature of the transformations themselves. The kind of transformation envisioned in De Finetti’s original result is of the following form: suppose you conduct some measurement procedure and write down the results in a table of values. These results are bound to random variables on the basis of their position in the table. Consider an alternative measurement procedure: we do exactly as before, but write the same numbers in different positions in the table. We are asked to accept that, if we have a good probability model of the first measurement procedure, and this model is unchanged by permutation of the random variables, then it is also a good probability model for the second procedure. This seems pretty plausible – intuitively, permuting a sequence of random variables seems to accomplish the same thing as permuting the measurement results that these random variables bind to – and its plausibility doesn’t depend on the details of the measurement procedure in question.

On the other hand, the kind of transformation envisioned in causal versions of

exchangeability are of the following nature: suppose you conduct a medical trial that involves administering treatment to a number of patients, and withholding treatment from a number of other patients. Now, consider an alternative procedure: first, you shuffle some patients between the “treatment administered” group and the “withheld” group, then you proceed as before. First, this is not a generic setup! Not all decision problems involve patients that can be shuffled. Secondly, it is not altogether clear that this transformation of the measurement procedure corresponds to a permutation of random variables. Here, we are not merely changing the order in which results are written into a table at the end of the experiment, but altering a seemingly more substantive aspect of the manner in which the experiment is carried out.

In this chapter we discuss *commutativity of exchange*, which is a symmetry of a conditional distribution to permutations of pairs of random variables. This can be understood in terms of the first kind of measurement procedure transformation: in particular, that the appropriate conditional distribution to model the procedure is unchanged by changing the order in which pairs of measurement results are written down. In the following chapter, Section 5.3, we consider modelling transformations like “shuffling patients in a medical experiment”.

4.2 Conditionally independent and identical response functions

Suppose a decision maker is implementing a decision procedure where they’ll make a choice and subsequently receive a sequence of paired values $(\mathcal{D}_i, \mathcal{Y}_i)$, with their objective depending on the output values yielded by \mathcal{Y}_i s only. Usually the \mathcal{D}_i s, which we call “inputs”, are under the decision maker’s control to some extent, but this might not always be the case. For example, perhaps the first m pairs come from data collected by someone else, where the decision maker has no control over inputs, and the next n depend on their own actions where they have complete control over the inputs.

Suppose the decision maker uses a probability set \mathbb{P}_C to model such a procedure, and variables (D_i, Y_i) are associated with the inputs and outputs. There are two different relationships between D_i and Y_i that might be of interest to the decision maker:

- For some choice α , $j > m$ and some fixed value of D_j , what are the *likely consequences* with regard to Y_j ?
- For some choice α , all $i \leq m$ with some fixed value of D_i , what is the *relative frequency* of different values of Y_i ?

The first is what the decision maker wants to know in order to make a good decision, and the second is something they can learn from the data before taking any actions. In particular, if the decision maker has a good reason to think that the two relationships should be (approximately) the same *and* be independent of the decision maker’s overall choice C , then they may reduce the overall problem

of choosing C to the problem of influencing the inputs under their control D_j for $j > m$ toward values that have been associated to with favourable consequences according to the past data.

The conditional independence of consequence Y_i from the choice C given the input D_i is important for this reduction; otherwise the decision maker needs to consider how Y_i depends on C as well as D_i . However, this independence is not required for the results in this chapter, and so we do not assume it. More generally, the results presented here do not show any particular method is appropriate for making decisions, and additional assumptions may be needed for that purpose.

In this chapter, we are interested in models \mathbb{P}_C where the probabilistic relationship between each D_i and the corresponding Y_i is unknown but identical for all indices i . To model this, we introduce a hypothesis H that represents this unknown relationship, and assert that the distribution of Y_i given (D_i, H) is identical for all i , independent of all data prior to i .

Definition 4.2.1 (Conditionally independent and identical response functions). A probability set \mathbb{P}_C on (Ω, \mathcal{F}) with variables $Y := (Y_i)_{i \in \mathbb{N}}$ and $D := (D_i)_{i \in \mathbb{N}}$ has *independent and identical response functions conditional on H* if for all i , $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (D_{[1,i]}, Y_{[1,i]}) | (H, C)$ and $\mathbb{P}_\alpha^{Y_i | D_i H} = \mathbb{P}_\alpha^{Y_j | D_j H}$ for all i, j .

We only require outputs Y_i to be independent of *previous* inputs and outputs, conditional on H and D_i . If D_i is selected based on previous data, then in general there may be relationships between D_j and Y_i for $j > i$ even after conditioning on D_i and H (e.g. D_j is chosen deterministically equal to Y_i for some $j > i$). However, for much of this chapter, we will focus on the simpler case where inputs are *weakly data-independent*, which means that conditional on H , the Y_i are also independent of future inputs. This allows for a kind of “pseudo-dependence” on past data, where inputs may be chosen as if an oracle told the decision maker the value of the usually unknown response function H , but not further depending on any particular previous data values. We explore relaxing this assumption in Section 4.5, although this work is only preliminary.

We show that for weakly data-independent models with conditionally independent and identical response functions, there is some variable W such that the conditional probability $\mathbb{P}_C^{Y|WD}$ is IO contractible. On the other hand, for data-dependent models, we instead require the *comb* (Section 4.5.1) $\mathbb{P}_C^{Y|D|W}$ IO contractible for some W .

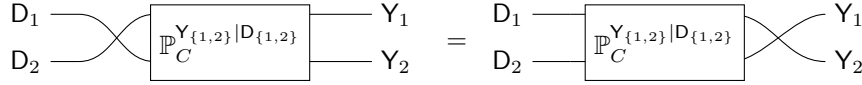
4.3 Symmetries of sequential conditional probabilities

In this section we define key technical terms, including symmetries of conditional probabilities, and prove the technical results IO contractibility and eventually prove the key theorems 4.3.22 and 4.3.27.

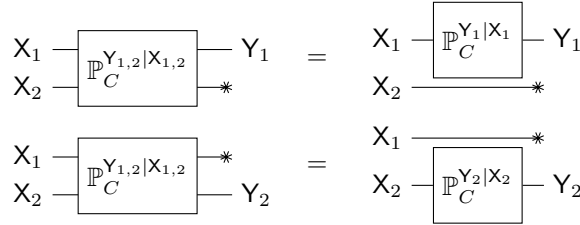
We introduce two basic symmetries: *exchange commutativity* and *locality*. The first says that permutations of a sequence of input-output pairs leaves a conditional probability unchanged, while the second says that the probability of an output does not depend on the value of any non-corresponding inputs. Note that the dependence that is ruled out by locality may be “physical” – for example, herd immunity makes each person’s likelihood of infection depend on the vaccination/recovery status of the rest of the population – or merely “epistemic”, where, for example, many people choosing to eat at one restaurant instead of a neighbouring one is evidence that the first serves better food that can be obtained without every sampling the food from either.

The assumptions of exchange commutativity and locality together make input-output contractibility, or IO contractibility for short. IO contractibility is equivalent to the condition that the conditional probabilities of every equally sized subsequence are equal.

Graphical notation can offer an intuitive picture of these two assumptions. In the simplified case of a sequence of length 2 (that is, $\mathbb{K} : X^2 \rightarrow Y^2$), exchange commutativity for two inputs and outputs is given by the following equality:



swapping the inputs is equivalent to applying the same swap to the outputs. Locality is given by the following pair of equalities:



and expresses the idea that the outputs are independent of the non-corresponding input, conditional on the corresponding input.

The definitions follow.

Call a model \mathbb{P}_C with sequential outputs Y and a corresponding sequence of inputs D a “sequential input-output model”.

Definition 4.3.1 (Sequential input-output model). A *sequential input-output model* is a triple (\mathbb{P}_C, D, Y) where \mathbb{P}_C is a probability set on (Ω, \mathcal{F}) , D is a sequence of “inputs” $D := (D_i)_{i \in \mathbb{N}}$ and Y is a corresponding sequence of “outputs” $Y = (Y_i)_{i \in \mathbb{N}}$ where $D_i : \Omega \rightarrow D$ and $Y_i : \Omega \rightarrow Y$.

Locality holds with respect to some auxiliary variable W when an output i is independent of future inputs, conditioned on the corresponding input i and W .

Definition 4.3.2 (Locality). Given a sequential input-output model (\mathbb{P}_C, D, Y) along with some $W : \Omega \rightarrow W$, for $\alpha \in C$ we say $\mathbb{P}_\alpha^{Y|WD}$ is *local* over W if for all $\alpha \in C$, $n \in \mathbb{N}$

$$\begin{array}{c}
 \begin{array}{ccc}
 & W & \\
 & \swarrow & \\
 D_{[n]} & \text{---} \boxed{\mathbb{P}_\alpha^{Y|WD}} & \text{---} Y_{[n]} \\
 \uparrow & & \uparrow \\
 D_{(n,\infty)} & & *
 \end{array}
 \end{array}
 \quad
 \begin{array}{c}
 \begin{array}{ccc}
 & W & \\
 & \swarrow & \\
 D_{[n]} & \text{---} \boxed{\mathbb{P}_\alpha^{Y|WD_{[n]}}} & \text{---} Y_{[n]} \\
 \uparrow & & \uparrow \\
 D_{(n,\infty)} & & *
 \end{array}
 \end{array}
 \\
 = \\
 \Longleftrightarrow \\
 \mathbb{P}_\alpha^{Y|WD} \left(\bigotimes_{i \in [n]} A_i \times Y^{\mathbb{N}} | w, x_{[n]}, x_{[n]^C} \right) = \mathbb{P}_C^{Y_{[n]}|WD_{[n]}} \left(\bigotimes_{i \in [n]} A_i | w, x_{[n]} \right) \\
 \forall A_i \in \mathcal{Y}, (x_{[n]}, x_{[n]^C}) \in \mathbb{N}, w \in W
 \end{array}$$

That is, $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e X_{(i,\infty)} | (W, X_i, C)$.

Exchange commutativity holds with respect to some auxiliary variable W when swapping input, output pairs doesn't alter the conditional distribution of outputs given inputs.

Definition 4.3.3 (Exchange commutativity). Given a sequential input-output model (\mathbb{P}_C, D, Y) along with some $W : \Omega \rightarrow W$, we say $\mathbb{P}_\alpha^{Y|WD}$ *commutes with exchange* over W if for all finite permutations $\rho : \mathbb{N} \rightarrow \mathbb{N}$

$$\mathbb{P}_\alpha^{Y_\rho|WD_\rho} = \mathbb{P}_\alpha^{Y|WD}$$

IO contractibility is the conjunction of both previous assumptions.

Definition 4.3.4 (IO contractibility). Given a sequential input-output model (\mathbb{P}_C, D, Y) along with some $W : \Omega \rightarrow W$, $\mathbb{P}_\alpha^{Y|WD}$ is *IO contractible* over W if it is local and commutes with exchange.

An input-output model (\mathbb{P}_C, D, Y) with $\mathbb{P}_\alpha^{Y|WD}$ IO contractible over W can be “contracted” to some subsequence of D and Y , and for any two subsequences $A, B \subset \mathbb{N}$, provided the subsequences are of equal length, the distribution of Y_A given W and X_A will be the same as the distribution of Y_B given W and X_B (Theorem 4.3.9). This feature is the motivation for the name *IO contractibility*.

Theorem 4.3.10 shows that exchange commutativity and locality are independent assumptions.

Before these theorems are proved, the following definition and Lemma will prove helpful.

All swaps can be written as a product of transpositions, so proving that a property holds for all finite transpositions is enough to show it holds for all finite swaps. It's useful to define a notation for transpositions.

Definition 4.3.5 (Finite transposition). Given two equally sized sequences $A = (a_i)_{i \in [n]}$, $B = (b_i)_{i \in [n]}$, $A \leftrightarrow B : \mathbb{N} \rightarrow \mathbb{N}$ is the permutation that sends the i th element of A to the i th element of B and vice versa. Note that $A \leftrightarrow B$ is its own inverse.

Lemma 4.3.6 is used to extend conditional probabilities of finite sequences to infinite ones.

Lemma 4.3.6 (Infinitely extended kernels). *Given a collection of Markov kernels $\mathbb{K}_i : W \times X^{\mathbb{N}} \rightarrow Y^i$ for all $i \in \mathbb{N}$, if we have for every $j > i$*

$$\mathbb{K}_j(id_{Y^i} \otimes del_{Y^{j-i}}) = \mathbb{K}_i \otimes del_{X^{j-i}} \quad (4.1)$$

then there is a unique Markov kernel $\mathbb{K} : X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ such that for all $i, j \in \mathbb{N}, j > i$

$$\mathbb{K}(id_{Y^i} \otimes del_{Y^{\mathbb{N}}}) = \mathbb{K}_i \otimes del_{X^{j-i}}$$

Proof. Take any $x \in X^{\mathbb{N}}$ and let $x_{|m} \in X^m$ be the first m elements of x . By Equation (4.1), for any $A_i \in \mathcal{Y}, i \in [m]$

$$\mathbb{K}_n(\bigtimes_{i \in [m]} A_i \times Y^{n-m} | x_{|n}) = \mathbb{K}_m(\bigtimes_{i \in [m]} A_i | x_{|m})$$

Furthermore, by the definition of the swap map for any permutation $\rho : [n] \rightarrow [n]$

$$\mathbb{K}_n \text{swap}_{\rho}(\bigtimes_{i \in [m]} A_{\rho(i)} \times Y^{n-m} | x_{|n}) = \mathbb{K}_n(\bigtimes_{i \in [m]} A_i \times Y^{n-m} | x_{|n})$$

thus by the Kolmogorov Extension Theorem (Çinlar, 2011), for each $x \in X^{\mathbb{N}}$ there is a unique probability measure $\mathbb{Q}_x \in \Delta(Y^{\mathbb{N}})$ satisfying

$$\mathbb{Q}_d(\bigtimes_{i \in [n]} A_i \times Y^{\mathbb{N}}) = \mathbb{K}_n(\bigtimes_{i \in [n]} A_{\rho(i)} | d_{|n}) \quad (4.2)$$

Furthermore, for each $\{A_i \in \mathcal{Y} | i \in \mathbb{N}\}$, $n \in \mathbb{N}$ note that for $p > n$

$$\begin{aligned} \mathbb{Q}_d(\bigtimes_{i \in [n]} A_i \times Y^{\mathbb{N}}) &\geq \mathbb{Q}_d(\bigtimes_{i \in [p]} A_i \times Y^{\mathbb{N}}) \\ &\geq \mathbb{Q}_d(\bigtimes_{i \in \mathbb{N}} A_i) \end{aligned}$$

so by the Monotone convergence theorem, the sequence $\mathbb{Q}_d(\bigtimes_{i \in [n]} A_i)$ converges as $n \rightarrow \infty$ to $\mathbb{Q}_d(\bigtimes_{i \in \mathbb{N}} A_i)$. $d \mapsto \mathbb{Q}_d^Z(\bigtimes_{i \in [n]} A_i)$ is measurable for all n , $\{A_i \in \mathcal{Y} | i \in \mathbb{N}\}$ by Equation (4.2), and so $d \mapsto \mathbb{Q}_d$ is also measurable.

Thus $d \mapsto \mathbb{Q}_d$ is the desired $\mathbb{P}_C^{Y^{\mathbb{N}} | D^{\mathbb{N}}} : D^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$. \square

Corollary 4.3.7. *Given $(\mathbb{P}_C, \Omega, \mathcal{F})$, $V : \Omega \rightarrow V$ and two pairs of sequences $(W, X) := (W_i, X_i)_{i \in \mathbb{N}}$ and $(Y, Z) := (Y_i, Z_i)_{i \in \mathbb{N}}$ with corresponding variables taking values in the same sets $W = Y$ and $X = Z$, if (\mathbb{P}_C, W, X) and (\mathbb{P}_C, Y, Z) are both local over W and*

$$\mathbb{P}^{X_{[n]} | VW_{[n]}} = \mathbb{P}^{Z_{[n]} | VY_{[n]}}$$

for all $n \in \mathbb{N}$ then

$$\mathbb{P}^{X | VW} = \mathbb{P}^{Z | VY}$$

Proof. By assumption of locality

$$\begin{aligned}\mathbb{P}^{X_{[n]}|VW_{[n]}} \otimes \text{del}_{W^{\mathbb{N}}} &= \mathbb{P}^{X|VW}(\text{id}_{X^n} \otimes \text{del}_{X^{\mathbb{N}}}) \\ \mathbb{P}^{Z_{[n]}|VY_{[n]}} \otimes \text{del}_{W^{\mathbb{N}}} &= \mathbb{P}^{Z|VY}(\text{id}_{X^n} \otimes \text{del}_{X^{\mathbb{N}}})\end{aligned}$$

hence for all $n, m > n$

$$\begin{aligned}\mathbb{P}^{X_{[m]}|VW_{[m]}}(\text{id}_{X^n} \otimes \text{del}_{X^{m-n}}) &= \mathbb{P}^{Z_{[m]}|VY_{[m]}}(\text{id}_{X^n} \otimes \text{del}_{X^{m-n}}) \\ &= \mathbb{P}^{X_{[n]}|VW_{[n]}} \otimes \text{del}_{W^{m-n}}\end{aligned}$$

and, in particular, by lemma 4.3.6, $\mathbb{P}^{X|VW}$ and $\mathbb{P}^{Z|VY}$ are the limits of the same sequence. \square

Lemma 4.3.8 (Alternative definition of exchange commutativity). *A sequential input-output model (\mathbb{P}_C, D, Y) along with some $W : \Omega \rightarrow W$ commutes with exchange over W if and only if for every α , every finite permutation $\rho : \mathbb{N} \rightarrow \mathbb{N}$ and corresponding swap map $\text{swap}_\rho : X^{\mathbb{N}} \rightarrow X^{\mathbb{N}}$*

$$\begin{array}{c} W \\ D \end{array} \begin{array}{|c|} \hline \mathbb{P}_\alpha^{Y|WD} \\ \hline \end{array} \begin{array}{c} \text{---} Y \end{array} = \begin{array}{c} W \\ D \end{array} \begin{array}{|c|} \hline \text{swap}_{\rho^{-1}} \\ \hline \end{array} \begin{array}{|c|} \hline \mathbb{P}_\alpha^{Y|WD} \\ \hline \end{array} \begin{array}{|c|} \hline \text{swap}_\rho \\ \hline \end{array} \begin{array}{c} \text{---} Y \end{array}$$

Proof. This follows from the fact that

$$\mathbb{P}_\alpha^{Y_\rho|WD_\rho} = \begin{array}{c} W \\ D \end{array} \begin{array}{|c|} \hline \text{swap}_{\rho^{-1}} \\ \hline \end{array} \begin{array}{|c|} \hline \mathbb{P}_\alpha^{Y|WD} \\ \hline \end{array} \begin{array}{|c|} \hline \text{swap}_\rho \\ \hline \end{array} \begin{array}{c} \text{---} Y \end{array}$$

To see this, note that

$$\begin{aligned}& \begin{array}{c} W \\ D \end{array} \begin{array}{|c|} \hline \text{swap}_{\rho^{-1}} \\ \hline \end{array} \begin{array}{|c|} \hline \mathbb{P}_\alpha^{Y|WD} \\ \hline \end{array} \begin{array}{|c|} \hline \text{swap}_\rho \\ \hline \end{array} \begin{array}{c} \text{---} Y \end{array} \left(\bigotimes_{i \in \mathbb{N}} A_i | w, (d_i)_{\mathbb{N}} \right) \\ &= \mathbb{P}_\alpha^{Y|WD} \left(\bigotimes_{i \in \mathbb{N}} A_{\rho^{-1}(i)} | w, (d_{\rho^{-1}(i)})_{\mathbb{N}} \right) \\ &= \mathbb{P}_\alpha^{Y_\rho|WD_\rho} \left(\bigotimes_{i \in \mathbb{N}} A_i | w, (d_i)_{\mathbb{N}} \right)\end{aligned}$$

\square

Theorem 4.3.9 provides an alternative characterization of IO contractibility in terms of the equality of the conditional distributions of sub-sequences.

Theorem 4.3.9 (Equality of equally sized contractions). *A sequential input-output model (\mathbb{P}_C, D, Y) is IO contractible over W if and only if for and every subsequence $A = (A_i)_{i \in |A|}$ and $B = (B_i)_{i \in |A|}$ with $i \neq j \implies A_i \neq A_j$ and $B_i \neq B_j$ and every α*

$$\begin{aligned}\mathbb{P}_\alpha^{Y_A|WD_{A \leftrightarrow [|A|]}} &= \mathbb{P}_\alpha^{Y_B|WD_{B \leftrightarrow [|A|]}} \\ &= \mathbb{P}_\alpha^{Y_A|WD_A} \otimes \text{del}_{D^{|\mathbb{N} \setminus A|}}\end{aligned}$$

where $[|A|]$ is the sequence $(1, 2, \dots, |A|)$ for finite A or $(1, 2, \dots)$ for infinite A .

Proof. Only if: If A is finite, then by exchange commutativity

$$\begin{aligned}\mathbb{P}_\alpha^{Y_A|\text{WD}_{A \leftrightarrow [n]}} &= \mathbb{P}_\alpha^{Y_{[n]}|\text{WD}} \\ &= \mathbb{P}_\alpha^{Y_B|\text{WD}_{B \leftrightarrow [n]}}\end{aligned}$$

if A is infinite, then we can take finite subsequences A_m that are the first m elements of A . Then

$$\begin{aligned}\mathbb{P}_\alpha^{Y_{A_m}|\text{WD}_{A_m \leftrightarrow [m]}} &= \mathbb{P}_\alpha^{Y_{[m]}|\text{WD}} \\ &= \mathbb{P}_\alpha^{Y_{B_m}|\text{WD}_{B_m \leftrightarrow [m]}}\end{aligned}$$

then by Corollary 4.3.7

$$\mathbb{P}_\alpha^{Y_A|\text{WD}_{A \leftrightarrow [n]}} = \mathbb{P}_\alpha^{Y_{B_m}|\text{WD}_{B_m \leftrightarrow [m]}}$$

By locality

$$\mathbb{P}_\alpha^{Y_A|\text{WD}_{A \leftrightarrow [n]}} = \mathbb{P}_\alpha^{Y_A|\text{WD}_A} \otimes \text{del}_{D|N \setminus A}$$

If: Taking $A = [n]$ for all n establishes locality, and taking $A = (\rho(i))_{i \in \mathbb{N}}$ for arbitrary finite permutation ρ establishes exchange commutativity. \square

Theorem 4.3.10 shows that neither locality nor exchange commutativity is implied by the other.

Theorem 4.3.10. *Exchange commutativity does not imply locality or vice versa.*

Proof. First, a model that exhibits exchange commutativity but not locality. Suppose $D = Y = \{0, 1\}$ and $\mathbb{P}_C^{Y|D} : D^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ is given by

$$\mathbb{P}_C^{Y|D}(\bigtimes_{i \in \mathbb{N}} A_i | (d_i)_{i \in \mathbb{N}}) = \prod_{i \in \mathbb{N}} \delta_{\lim_{n \rightarrow \infty} \sum_{i=1}^n \frac{d_i}{n}}(A_i)$$

then

$$\begin{aligned}\mathbb{P}_C^{Y_\rho|D_\rho}(\bigtimes_{i \in \mathbb{N}} A_i | (d_i)_{i \in \mathbb{N}}) &= \prod_{i \in \mathbb{N}} \delta_{\lim_{n \rightarrow \infty} \sum_{i=1}^n \frac{d_{\rho^{-1}(i)}}{n}}(A_{\rho^{-1}(i)}) \\ &= \mathbb{P}_C^{Y|D}(\bigtimes_{i \in \mathbb{N}} A_i | (d_i)_{i \in \mathbb{N}})\end{aligned}$$

so (\mathbb{P}_C, D, Y) commutes with exchange, but

$$\begin{aligned}\mathbb{P}_C^{Y_1|D}(\bigtimes_{i \in \mathbb{N}} A_i | 0, 1, 1, 1, \dots) &= \delta_1(A_i) \\ \mathbb{P}_C^{Y_1|D}(\bigtimes_{i \in \mathbb{N}} A_i | 0, 0, 0, 0, \dots) &= \delta_0(A_i)\end{aligned}$$

so (\mathbb{P}_C, D, Y) is not local.

Next, a Markov kernel that satisfies locality but does not commute with exchange. Suppose again $D = Y = \{0, 1\}$ and $\mathbb{P}_C^{Y|D} : D^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ is given by

$$\mathbb{P}_C^{Y|D}(\bigtimes_{i \in \mathbb{N}} A_i | (d_i)_{i \in \mathbb{N}}) = \prod_{i \in \mathbb{N}} \delta_i(A_i)$$

then

$$\begin{aligned} \mathbb{P}_C^{Y_{\rho}|D_{\rho}}(\bigtimes_{i \in \mathbb{N}} A_i | (d_i)_{i \in \mathbb{N}}) &= \prod_{i \in \mathbb{N}} \delta_i(A_{\rho^{-1}(i)}) \\ &\neq \prod_{i \in \mathbb{N}} \delta_i(A_i) \\ &= \mathbb{P}_C^{Y|D}(\bigtimes_{i \in \mathbb{N}} A_i | (d_i)_{i \in \mathbb{N}}) \end{aligned}$$

so (\mathbb{P}_C, D, Y) does not commute with exchange but for all n

$$\begin{aligned} \mathbb{P}_C^{Y_{[n]}|D}(\bigtimes_{i \in [n]} A_i | (d_i)_{i \in \mathbb{N}}) &= \prod_{i \in [n]} \delta_i(A_{\rho^{-1}(i)}) \\ &= \mathbb{P}_C^{Y_{[n]}|D}(\bigtimes_{i \in [n]} A_i | (0)_{i \in \mathbb{N}}) \end{aligned}$$

so (\mathbb{P}_C, D, Y) is local. □

Theorem 4.3.10 presents abstract counterexamples to show that the assumptions of exchange commutativity and locality are independent. For some more practical examples, a model of the treatment of several patients who are known to have different illnesses might satisfy consequence locality but not exchange commutativity. Patient B's treatment can be assumed not to affect patient A, but the same results would not be expected from giving patient A's treatment to patient B as from giving patient A's treatment to patient A.

A model of strategic behaviour could satisfy exchange commutativity but not locality. Suppose a decision maker is observing people playing a game where they press a red or green button, and (for reasons mysterious to the decision maker), receive a payout randomly of 0 or \$100. The decision maker might reason that the results should be the same no matter who presses a button, but also that people will be more likely to press the red button if the red button tends to give a higher payout. In this case, the decision maker's prediction for the payout of the i th attempt given the red button has been pressed will be higher if the proportion of red button presses in the entire dataset is higher. There are other reasons why exchange commutativity might hold but not locality – Dawid (2000) offers the alternative example of herd immunity in vaccination campaigns. In this case, the overall proportion of the population vaccinated will affect the disease prevalence over and above an individual's vaccination status.

Although locality could be described as an assumption that there is no interference between inputs and outputs of different indices, it actually allows

for some models with certain kinds of interference between inputs and non-corresponding. For example: consider an experiment where I first flip a coin and record the results of this flip as the outcome Y_1 of “step 1”. Subsequently, I can either copy the outcome from step 1 to the result for “step 2” (this is the input $D_1 = 0$), or flip a second coin use this as the input for step 2 (this is the input $D_1 = 1$). D_2 is an arbitrary single-valued variable. Then for all d_1, d_2

$$\begin{aligned}\mathbb{P}^{Y_1|D}(y_1|d_1, d_2) &= 0.5 \\ \mathbb{P}^{Y_2|D}(y_2|d_1, d_2) &= 0.5\end{aligned}$$

Thus the marginal distribution of both experiments in isolation is Bernoulli(0.5) no matter what choices I make, but the input D_1 affects the joint distribution of the results of both steps, which is not ruled out by locality.

4.3.1 Representation of IO contractible models

Theorem 4.3.18 shows that a IO contractible conditional distribution can be represented as the product of a column exchangeable probability distribution and a “lookup function” or “switch”. This lookup function is also used in the representation of potential outcomes models (see, for example, Rubin (2005)), but potential outcomes also carry an interpretation that is absent here.. This theorem allows De Finetti’s theorem to be applied to the column exchangeable probability distribution, which is a key step in proving the main result (Theorem 4.3.22).

We reuse a number of concepts in the following work: models with sequences of inputs and outputs, “tabulated” representations of conditional probabilities and “hypothesis” or “directing measures” defined as the limit of relative frequencies.

Definition 4.3.11 (Count of input values). Given a sequential input-output model (\mathbb{P}_C, D, Y) on (Ω, \mathcal{F}) with countable D , $\#_j^k$ is the variable

$$\#_j^k := \sum_{i=1}^{k-1} \mathbb{I}[D_i = j]$$

In particular, $\#_j^k$ is equal to the number of times $D_i = j$ over all $i < k$.

Definition 4.3.12 (Tabulated conditional distribution). Given a sequential input-output model (\mathbb{P}_C, D, Y) on (Ω, \mathcal{F}) , define the tabulated conditional distribution $Y^D : \Omega \rightarrow Y^{\mathbb{N} \times D}$ by

$$Y_{ij}^D = \sum_{k=1}^{\infty} \mathbb{I}[\#_j^k = i - 1] \mathbb{I}[D_k = j] Y_k$$

That is, the (i, j) -th coordinate of $Y^D(\omega)$ is equal to the coordinate $Y_k(\omega)$ for which the corresponding $D_k(\omega)$ is the i th instance of the value j in the sequence $(D_1(\omega), D_2(\omega), \dots)$, or 0 if there are fewer than i instances of j in this sequence.

Definition 4.3.13 (Measurable set of probability distributions). Given a measurable set (Ω, \mathcal{F}) , the measurable set of distributions on Ω , $\mathcal{M}_1(\Omega)$, is the set of all probability distributions on Ω equipped with the coarsest σ -algebra such that the evaluation maps $\eta_B : \nu \mapsto \nu(B)$ are measurable for all $B \in \mathcal{F}$.

We define the *directing random measure* of a sequence of variables as the limit of normalised partial sums of variables in the sequence. We refer to directing random measures with the letter H by default. We also define H in the case that the relevant limit does not exist for completeness, although we are only interested in cases where it is well-defined. Definition 4.3.14 reduces to the definition of a directing random measure given in Kallenberg (2005a) when we consider a probability space instead of a probability set.

Definition 4.3.14 (Directing random measure). Given a probability set $(\mathbb{P}_C, \Omega, \mathcal{F})$ and a sequence $\mathbf{X} := (\mathbf{X}_i)_{i \in \mathbb{N}}$, the directing random measure of \mathbf{X} written $H : \Omega \rightarrow \mathcal{M}_1(X)$ is the function

$$H := A \mapsto \begin{cases} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbb{1}_A(\mathbf{X}_i) & \text{this limit exists for all } A \in \mathcal{F} \\ \llbracket A = X \rrbracket & \text{otherwise} \end{cases}$$

Given two variable sequences (D, Y) , which we call the inputs and outputs respectively, we define the *directing random conditional* as the directing random measure of the “tabulated conditional” Y^D , interpreted as a sequence of column vectors $((Y_{1j}^D)_{j \in D}, (Y_{2j}^D)_{j \in D}, \dots)$.

Definition 4.3.15 (Directing random conditional). Given a sequential input-output model (\mathbb{P}_C, D, Y) , we will say the directing random measure $H : \Omega \rightarrow \mathcal{M}_1(Y^D)$ is the function

$$H := \bigotimes_{j \in D} A_j \mapsto \begin{cases} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \prod_{j \in D} \mathbb{1}_{A_j}(Y_{ij}^D) & \text{this limit exists} \\ \llbracket \bigotimes_{j \in D} A_j = Y^D \rrbracket & \text{otherwise} \end{cases}$$

We say a model observes data-independence when future inputs are independent of outputs conditional on past inputs and the directing measure H .

Definition 4.3.16 (Data-independent). A sequential input-output model (\mathbb{P}_C, D, Y) is weakly data-independent if $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e D_{(i, \infty]} | (H, D_{[1, i]}, C)$.

A finite permutation of rows is a function that independently permutes a finite number of elements in each row of a table. A special case of such a function is one that swaps entire columns (that is, a permutation of rows that applies the same permutation to each row).

Definition 4.3.17 (Permutation of rows). Given a sequence of indices $(i, j)_{i \in \mathbb{N}, j \in D}$ a finite permutation of rows is a function $\eta : \mathbb{N} \times D \rightarrow \mathbb{N} \times D$ such that for each $j \in D$, $\eta_j := \eta(\cdot, j)$ is a finite permutation $\mathbb{N} \rightarrow \mathbb{N}$ and $\eta(i, j) = (\eta_j(i), j)$.

There is an assumption in the following theorem that the set of input sequences in which each value appears infinitely often has measure 1, which is needed in the main Theorem 4.3.22 to guarantee that the hypothesis \mathbf{H} is a function of the observable variables.

Theorem 4.3.18. *Suppose a sequential input-output model (\mathbb{P}_C, D, Y) is given with D countable and, letting $E \subset D^{\mathbb{N}}$ be the set of all sequences for which each $j \in D$ occurs infinitely often, $\mathbb{P}_\alpha^D(E) = 1$ for all α . Then for some W , $\mathbb{P}_\alpha^{Y|WD}$ is IO contractible if and only if for all α*

$$\begin{aligned} \mathbb{P}_\alpha^{Y|WD} &= \begin{array}{c} W \\ \boxed{\mathbb{P}_\alpha^{Y^D|W}} \\ D \end{array} \xrightarrow{\quad} \boxed{\mathbb{F}_{lu}} \xrightarrow{\quad} Y \\ &\iff \\ \mathbb{P}_\alpha^{Y|WD} \left(\bigotimes_{i \in \mathbb{N}} A_i | w, (d_i)_{i \in \mathbb{N}} \right) &= \mathbb{P}_\alpha^{(Y_{id_i}^D)_{i \in \mathbb{N}} | W} \left(\bigotimes_{i \in \mathbb{N}} A_i | w \right) \quad \forall A_i \in \mathcal{Y}^{\mathbb{N}}, w \in W, d_i \in D \end{aligned}$$

Where \mathbb{F}_{lu} is the Markov kernel associated with the lookup map

$$\begin{aligned} lu : X^{\mathbb{N}} \times Y^{\mathbb{N} \times D} &\rightarrow Y \\ ((x_i)_{i \in \mathbb{N}}, (y_{ij})_{i, j \in \mathbb{N} \times D}) &\mapsto (y_{id_i})_{i \in \mathbb{N}} \end{aligned}$$

and for any finite permutation of rows $\eta : \mathbb{N} \times D \rightarrow \mathbb{N} \times D$

$$\mathbb{P}_\alpha^{(Y_{ij}^D)_{i \in \mathbb{N}, j \in D} | W} = \mathbb{P}_\alpha^{(Y_{\eta(i,j)}^D)_{i \in \mathbb{N}, j \in D} | W} \quad (4.3)$$

Proof. Only if: Note that at most one of $\llbracket \#_j^k = i - 1 \rrbracket \llbracket D_k = j \rrbracket$ and $\llbracket \#_j^l = i - 1 \rrbracket \llbracket D_l = j \rrbracket$ can be greater than 0 for $k \neq l$ and, by assumption, $\sum_{j \in D} \sum_{k \in \mathbb{N}} \llbracket \#_j^k = i - 1 \rrbracket \llbracket D_k = j \rrbracket = 1$ almost surely (that is, for any i, j there is some k such that D_k is the i th occurrence of j). Define $R_k : \Omega \rightarrow \mathbb{N} \times D$ by $\omega \mapsto \arg \max_{i \in \mathbb{N}, j \in D} \llbracket \#_j^k = i - 1 \rrbracket \llbracket D_k = j \rrbracket(\omega)$ (i.e. R_k returns the (i, j) pair where j is the value of D_k and i is the count of j occurrences up to D_k). Let $R : \mathbb{N} \rightarrow \mathbb{N} \times D$ by $k \mapsto R_k$. R is almost surely bijective and

$$\begin{aligned} Y^D &:= (Y_{ij}^D)_{i \in \mathbb{N}, j \in D} \\ &= (Y_{R^{-1}(i,j)}^D)_{i \in \mathbb{N}, j \in D} \\ &=: Y_{R^{-1}|W} \end{aligned}$$

By construction, $D_{R^{-1}(i,j)} = j$ almost surely; that is, $D_{R^{-1}}$ is a single-valued variable. In particular, it is almost surely equal to $e := (e_{ij})_{i \in \mathbb{N}, j \in D}$ such that $e_{ij} = j$ for all i . Hence

$$\begin{aligned} \mathbb{P}_\alpha^{Y^D | WD_{R^{-1}}} (A | w, d) &= \mathbb{P}_\alpha^{Y_{R^{-1}} | WD_{R^{-1}}} (A | w, d) \\ &\stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_\alpha^{Y_{R^{-1}} | WD_{R^{-1}}} (A | e, w) \end{aligned} \quad (4.4)$$

$$= \mathbb{P}_\alpha^{Y^D} (A | w) \quad (4.5)$$

for any $d \in D^{\mathbb{N}}$. Equation (4.4) implies $Y^D \perp\!\!\!\perp D|(W, C)$.

Now,

$$\mathbb{P}_{\alpha}^{Y_{R^{-1}}|WD_{R^{-1}}}(A|w, d) = \int_R \mathbb{P}_{\alpha}^{Y_{\rho}|WD_{\rho}}(A|d) \mathbb{P}_{\alpha}^{R^{-1}|WD_{R^{-1}}}(\mathrm{d}\rho|w, d) \quad (4.6)$$

For each ρ , define $\rho^n : \mathbb{N} \rightarrow \mathbb{N}$ as the finite permutation that agrees with ρ on the first n indices and is the identity otherwise. By IO contractibility, for $n \in \mathbb{N}$

$$\begin{aligned} \mathbb{P}^{Y_{\rho^n([n])}|D_{\rho^n([n])}} &= \mathbb{P}^{Y_{\rho([n])}|D_{\rho([n])}} \\ &= \mathbb{P}^{Y_{[n]}|D_{[n]}} \end{aligned}$$

By Corollary 4.3.7, it must therefore be the case that

$$\mathbb{P}^{Y|D} = \mathbb{P}^{Y_{\rho}|D_{\rho}}$$

Then from Equation (4.6)

$$\begin{aligned} \mathbb{P}_{\alpha}^{Y_{R^{-1}}|WD_{R^{-1}}}(A|w, d) &\stackrel{\mathbb{P}_C}{\cong} \int_R \mathbb{P}_{\alpha}^{Y_{\rho}|WD_{\rho}}(A|d) \mathbb{P}_{\alpha}^{R^{-1}|WD_{R^{-1}}}(\mathrm{d}\rho|w, d) \\ &\stackrel{\mathbb{P}_C}{\cong} \int_R \mathbb{P}_C^{Y|WD}(A|w, d) \mathbb{P}_{\alpha}^{R^{-1}|WD_{R^{-1}}}(\mathrm{d}\rho|w, d) \\ &\stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y|WD}(A|w, d) \end{aligned} \quad (4.7)$$

for all $i, j \in \mathbb{N}$. Then by Equation (4.5) and Equation (4.7)

$$\mathbb{P}_{\alpha}^{Y^D|W}(A|w) = \mathbb{P}_{\alpha}^{Y|WD}(A|w, e) \quad (4.8)$$

Take some $d \in D^{\mathbb{N}}$. From Equation (4.8) and IO contractibility of $\mathbb{P}_C^{Y|WD}(A|e)$,

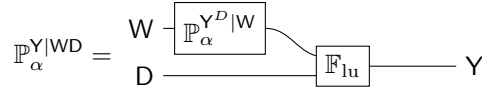
$$\begin{aligned} (\mathbb{P}_{\alpha}^{Y^D|W} \otimes \mathrm{id}_D) \mathbb{F}_{tu}(A|w, d) &= \mathbb{P}_{\alpha}^{(Y_{id_i}^D)_{i \in \mathbb{N}}|W}(A|d) \\ &= \mathbb{P}_{\alpha}^{(Y_{id_i})_{i \in \mathbb{N}}|WD}(A|w, e) \\ &= \mathbb{P}_{\alpha}^{(Y_{id_i})_{i \in \mathbb{N}}|W(D_{id_i})_{i \in \mathbb{N}}}(A|w, (e_{id_i})_{i \in \mathbb{N}}) \\ &= \mathbb{P}_{\alpha}^{Y|WD}(A|w, (e_{id_i})_{i \in \mathbb{N}}) \\ &= \mathbb{P}_{\alpha}^{Y|WD}(A|w, (d_i)_{i \in \mathbb{N}}) \end{aligned}$$

Furthermore, for some finite permutation within columns $\eta : \mathbb{N} \times D \rightarrow \mathbb{N} \times D$,

note that $e_{\eta(i,j)} = j$ and hence $(e_{\eta(i,j)})_{i \in \mathbb{N}, j \in D} = e$. Thus

$$\begin{aligned}
\mathbb{P}_\alpha^{(Y_{\eta(i,j)}^D)^{\mathbb{N} \times D} | W} (A|w) &= \mathbb{P}_\alpha^{(Y^D)^{\mathbb{N} \times D} | W} \text{swap}_\eta(A|w) \\
&= \mathbb{P}_\alpha^{Y | WD} \text{swap}_\eta(A|w, e) && \text{from Eq. (4.8)} \\
&= \mathbb{P}_\alpha^{Y_\eta | WD} (A|w, e) \\
&= \mathbb{P}_\alpha^{Y | WD_{\eta^{-1}}} (A|w, e) && \text{by exchange commutativity} \\
&= \mathbb{P}_\alpha^{Y | WD} (A|w, (e_{\eta^{-1}(i,j)})_{i \in \mathbb{N}, j \in D}) \\
&= \mathbb{P}_\alpha^{Y | WD} (A|w, e) \\
&= \mathbb{P}_\alpha^{(Y_{ij}^D)^{\mathbb{N} \times D} | W} (A|w) && \text{from Eq. (4.8)}
\end{aligned}$$

If: Suppose



where $\mathbb{P}_\alpha^{Y^D | W}$ satisfies Equation (4.3).

Consider any two $d, d' \in D^{\mathbb{N}}$ such that for some $S, T \subset \mathbb{N}$ with $|S| = |T| = n$, $d_S = d'_T$. Let $S \leftrightarrow T$ be the transposition that swaps the i th element of S with the i th element of T for all i .

$$\begin{aligned}
\mathbb{P}_\alpha^{Y_S | WD} \left(\bigtimes_{i \in [n]} A_i | w, d \right) &= \mathbb{P}_\alpha^{(Y_{id_i}^D)_{i \in S} | W} \left(\bigtimes_{i \in [n]} A_i | w \right) \\
&= \mathbb{P}_\alpha^{(Y_{S \leftrightarrow T(i)d_i}^D)_{i \in S} | W} \left(\bigtimes_{i \in [n]} A_i | w \right) \\
&= \mathbb{P}_\alpha^{(Y_{id_{S \leftrightarrow T(i)}}^D)_{i \in T} | W} \left(\bigtimes_{i \in [n]} A_i | w \right) \\
&= \mathbb{P}_\alpha^{(Y_{id'_i}^D)_{i \in T} | W} \left(\bigtimes_{i \in [n]} A_i | w \right) \\
&= \mathbb{P}_\alpha^{Y_T | WD} \left(\bigtimes_{i \in [n]} A_i | w, d' \right)
\end{aligned}$$

and, in particular, taking $T = [n]$

$$= \mathbb{P}_\alpha^{Y_{[n]} | WD} \left(\bigtimes_{i \in [n]} A_i | w, d' \right)$$

but d' is an arbitrary sequence such that the T elements match the S elements of d , so this holds for any other d'' whose T elements also match the S elements

of d . That is

$$\mathbb{P}_\alpha^{Y_S | \text{WD}}(\bigotimes_{i \in [n]} A_i | w, d) = (\mathbb{P}_\alpha^{Y_{[n]} | \text{WD}_{[n]}} \otimes \text{del}_{D^\mathbb{N}})(\bigotimes_{i \in [n]} A_i | w, d')$$

so \mathbb{K} is IO contractible by Theorem 4.3.9. \square

As a consequence of Theorem 4.3.18 along with De Finetti's representation theorem, we can say that given (\mathbb{P}_C, D, Y) IO contractible, the columns of Y^D are independent conditional on the directing random conditional H .

Lemma 4.3.19. *Suppose a sequential input-output model (\mathbb{P}_C, D, Y) is given with D countable and, letting $E \subset D^\mathbb{N}$ be the set of all sequences for which each $j \in D$ occurs infinitely often, $\mathbb{P}_\alpha^D(E) = 1$ for all α and for some W , $\mathbb{P}_\alpha^{Y | \text{WD}}$ is IO contractible. Then, letting H be the directing random conditional of (\mathbb{P}_C, D, Y) (Definition 4.3.15) and $Y_{iD}^D := (Y_{ij}^D)_{j \in D}$, we have for all $i \in \mathbb{N}$, $Y_{iD}^D \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{N \setminus \{i\}D}^D, W) | (H, C)$ and $\mathbb{P}_\alpha^{Y_{iD}^D} = \mathbb{P}_\alpha^{Y_{kD}^D}$ and*

$$\mathbb{P}_\alpha^{Y_{iD}^D | H}(A | \nu) \cong \nu(A)$$

Proof. Fix $w \in W$ and consider $\mathbb{P}_{\alpha, w}^{Y^D} := \mathbb{P}_\alpha^{Y^D | W}(\cdot | w)$. From Theorem 4.3.18, we have the exchangeability of the sequence $(Y_{1D}^D, Y_{2D}^D, \dots)$ with respect to $(\mathbb{P}_{\alpha, w}, \Omega, \mathcal{F})$ as a special case of the invariance of $\mathbb{P}_\alpha^{(Y_{ij}^D)_{N \times D} | W}$ to permutations of rows. By the column exchangeability of $\mathbb{P}_{\alpha, w}^{Y^D}$, from Kallenberg (2005a, Prop. 1.4) (where H is what Kallenberg calls the directing random measure)

$$\mathbb{P}_{\alpha, w}^{Y^D | H} = H \longrightarrow \begin{array}{c} \bullet \\ \boxed{\mathbb{P}_{S_0 | H} - S_i} \\ i \in \mathbb{N} \end{array}$$

Hence

$$\mathbb{P}_\alpha^{Y^D | HW} = H \longrightarrow \begin{array}{c} \bullet \\ \boxed{\mathbb{P}_{S_0 | H} - S_i} \\ i \in \mathbb{N} \end{array}$$

$W \longrightarrow *$

which yields $Y^D \perp\!\!\!\perp_{\mathbb{P}_C}^e W | (H, C)$. Further application of Kallenberg (2005a, Prop. 1.4) yields $Y_{iD}^D \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{N \setminus \{i\}D}^D, W) | (H, C)$ and

$$\mathbb{P}_\alpha^{Y_{iD}^D | H}(A | \nu) \cong \nu(A)$$

\square

If the conditions of Theorem 4.3.18 are satisfied, we do not need the full sequence of pairs (D, Y) to calculate H , any subsequence that satisfies the condition that each value of D occurs infinitely often is sufficient.

Theorem 4.3.20. *Suppose a sequential input-output model (\mathbb{P}_C, D, Y) is given with D countable. Consider an infinite set $A \subset \mathbb{N}$, and let $D_A := (D_i)_{i \in A}$ and $Y_A := (Y_i)_{i \in A}$, and letting $E \subset D^{\mathbb{N}}$ be the set of all sequences for which each $j \in D$ occurs infinitely often, suppose $\mathbb{P}_\alpha^{D_A}(E) = 1$ for all α . Suppose also that for some W , $\mathbb{P}^{Y|WD}$ is IO contractible. Then H_A , the directing random conditional of (\mathbb{P}_C, D_A, Y_A) is almost surely equal to H , the directing random conditional of (\mathbb{P}_C, D, Y) .*

Proof. The strategy we will pursue is to show that an arbitrary subsequence of (D_i, Y_i) pairs induces a random contraction of the rows of Y^D . Then we show that the contracted version of Y^D has the same distribution as the original, and consequently the normalised partial sums converge to the same limit.

Define $Y^{D,A}$ as the tabulated conditional of (D_A, Y_A) , i.e. let $\#_j^{A,k}$ be the count restricted to A :

$$\#_j^{A,k} := \sum_{i \in A}^{k-1} \mathbb{I}[D_i = j]$$

then

$$\begin{aligned} Y_{ij}^{D,A} &:= \sum_{k \in A} \mathbb{I}[\#_j^{A,k} = i - 1] \mathbb{I}[D_k = j] Y_k \\ &= \sum_{k \in A} \mathbb{I}[\#_j^{A,k} = i - 1] \mathbb{I}[D_k = j] Y_{R_k j}^D \end{aligned}$$

That is, defining $Q : \mathbb{N} \rightarrow \mathbb{N}$ by $i \mapsto \sum_{k \in A} \mathbb{I}[\#_j^{A,k} = i - 1] \mathbb{I}[D_k = j] R_k$ then

$$Y_{ij}^{D,A} = Y_{Q(i)j}^D \quad (4.9)$$

where $Q(i) \in \mathbb{N}$ by the assumption that each value of D occurs infinitely often in A (otherwise $Q(i)$ might be 0).

Equation (4.9) is what is meant by “the subsequence (D_A, Y_A) induces a random contraction over the rows of Y^D ”. We will now show that $Y^{D,A}$ has the same distribution as Y^D .

Let $\text{con}_q : Y^{\mathbb{N} \times D} \rightarrow Y^{\mathbb{N} \times D}$ be the Markov kernel associated with the function that sends $(Y_{ij}^D)_{i \in \mathbb{N}, j \in D}$ to $(Y_{q(i)j}^D)_{i \in \mathbb{N}, j \in D}$. Then for any $B \in \mathcal{Y}^{\mathbb{N} \times D}$, w, q :

$$\begin{aligned} \mathbb{P}_\alpha^{Y^{D,A}|WQ}(B|w, q) &= \mathbb{P}_\alpha^{Y^D|W} \text{con}_q(B|w) \\ &= \mathbb{P}^{Y|WD} \text{con}_q(B|w, e) && \text{by Eq.(4.8)} \\ &= \mathbb{P}^{Y|WD}(B|w, e) && \text{by Theorem 4.3.9} \\ &= \mathbb{P}_\alpha^{Y^D|W}(B|w) && \text{by Eq.(4.8)} \end{aligned} \quad (4.10)$$

Finally, take H^A the directing random measure of $Y^{D,A}$. We conclude from the equality Eq. (4.10) and from the fact that there is a one-to-one map from directing random measures to exchangeable distributions that $H^A \stackrel{\mathbb{P}_\alpha}{\cong} H$. \square

The following is a technical lemma that will be used in Theorem 4.3.22.

Lemma 4.3.21. *Suppose a sequential input-output model (\mathbb{P}_C, D, Y) is given with D countable and, letting $E \subset D^\mathbb{N}$ be the set of all sequences for which each $j \in D$ occurs infinitely often, $\mathbb{P}_\alpha^D(E) = 1$ for all α , for some W , $\mathbb{P}_\alpha^{Y|WD}$ is IO contractible, $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e W|(H, D, C)$ and for all α*

$$\mathbb{P}_\alpha^{Y|WD} = \begin{array}{c} W \text{---} \boxed{\mathbb{P}_\alpha^{Y^D|W}} \text{---} \boxed{\mathbb{F}_{lu}} \text{---} Y \\ D \text{---} \boxed{\mathbb{F}_{lu}} \end{array}$$

then

$$\mathbb{P}_\alpha^{Y|HD} = \begin{array}{c} H \text{---} \boxed{\mathbb{P}_\alpha^{Y^D|H}} \text{---} \boxed{\mathbb{F}_{lu}} \text{---} Y \\ D \text{---} \boxed{\mathbb{F}_{lu}} \end{array}$$

Proof. Y^D is a function of Y and D (see Definition 4.3.12) and H is a function of Y^D . Say $f : Y \times D \rightarrow H$ is such that $H = f(Y, D)$ (see Definition 4.3.14). Then (using $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e W|(H, D, C)$)

$$\mathbb{P}_\alpha^{YH|WD} = \begin{array}{c} W \text{---} \boxed{\mathbb{P}_\alpha^{Y^D|W}} \text{---} \boxed{\mathbb{F}_{lu}} \text{---} Y \\ D \text{---} \boxed{\mathbb{F}_{lu}} \text{---} \boxed{\mathbb{F}_f} \text{---} H \end{array} \quad (4.11)$$

For $d \in D^\mathbb{N}$, take $[d = j]_i$ to be the i th coordinate of d equal to $j \in D$ and $\#_{[d=j]_i}$ to be the position in d of $[d = j]_i$. Concretely, f is given by

$$f(y, d) = \bigtimes_{j \in D} A_j \mapsto \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \prod_{j \in D} \mathbb{1}_{A_j}(y_{\#_{[d=j]_i}})$$

where the limit exists. Note that for $y^D \in Y^{D \times \mathbb{N}}$ we have

$$f \circ \text{lu}(y^D, d) = \bigtimes_{j \in D} A_j \mapsto \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \prod_{j \in D} \mathbb{1}_{A_j}(y_{\#_{[d=j]_i}}^D)$$

Let $G := f \circ \text{lu}(Y^D, d)$ for some $d \in D^\mathbb{N}$.

Define $g : Y^{D \times \mathbb{N}} \rightarrow H$ by

$$g(y^D) := \bigtimes_{j \in D} A_j \mapsto \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \prod_{j \in D} \mathbb{1}_{A_j}(y_{ij}^D)$$

and note that $g(Y^D) = H$. For some $A \in \mathcal{Y}^D$ let $G_A := G(A)$ and $H_A := H(A)$. Consider

$$\begin{aligned} \mathbb{P}_\alpha(G_A \bowtie H_A) &= \int_H \mathbb{P}_\alpha^{G_A|H}(\nu(A)|\nu) \mathbb{P}_\alpha^H(d\nu) \\ &= \int_H \mathbb{P}_\alpha^{Y^D|H} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \prod_{j \in D} \mathbb{1}_{A_j}(y_{\#_{[d=j]_i}}) \bowtie \nu(A) | \nu \right) \mathbb{P}_\alpha^H(d\nu) \end{aligned}$$

but by Lemma 4.3.19, the sequence $(Y_{iD}^D)_{i \in \mathbb{N}}$ are mutually independent conditional on H and for all α , $\mathbb{P}_\alpha^{Y_{iD}^D|H}(A|\nu) \stackrel{\mathbb{P}_C}{\cong} \nu(A)$. Thus, by the law of large numbers

$$\mathbb{P}_\alpha^{Y^D|H} \left(\left[\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \prod_{j \in D} \mathbb{1}_{A_j}(y_{\#_{[d=j]_i}}) \right] \bowtie \nu(A) | \nu \right) = 1$$

In particular, by Eq. (4.11)

$$\begin{aligned} \mathbb{P}_\alpha^{YH|WD} &= \begin{array}{c} W \text{---} \boxed{\mathbb{P}_\alpha^{Y^D|W}} \text{---} \boxed{\mathbb{F}_{lu}} \text{---} Y \\ D \text{---} \boxed{\mathbb{F}_{lu}} \text{---} \boxed{\mathbb{F}_f} \text{---} H \end{array} \\ &= \begin{array}{c} W \text{---} \boxed{\mathbb{P}_\alpha^{Y^D|W}} \text{---} \boxed{\mathbb{F}_{lu}} \text{---} Y \\ D \text{---} \boxed{\mathbb{F}_{lu}} \text{---} \boxed{\mathbb{F}_g} \text{---} H \end{array} \end{aligned}$$

noting that $\mathbb{F}_g \otimes \text{del}_W = \mathbb{P}_\alpha^{H|Y^D W}$ (as H is by definition a function of Y^D)

$$\mathbb{P}_\alpha^{YH|WD} = \begin{array}{c} W \text{---} \boxed{\mathbb{P}_\alpha^{H|W}} \text{---} \boxed{\mathbb{P}_\alpha^{Y^D|WH}} \text{---} \boxed{\mathbb{F}_{lu}} \text{---} Y \\ D \text{---} \boxed{\mathbb{F}_{lu}} \text{---} H \end{array}$$

and by the fact that $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e W|(H, D, C)$ and Y^D is a function of Y and D , we have $(Y^D, Y) \perp\!\!\!\perp_{\mathbb{P}_C}^e W|(H, C)$ by contraction. Then by decomposition, $Y^D \perp\!\!\!\perp_{\mathbb{P}_C}^e W|(H, C)$ also, and so

$$\mathbb{P}_\alpha^{YH|WD} = \begin{array}{c} W \text{---} \boxed{\mathbb{P}_\alpha^{H|W}} \text{---} \boxed{\mathbb{P}_\alpha^{Y^D|H}} \text{---} \boxed{\mathbb{F}_{lu}} \text{---} Y \\ D \text{---} \boxed{\mathbb{F}_{lu}} \text{---} H \end{array}$$

and so by higher order conditionals,

$$\mathbb{P}_\alpha^{Y|HD} = \begin{array}{c} \text{H} \text{---} \boxed{\mathbb{P}_\alpha^{Y^D|H}} \text{---} \boxed{\mathbb{F}_{\text{lu}}} \text{---} Y \\ \text{D} \text{---} \boxed{\mathbb{F}_{\text{lu}}} \end{array}$$

□

Theorem 4.3.22 is the main result of this section. It shows that sequential input-output model (\mathbb{P}_C, D, Y) is IO contractible over some W if and only if there is some hypothesis H such that the Y_i s are related to the D_i s by conditionally independent and identical response functions (subject to a support assumption).

Note that property (2) is equivalent to the conjunction of conditionally independent and identical response functions (Def 4.2.1) and weak data-independence (Def 4.3.16).

Theorem 4.3.22 (Representation of IO contractible models). *Suppose a sequential input-output model (\mathbb{P}_C, D, Y) with sample space (Ω, \mathcal{F}) is given with D countable and, letting $E \subset D^\mathbb{N}$ be the set of all sequences for which each $j \in D$ occurs infinitely often, $\mathbb{P}_\alpha^D(E) = 1$ for all α . Then the following are equivalent:*

1. *There is some W such that $\mathbb{P}^{Y|WD}$ is IO contractible and $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e W | (H, D, C)$*
2. *For all i , $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{\neq i}, D_{\neq i}, C) | (H, D_i)$ and for all i, j*

$$\mathbb{P}_C^{Y_i|HD_i} = \mathbb{P}_C^{Y_j|HD_j}$$

3. *There is some $\mathbb{L} : H \times X \rightarrow Y$ such that for all α ,*

$$\mathbb{P}_\alpha^{Y|HD} = \begin{array}{c} \text{H} \text{---} \bullet \text{---} \boxed{\mathbb{L}} \text{---} Y_i \\ \text{D}_i \text{---} \boxed{\mathbb{L}} \\ i \in \mathbb{N} \end{array}$$

Proof. As a preliminary, we will show

$$\mathbb{F}_{\text{lu}} = \begin{array}{c} \boxed{\begin{array}{c} Y^D \text{---} \boxed{\mathbb{F}_{\text{lus}}} \text{---} Y \\ D \text{---} \boxed{\mathbb{F}_{\text{lus}}} \\ i \in \mathbb{N} \end{array}} \end{array} \quad (4.12)$$

where $\text{lus} : D \times Y^D \rightarrow Y$ is the single-shot lookup function

$$((y_i)_{i \in D}, d) \mapsto y_d$$

Recall that lu is the function

$$((d_i)_{i \in \mathbb{N}}, (y_{ij})_{i,j \in \mathbb{N} \times D}) \mapsto (y_{id_i})_{i \in \mathbb{N}}$$

By definition, for any $\{A_i \in \mathcal{Y} | i \in \mathbb{N}\}$

$$\begin{aligned}
 \mathbb{F}_{\text{lu}}\left(\bigotimes_{i \in \mathbb{N}} A_i | (d_i)_{i \in \mathbb{N}}, (y_{ij})_{i \in \mathbb{N} \times D}\right) &= \delta_{(y_{id_i})_{i \in \mathbb{N}}} \left(\bigotimes_{i \in \mathbb{N}} A_i\right) \\
 &= \prod_{i \in \mathbb{N}} \delta_{y_{id_i}}(A_i) \\
 &= \prod_{i \in \mathbb{N}} \mathbb{F}_{\text{evs}}(A_i | d_i, (y_{ij})_{j \in D}) \\
 &= \left(\bigotimes_{i \in \mathbb{N}} \mathbb{F}_{\text{evs}}\right) \left(\bigotimes_{i \in \mathbb{N}} A_i | (d_i)_{i \in \mathbb{N}}, (y_{ij})_{i, j \in \mathbb{N} \times D}\right)
 \end{aligned}$$

which is what we wanted to show.

(1) \implies (3): Let $H := \mathcal{M}_1(Y^D)$ and define $\mathbb{M} : H \rightarrow Y^X$ by $\mathbb{M}(A|h) = h(A)$ for all $A \in \mathcal{Y}^X$, $h \in H$. Define $\mathbb{Y}^D : \Omega \rightarrow Y^{\mathbb{N} \times D}$ as in Theorem 4.3.18. Fix $w \in W$ and let $\mathbb{P}_{\alpha, w}^{Y^D} := \mathbb{P}_{\alpha}^{Y^D|W}(\cdot|w)$. By the column exchangeability of $\mathbb{P}_{\alpha, w}^{Y^D}$, from Kallenberg (2005a, Prop. 1.4) there is a directing random measure $\mathbb{H} : Y^{X \times \mathbb{N}} \rightarrow H$ such that

$$\mathbb{P}_{\alpha, w}^{Y^D|H} = \boxed{\begin{array}{c} \text{H} \text{---} \bullet \text{---} \boxed{\mathbb{M}} \text{---} \mathbb{Y}_i^D \\ i \in \mathbb{N} \end{array}}$$

as the right hand side does not depend on w

$$\begin{array}{c} W \longrightarrow * \\ \\ \mathbb{P}_{\alpha}^{Y^D|WH} = \boxed{\begin{array}{c} \text{H} \text{---} \bullet \text{---} \boxed{\mathbb{M}} \text{---} \mathbb{Y}_i^D \\ i \in \mathbb{N} \end{array}} \end{array} \quad (4.13)$$

By Theorem 4.3.18, for each $w \in W$

$$\mathbb{P}_{\alpha}^{Y|WD} = \begin{array}{c} W \text{---} \boxed{\mathbb{P}_{\alpha}^{Y^D|W}} \\ D \text{---} \boxed{\mathbb{F}_{\text{lu}}} \end{array} \longrightarrow Y$$

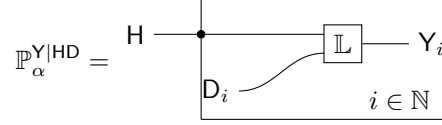
and so by Lemma 4.3.21

$$\mathbb{P}_{\alpha}^{Y|HD} = \begin{array}{c} \text{H} \text{---} \boxed{\mathbb{P}_{\alpha}^{Y^D|H}} \\ D \text{---} \boxed{\mathbb{F}_{\text{lu}}} \end{array} \longrightarrow Y \quad (4.14)$$

By the assumption $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e W | (H, D, C)$, we can substitute Equations (4.13) and (4.12) into (4.14) for

$$\mathbb{P}_{\alpha}^{Y|HD} = \boxed{\begin{array}{c} \text{H} \text{---} \bullet \text{---} \boxed{\mathbb{L}} \text{---} Y_i \\ D_i \text{---} \curvearrowright \\ i \in \mathbb{N} \end{array}}$$

(3) \implies (2): If



then by the definition of higher order conditionals, for any $i \in \mathbb{N}$ and any $\alpha \in C$

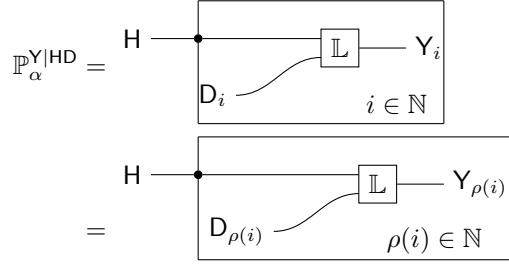
$$\mathbb{P}_\alpha^{Y_i|HD_i Y_{\neq i} D_{\neq i}} \stackrel{\mathbb{P}_C}{\cong} \mathbb{L} \otimes \text{del}_{Y^{\mathbb{N}} \times X^{\mathbb{N}}}$$

hence $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{\neq i}, D_{\neq i}, C) | (H, D_i)$

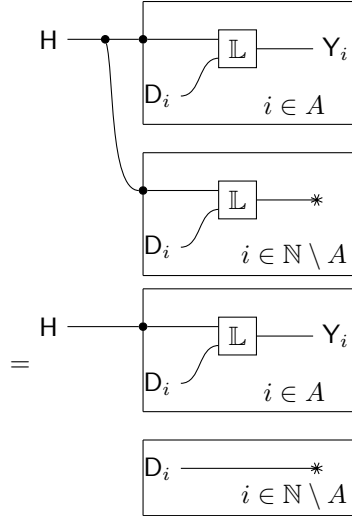
(2) \implies (1): Take $W := H$, which implies $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e W | (H, D, C)$ immediately. Take $\mathbb{L} := \mathbb{P}_C^{Y_i|H X_i}$ for arbitrary i (by assumption, they are all the same). Then, by assumption $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{[1,i]}, D_{[1,i]}, C) | (H, D_i)$, for all α

$$\mathbb{P}_\alpha^{Y_i|HD_i Y_{[1,i]} D_{[1,i]}} \stackrel{\mathbb{P}_C}{\cong} \mathbb{L} \otimes \text{del}_{Y^{i-1} \times X^{i-1}}$$

and so by higher order conditionals



hence (\mathbb{P}_C, D, Y) is exchange commutative over H . Furthermore, take $A \subset \mathbb{N}$. Then



so (\mathbb{P}_C, D, Y) is also local over H . □

4.3.2 Conditionally independent and identical responses with data-independent inputs

Theorem 4.3.22 says that a sequential input-output model (\mathbb{P}_C, D, Y) features conditionally independent and identical response functions $\mathbb{P}_C^{Y_i|HD_i}$ and is weakly data independent if and only if there is some W such that $\mathbb{P}^{Y|WD}$ is IO contractible over W , and $Y \perp\!\!\!\perp_{\mathbb{P}_C}^E W|(H, C)$ (see Definition 4.3.14 for the definition of H).

A simple special case to consider is when W is single valued – that is, when $\mathbb{P}_C^{Y|D}$ is IO contractible. As Theorem 4.3.23 shows, this corresponds to the models where the inputs D are strongly data-independent and everywhere independent of the hypothesis H . We can also consider the case where (\mathbb{P}_C, D, Y) is only exchange commutative over $*$. This corresponds to models where the inputs D are data-independent and the hypothesis H depends on a symmetric function of the inputs D (under some side conditions).

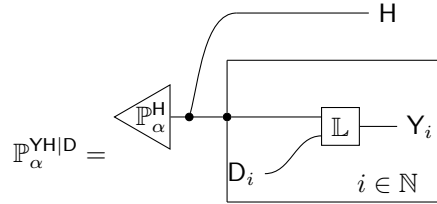
Theorem 4.3.23 (Data-independent IO contractibility). *Suppose a sequential input-output model (\mathbb{P}_C, D, Y) with sample space (Ω, \mathcal{F}) is given with D countable and, letting $E \subset D^{\mathbb{N}}$ be the set of all sequences for which each $j \in D$ occurs infinitely often, $\mathbb{P}_C^D(E) = 1$ for all α . Then the following are equivalent:*

1. $\mathbb{P}^{Y|D}$ is IO contractible
2. For all i , $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^E (Y_{\neq i}, D_{\neq i}, C)|(H, D_i)$, for all i, j

$$\mathbb{P}_C^{Y_i|HD_i} = \mathbb{P}_C^{Y_j|HD_j}$$

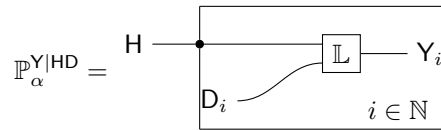
$$, H \perp\!\!\!\perp_{\mathbb{P}_C}^E D|C \text{ and for all } i \ D_i \perp\!\!\!\perp_{\mathbb{P}_C}^E D_{(i, \infty]}|(D_{[1, i]}, C)$$

3. There is some $\mathbb{L} : H \times X \rightarrow Y$ such that for all α ,

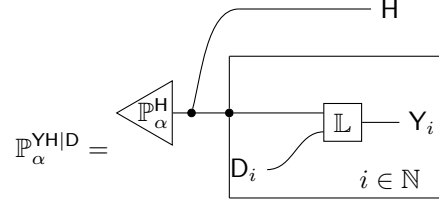


Proof. (1) \implies (3) From Theorem 4.3.18, (\mathbb{P}_C, D, Y) IO contractible over W implies $Y^D \perp\!\!\!\perp_{\mathbb{P}_C}^E D|(W, C)$ which in turn implies $H \perp\!\!\!\perp_{\mathbb{P}_C}^E D|(W, C)$. If $\mathbb{P}^{Y|D}$ is IO contractible, then $H \perp\!\!\!\perp_{\mathbb{P}_C}^E D$.

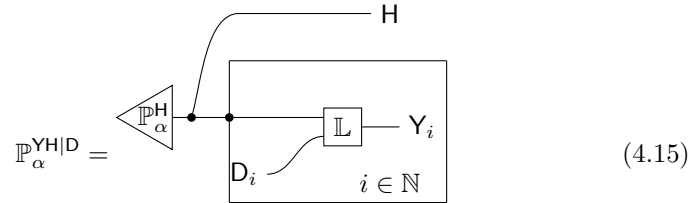
From Theorem 4.3.22 we have $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^E (Y_{[1, i]}, D_{[1, i]}, C)|(H, D_i)$ and $\mathbb{P}_C^{Y_i|HD_i} = \mathbb{P}_C^{Y_j|HD_j}$ and



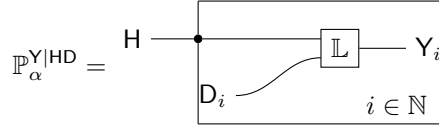
Noting that $H \perp\!\!\!\perp_{\mathbb{P}_C}^e D$, we can write



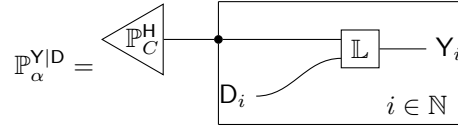
(3) \implies (2) From



we have



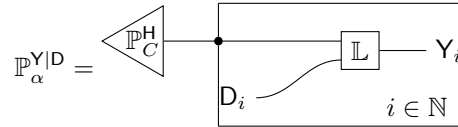
and $H \perp\!\!\!\perp_{\mathbb{P}_C}^e D|C$, so we get all elements of (2) immediately except $D_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{[1,i]}|(D_{[1,i]}, C))$. But marginalising Equation (4.15) over H yields



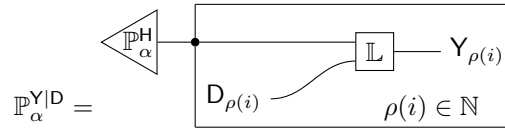
and, in particular, for any $A \subset \mathbb{N}$

$$\mathbb{P}_\alpha^{Y_A|D_{(A, \mathbb{N} \setminus A)}} = \mathbb{P}_\alpha^{Y_A|D_A} \otimes \text{del}_{D_{|\mathbb{N} \setminus A|}}$$

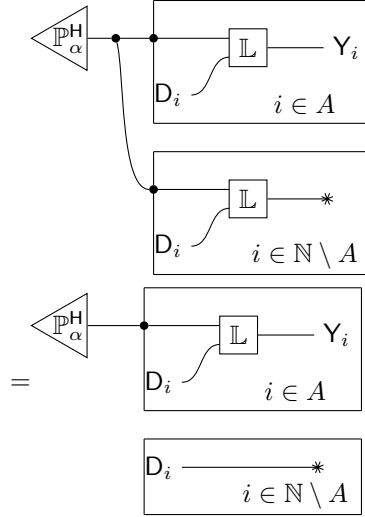
hence, taking $A = \{i\}$, $D_{\mathbb{N} \setminus \{i\}} \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_i)|(D_i, C)$ which implies $D_{(i, \infty]} \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_i)|(D_{[1,i]}, C)$. (2) \implies (1) From Theorem 4.3.22 and $H \perp\!\!\!\perp_{\mathbb{P}_C}^e D|C$,



and the argument from here is analogous to the section “(2) \implies (1)” from Theorem 4.3.22. In particular



hence $\mathbb{P}_C^{Y|D}$ is exchange commutative. Furthermore, take $A \subset \mathbb{N}$. Then



□

Lemma 4.3.24 (Exchangeable to dominated).

Given $\mathbb{P}_C^{Y|D}$ exchange commutative, we can show that $\mathbb{P}_C^{Y|WD}$ is IO contractible over where W is some symmetric function of D . This W therefore satisfies $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e W|(H, D, C)$, and so exchange commutativity of $\mathbb{P}_C^{Y|D}$ is enough to establish (\mathbb{P}_C, D, Y) has conditionally independent and identical response conditionals.

Theorem 4.3.25. *If $\mathbb{P}_C^{Y|D}$ is exchange commutative, and for each α \mathbb{P}_α^D is absolutely continuous with respect to some exchangeable distribution in $\Delta(D^{\mathbb{N}})$ then there is some W symmetric over D such that (\mathbb{P}_C, D, Y) is IO contractible over W and $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e W|(D_i, H, C)$.*

Proof. Take all sets in $\mathcal{D}^{\mathbb{N}}$ invariant under any finite permutation, and call this the *exchangeable σ -algebra* \mathcal{E} (Kallenberg, 2005a, pg. 29). Let $\mathcal{E}_D := D^{-1}(\mathcal{E})$.

Consider $\mathbb{P}_\alpha^{Y_{[n]}|D}$ for some $n \in \mathbb{N}$. By assumption of exchange commutativity, for any permutation $\rho_{>n}$ that only affects indices after the n th, $\mathbb{P}_\alpha^{Y_{[n]}|D} = \mathbb{P}_\alpha^{Y_{[n]}|D_{\rho_{>n}}}$. That is, $\mathbb{P}_\alpha^{Y_{[n]}|D}$ is $\mathcal{E}_{D_{(n,\infty)}} \vee \sigma(D_{[n]})$ -measurable.

Let \mathcal{T}_D be the tail σ -algebra generated by D defined as the intersection $\mathcal{T}_D := \bigcap_{i=1}^{\infty} \sigma(D_{[i,\infty)})$. Note that $\mathcal{T}_{D_{(n,\infty)}} = \mathcal{T}_D$. By Kallenberg (2005a, Corollary 1.6) and the assumption that \mathbb{P}_α^D is dominated by an exchangeable distribution, $\mathcal{E}_{D_{(n,\infty)}} = \mathcal{T}_{D_{(n,\infty)}}$ almost surely for any n . But $\mathcal{T}_{D_{(n,\infty)}} = \mathcal{T}_D$, so $\mathcal{E}_{D_{(n,\infty)}} = \mathcal{T}_D$ almost surely for any n , and thus $\mathcal{E}_{D_{(n,\infty)}} = \mathcal{E}_D$ almost surely.

Thus $\mathbb{P}_\alpha^{Y_{[n]}|D}$ is $\mathcal{E}_D \vee \sigma(D_{[n]})$ -measurable. By Kallenberg (2005a, Corollary 1.6) again, there is a random J taking values in the set of distributions on D

$$\mathbb{P}_\alpha^{\mathbf{Y}_{[n]}|\mathbf{D}^J} = \mathbb{P}_\alpha^{\mathbf{Y}_{[n]}|\mathbf{D}_{[n]}^J} \otimes \text{erase}_{DX^N}$$

It remains to be shown that $\mathbb{Y}_i \perp\!\!\!\perp_{\mathbb{P}_C} \mathbb{J} \mid (\mathbf{D}_i, \mathbf{H}, \mathbf{C})$. Recall $\mathbb{P}_\alpha^{\mathbf{D}}$ is dominated by an exchangeable distribution, call it ν . Because $\mathbb{P}_C^{\mathbf{Y} \mid \mathbf{D}}$ is exchange commutative, $\nu \mathbb{P}_C^{\mathbf{Y} \mid \mathbf{D}}$ is also exchangeable, and thus $\mathbb{P}_\alpha^{\mathbf{D}} \mathbb{P}_C^{\mathbf{Y} \mid \mathbf{D}}$ is dominated by an exchangeable distribution.

$\mathbb{P}_\alpha^{Y|HD} =$

Diagram illustrating the definition of $\mathbb{P}_\alpha^{Y|HD}$. A horizontal line labeled H enters a box. Inside the box, the line has a control dot connected by a curved line to a box labeled L . The output of the box is labeled Y_i . Below the box, the label D_i is connected to the control line. To the right of the box, the label $i \in \mathbb{N}$ is shown.

9

$$\mathbb{P}_{\alpha}^{Y_i|D_i Y_A, D_A} = \mathbb{P}_{\alpha}^{Y_j|D_j R V Y_B D_B}$$

Proof. If: By Theorem 4.3.22, $\mathbb{P}_\alpha^{\mathbf{Y}^{\text{HD}}}$ is IO contractible. By Theorem 4.3.20, \mathbf{H} is almost surely a function of both (D_A, \mathbf{Y}_A) and (D_B, \mathbf{Y}_B) and, furthermore, $\mathbf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (D_A, \mathbf{Y}_A) | (D_i, \mathbf{H}, C)$, $\mathbf{Y}_j \perp\!\!\!\perp_{\mathbb{P}_C}^e (D_B, \mathbf{Y}_B) | (D_j, \mathbf{H}, C)$. Hence there is some $f : D^{\mathbb{N}} \times Y^{\mathbb{N}} \rightarrow H$ such that for all $E \in \mathcal{Y}$, $d_i \in D$, $d \in D^{\mathbb{N}}$, $y \in Y^{\mathbb{N}}$

Only if: By construction

4.4 Discussion

4.4.1 Simple symmetries vs strategic behaviour

The previous section established a number of symmetries of input-output models that either imply, or are equivalent to (under some side conditions) conditionally independent and identical responses. Theorem 4.3.22 shows that for weakly data-independent models, conditionally independent and identical responses is equivalent to IO contractibility over the directing random conditional \mathbf{H} . Where $\mathbb{P}_\alpha^{\mathbf{YD}}$ is dominated by an exchangeable measure for every α , Theorem 4.3.27 establishes the alternative condition that, loosely speaking, the conditional distribution of any output given the corresponding input and an infinite sequence of additional input-output pairs is identical.

These general results both establish a symmetry of the conditional distribution of outputs after conditioning on some “long run limit” (either \mathbf{H} or an infinite sequence of input-output pairs). This makes the results less tidy than the classic result for conditionally independent and identically distributed sequences, which required only that the distribution of the sequence be symmetric to permutation, and no conditioning on long-run limits.

We can consider simpler versions of exchange commutativity or IO contractibility that omit the conditioning on the long-run limit. This is where we choose $W = *$ in Definitions 4.3.3 and 4.3.4 respectively. Theorem 4.3.22 and Corollary 4.3.26 respectively establish that these simpler symmetries are sufficient for conditionally independent and identical responses. We present a few examples to show that these simpler symmetries are not necessary for this property, however. The basic idea in these examples is that, even with conditionally independent and identical responses, inputs could be chosen strategically and different inputs could be chosen according to different strategies.

Example 1: purely passive observation Purely passive observations can be modeled with a single-element probability set \mathbb{P}_C where $|\mathbb{P}_C| = 1$. In this case, a model that is exchangeable over the sequence of pairs $\mathbf{YD} := (D_i, Y_i)_{i \in \mathbb{N}}$ has (\mathbb{P}_C, D, Y) exchange commutative over $*$. This follows from the fact that

$$\begin{aligned} \mathbb{P}_C^{\mathbf{YD}} &= \mathbb{P}_C^{(\mathbf{YD})_\rho} \\ \implies \mathbb{P}_C^{\mathbf{Y|D}} &= \mathbb{P}_C^{Y_\rho | D_\rho} \end{aligned}$$

thus by Corollary 4.3.26, (\mathbb{P}_C, D, Y) features conditionally independent and identical response functions. Note that $\mathbb{P}_C^{\mathbf{Y|D}}$ is not necessarily IO contractible. Suppose there is a machine with two arms $D = \{0, 1\}$, one of which pays out \$100 and the other that pays out nothing. A decision maker (DM) doesn’t know which is which, but the DM watches a sequence of people operate the machine who almost all do know which one is good. The DM is sure that they all want the money, and that they will pull the good arm $1 - \epsilon$ of the time independent of every other trial. Set the hypotheses \mathbf{H} to “0 is good” and “1 is good” (which

we'll just refer to as $\{0, 1\}$), with 50% probability on each initially. Then

$$\begin{aligned}\mathbb{P}_C^{Y_2|D_2}(100|1) &= \sum_{0,1} \mathbb{P}_C^{Y_2|D_2H}(100|1, 0) \mathbb{P}_C^{H|D_2}(0|1) + \mathbb{P}_C^{Y_2|D_2H}(100|1, 1) \mathbb{P}_C^{H|D_2}(1|1) \\ &= 1 - \epsilon\end{aligned}$$

but

$$\begin{aligned}\mathbb{P}_C^{Y_2|D_1D_2}(100|0, 1) &= \sum_{0,1} \mathbb{P}_C^{Y_2|D_1D_2H}(100|0, 1, 0) \mathbb{P}_C^{H|D_1D_2}(0|0, 1) + \mathbb{P}_C^{Y_2|D_1D_2H}(100|0, 1, 1) \mathbb{P}_C^{H|D_1D_2}(1|0, 1) \\ &= 0.5\end{aligned}$$

Example 2: all inputs chosen by the decision maker Consider the previous example, except instead of watching knowledgeable operators, the DM will pull each lever themselves, and they will decide in advance on the sequence of pulls. We suppose that the DM's model reflects precisely their knowledge of H when they choose the sequence D , and so H has no dependence on D .

$$\begin{aligned}\mathbb{P}_C^{Y_2|D_2}(100|1) &= \sum_{0,1} \mathbb{P}_C^{Y_2|D_2H}(100|1, 0) \mathbb{P}_C^H(0) + \mathbb{P}_C^{Y_2|D_2H}(100|1, 1) \mathbb{P}_C^H(1) \\ &= 0.5 \\ \mathbb{P}_C^{Y_2|D_1D_2}(100|0, 1) &= \sum_{0,1} \mathbb{P}_C^{Y_2|D_1D_2H}(100|0, 1, 0) \mathbb{P}_C^H(0) + \mathbb{P}_C^{Y_2|D_1D_2H}(100|0, 1, 1) \mathbb{P}_C^H(1) \\ &= 0.5\end{aligned}$$

so here the decision maker has adopted a model where $\mathbb{P}_C^{Y|D}$ is IO contractible.

Example 3: mixing strategies A decision maker might be in the position of having both observational and experimental data. Modify the machine from the previous example so that the good lever pays out \$100 $0.5 + \epsilon$ of the time, and the bad lever pays out $0.5 - \epsilon$ of the time and (as before) the DM's prior probability that each lever is the good one is 0.5. Suppose the DM from the previous examples observes a sequence of strangers operating the machine, the results associated with the sequence of pairs $(D_i, Y_i)_{i \in \mathbb{N}}$, and also operates the machine themselves according to a plan fixed in advance, the results associated with the sequence of pairs $(E_i, Z_i)_{i \in \mathbb{N}}$.

If, in this situation, the DM were to adopt a model $(\mathbb{P}_C, (D, E), (Y, Z))$ such that $\mathbb{P}_C^{YZ|DE}$ is IO contractible over $*$, understanding (D, E, Y, Z) to be a single sequence of pairs, then by Theorem 4.3.9 implies, for some $n \in \mathbb{N}$ and any choice of actions by the DM α ,

$$\mathbb{P}_\alpha^{Z_i|E_iD_{[n]}Y_{[n]}} = \mathbb{P}_\alpha^{Y_i|D_iE_{[n]}Z_{[n]}}$$

That is, there is a symmetry between predicting the consequences of one of the DM's inputs from the DM's passive observations and predicting the outputs of one of the passive observations from the DM's input-output pairs. However, this might not be appropriate - while the DM is ignorant about which lever is better, the others who operate the machine might not be. If the DM supposes that the strangers are knowledgeable regarding the better lever, then he will take the stranger's having chosen a certain lever as evidence that that lever is the better one, while he will not treat his own choice of lever in the same way. Thus, for example,

$$\mathbb{P}_\alpha^{Z_i|E_i D_{[2]} Y_{[2]}}(100|1, 1, 1, 0, 100) > \mathbb{P}_\alpha^{Y_i|E_i E_{[2]} Z_{[2]}}(100|1, 1, 1, 0, 100)$$

In this case, the DMs model is not even exchange commutative over $*$.

4.4.2 Implications of IO contractibility

Theorem 4.3.27 establishes a necessary condition for conditionally independent and identical response functions: the conditional distributions of every output given the corresponding input and a suitable infinite sequence of other input-output pairs are identical. The following two examples substantiate the claims made at the beginning of this chapter: that conditionally independent and identical response functions imply, under appropriate conditions, that experimental and observational data is interchangeable and that experimental data predicts the outcomes of a decision maker's choices just as well as it predicts held out experimental outputs.

It is common to find statements to the effect that it is *hard* to assess whether conditionally independent and identical responses are reasonable to assume, and may require expert knowledge. However, we propose that, in fact, it's often easy to reject this assumptions. One might respond that we might still accept that the condition is close to holding, and in this case it's may often be possible to make good decisions by reasoning as if it holds precisely. However, this begs the question: in what sense is it "close" to holding? In other words, if we want to relax this assumption, what do we relax it to?

some references

A key question is thus: how do we formulate weaker assumptions that are more widely acceptable than the assumption of conditionally independent and identical response functions? This is explored in Chapter 5.

Example 4: experimental and observational data Suppose we have two sequences of binary pairs $((D, X), Y) := ((D_i, X_i), Y_i)_{i \in \mathbb{N}}$ the D_i s represent whether patient i was given a particular medicine. The D_i s were assigned uniformly according to some source of randomness for even $i \geq 2$, while what exactly determined the D_j for odd j is not known and is likely to have involved patient or doctor discretion. The X_i s are covariates, and the Y_i s record binarized outcomes of the treatment. D_0 is up to the decision maker, set deterministically according to $\alpha \in 0, 1$. Within both the even and the odd indices of D both options are taken infinitely often with probability 1.

According to Theorem 4.3.27, the assumption of conditionally independent and identical responses applied to $((D, X), Y)$ implies

$$\begin{aligned} \mathbb{P}_\alpha^{Y_0|D_0X_0D_{\text{odds}}X_{\text{odds}}Y_{\text{odds}}} &= \mathbb{P}_\alpha^{Y_0|D_0D_{\text{evens}\setminus\{0\}}X_{\text{evens}\setminus\{0\}}Y_{\text{evens}\setminus\{0\}}} \\ &= \mathbb{P}_\alpha^{Y_2|D_2X_2X_{\text{evens}\setminus\{0,2\}}Y_{\text{evens}\setminus\{0,2\}}} \\ &= \mathbb{P}_\alpha^{Y_2|D_2X_2X_{\text{odds}}Y_{\text{odds}}} \end{aligned}$$

That is, under this assumption, four problems are deemed identical:

- Predicting a held-out experimental outcome from the experimental data
- Predicting a held-out experimental outcome from the observational data
- Predicting the outcome of the decision maker’s input from the experimental data
- Predicting the outcome of the decision maker’s input from the observational data

But the proposition that these problems are *identical* is hard to swallow: despite the obvious differences in the procedures used to obtain the various sequences of pairs, such an assumption nevertheless holds that these differences cannot possibly lead to any differences between the problems discussed.

In practice, when both experimental and observational data are available, they are *not* assumed to be interchangeable in this sense – in fact, the question of how well the observational data predicts experimental outputs is one of substantial interest Eckles and Bakshy (2021); Gordon et al. (2018, 2022).

4.5 Conditionally independent and identical response functions with data-dependent inputs

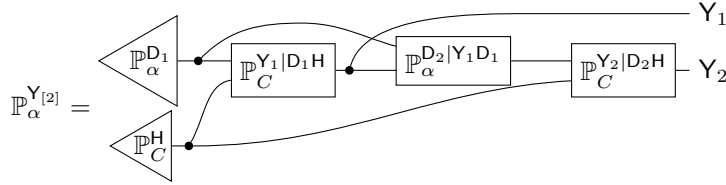
The results of the previous section concern “just-do” models where actions have not dependence on previous data. Decision problems of interest actually have actions that depend on data – what’s really wanted are “see-do” models of some variety (see Definition 3.2.14). Here, Theorem 4.3.23 is generalised to sequential see-do models with the use of *probability combs*. This work is preliminary, and in particular this generalisation doesn’t lend itself to an easy interpretation that we are aware of.

To begin with an example, consider a probability set (\mathbb{P}_C, D, Y) with $D := (D_i)_{i \in \mathbb{N}}$ and $Y := (Y_i)_{i \in \mathbb{N}}$ as usual, and take a subsequence $(D_i, Y_i)_{i \in [2]}$ of length 2. Suppose \mathbb{P}_C features conditionally independent and identical response functions – that is, the following holds:

$$\begin{aligned} Y_i &\perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{<i}, D_{<i}, C) | HD_i & \forall i \in \mathbb{N} \\ \wedge \mathbb{P}_C^{Y_i | HD_i} &= \mathbb{P}_C^{Y_0 | HD_0} & \forall i \in \mathbb{N} \end{aligned}$$

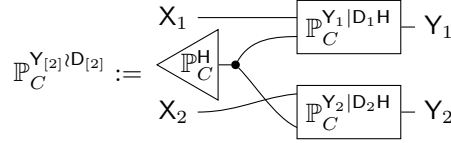
and, for simplicity, assume $H \perp\!\!\!\perp_{\mathbb{P}_C}^c (D, C)$ also.

Then, for arbitrary $\alpha \in C$



note that D_2 depends on Y_1 and D_1 . $\mathbb{P}_\alpha^{D_2|Y_1 D_1}$ has been “inserted” between the response conditionals $\mathbb{P}_C^{Y_1|D_1 H}$ and $\mathbb{P}_C^{Y_2|D_2 H}$.

Given $\mathbb{P}_C^{Y_1|D_1 H}$ and $\mathbb{P}_C^{Y_2|D_2 H}$, define the comb



then $\mathbb{P}_C^{Y_{[2]} \wr D_{[2]}}$ is IO contractible. $\mathbb{P}_C^{Y_{[2]} \wr D_{[2]}}$ is *not* a uniform conditional probability; in general

$$\mathbb{P}_\alpha^{D_1 D_2} \mathbb{P}_C^{Y_{[2]} \wr D_{[2]}} \neq \mathbb{P}_\alpha^{Y_1 Y_2}$$

4.5.1 Combs

Combs generalise conditional probabilities in this sense: given a conditional distribution and a marginal distribution of the right type, joining them together (with the semidirect product 2.2.23) I get a marginal distribution of a different type. Define “1-combs” as conditional probabilities and “0-combs” as conditional distributions. Then the previous observation can be restated as: given a 1-comb and a 0-comb of the right type, joining them together yields a 0-comb of a different type. Higher order combs generalise this: given an n -comb and an $n - 1$ -comb of the right type, joining them yields an $n - 1$ comb.

Joining combs uses an “insert” operation (Definition 4.5.4). A graphical

depiction of this operation gives some intuition for why it is called “insert”:

$$\begin{aligned}
 \mathbb{P}_\alpha^{Y_1 D_2 Y_2 | D_1} &= \text{insert}(\mathbb{P}_\alpha^{D_2 | D_1 Y_1}, \mathbb{P}_C^{Y_{[2]} | D_{[2]}}) \\
 &= \text{Diagram (4.16)} \\
 &= \text{Diagram (4.17)}
 \end{aligned}
 \tag{4.17}$$

While Equation (4.16) is a well-formed string diagram in the category of Markov kernels, Equation (4.17) is not. In the case that all the underlying sets are discrete, Equation (4.17) can be defined using an extended string diagram notation appropriate for the category of real-valued matrices (Jacobs et al., 2019), though we do not introduce this extension here.

Formal definitions of combs and both notations follow. As with conditional probabilities, a *uniform* n -comb $\mathbb{P}_C^{Y_{[n]} | X_{[n]}}$ is a Markov kernel that satisfies the definition of an n -comb for each $\alpha \in C$.

Definition 4.5.1 (n -Comb). Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ with variables $Y_i : \Omega \rightarrow Y$, $D_i : \Omega \rightarrow D$ for $i \in [n]$ and $W : \Omega \rightarrow W$, the uniform n -comb $\mathbb{P}^{Y_{[n]} | D_{[n]} | W} : W \times D^n \rightarrow Y^n$ is the Markov kernel given by the recursive definition

$$\begin{aligned}
 \mathbb{P}^{Y_1 | D_1 | W} &= \mathbb{P}^{Y_1 | D_1 W} \\
 &= \text{Diagram} \\
 \mathbb{P}^{Y_{[m]} | D_{[m]} | W} &=
 \end{aligned}$$

Definition 4.5.2 (\mathbb{N} -comb). Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ with variables $Y_i : \Omega \rightarrow Y$ and $D_i : \Omega \rightarrow D$, for $i \in \mathbb{N}$ and $W : \Omega \rightarrow W$, the \mathbb{N} -comb $\mathbb{P}^{Y_{\mathbb{N}} | D_{\mathbb{N}} | W} : W \times D^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ is the Markov kernel such that for all $n \in \mathbb{N}$

$$\mathbb{P}^{Y_{\mathbb{N}} | D_{\mathbb{N}} | W}[\text{id}_{Y^n} \otimes \text{del}_{Y^{\mathbb{N}}}] = \mathbb{P}^{Y_{[n]} | D_{[n]} | W} \otimes \text{del}_{Y^{\mathbb{N}}}$$

image alignment

Theorem 4.5.3 (Existence of \mathbb{N} -combs). *Given a probability set \mathbb{P} with variables $Y_i : \Omega \rightarrow Y$ and $D_i : \Omega \rightarrow D$ for $i \in \mathbb{N}$ and $W : \Omega \rightarrow W$, D, Y, W standard measurable, a uniform \mathbb{N} -comb $\mathbb{P}^{Y_{\mathbb{N}} \wr D_{\mathbb{N}} | W} : W \times D^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ exists.*

Proof. For each $n \in \mathbb{N}$ $m < n$, we have

$$\mathbb{P}^{Y_{[n]} \wr D_{[n]} | W} [\text{id}_{Y_{n-m}} \otimes \text{del}_{Y_m}] = \mathbb{P}^{Y_{[n-m]} \wr D_{[n-m]}} \otimes \text{del}_{Y_m}$$

and each m and n comb exists because the requisite conditional probabilities exist. Therefore the existence of $\mathbb{P}^{Y_{\mathbb{N}} \wr D_{\mathbb{N}}}$ is a consequence of Lemma 4.3.6. \square

For discrete sets, the insert operation has a compact definition:

Definition 4.5.4 (Comb insert - discrete). Given an n -comb $\mathbb{P}_{\alpha}^{Y_{[n]} \wr D_{[n]}}$ and an $n-1$ comb $\mathbb{P}_{\alpha}^{D_{[2,n]} \wr Y_{[n-1]} | D_1}$ with (D, \mathcal{D}) and (Y, \mathcal{Y}) discrete, for all $y_i \in Y$ and $d_i \in D$

$$\begin{aligned} & \text{insert}(\mathbb{P}_{\alpha}^{D_{[2,n]} \wr Y_{[n-1]} | D_1}, \mathbb{P}_{\alpha}^{Y_{[n]} \wr D_{[n]}})(y_{[n]}, d_{[2,n]} | d_1) \\ &= \mathbb{P}_{\alpha}^{Y_{[n]} \wr D_{[n]}}(y_{[n]} | d_{[n]}) \mathbb{P}_{\alpha}^{D_{[2,n]} \wr Y_{[n-1]} | D_1}(d_{[n]} | d_1, y_{[n-1]}) \end{aligned}$$

Inserting a comb into a comb (of appropriate dimensions) yields a conditional probability.

Theorem 4.5.5. *Given an n -comb $\mathbb{P}_{\alpha}^{Y_{[n]} \wr D_{[n]}}$ and an $n-1$ comb $\mathbb{P}_{\alpha}^{D_{[2,n]} \wr Y_{[n-1]} | D_1}$, (D, \mathcal{D}) and (Y, \mathcal{Y}) discrete,*

$$\begin{aligned} & \text{insert}(\mathbb{P}_{\alpha}^{D_{[2,n]} \wr Y_{[n-1]} | D_1}, \mathbb{P}_{\alpha}^{Y_{[n]} \wr D_{[n]}}) \\ &= \mathbb{P}_{\alpha}^{Y_{[n]} \wr D_{[2,n]} | D_1} \end{aligned}$$

Proof. Take $Y_{[0]} = D_{n+1} = *$, and

$$\begin{aligned} & \text{insert}(\mathbb{P}_{\alpha}^{D_{[2,n]} \wr Y_{[n-1]} | D_1}, \mathbb{P}_{\alpha}^{Y_{[n]} \wr D_{[n]}})(y_{[n]}, d_{[2,n]} | d_1) \\ &= \mathbb{P}_{\alpha}^{Y_{[n]} \wr D_{[n]}}(y_{[n]} | d_{[n]}) \mathbb{P}_{\alpha}^{D_{[2,n]} \wr Y_{[n-1]} | D_1}(d_{[2,n]} | d_1, y_{[n-1]}) \\ &= \prod_{i=1}^n \mathbb{P}_{\alpha}^{Y_{[i]} | D_{[i]} Y_{[i-1]}}(y_i | d_{[i]}, y_{[i-1]}) \mathbb{P}_{\alpha}^{D_{i+1} | D_{[i]} Y_{[i-1]}}(d_i | d_{[i-1]}, y_{[i-1]}) \\ &= \mathbb{P}_{\alpha}^{Y_{[n]} \wr D_{[2,n]}}(y_{[n]}, d_{[n]} | d_1) \end{aligned}$$

\square

Aside: combs are the output of the “fix” operation

There is a relationship between combs and the “fix” operation defined in Richardson et al. (2017). In particular, suppose we have a probability \mathbb{P}_{α} and a comb

$\mathbb{P}_\alpha^{Y_{[2]}|D_{[2]}}$. Then (assuming discrete sets)

$$\begin{aligned} \mathbb{P}_\alpha^{Y_{[2]}|D_{[2]}}(y_1, y_2 | d_1, d_2) &= \mathbb{P}_\alpha^{Y_1|D_1}(y_1 | d_1) \mathbb{P}_\alpha^{Y_2|D_2}(y_2 | d_2) \\ &= \frac{\mathbb{P}_\alpha^{Y_1|D_1}(y_1 | d_1) \mathbb{P}_\alpha^{D_2|Y_1 D_1}(d_2 | y_1, d_1) \mathbb{P}_\alpha^{Y_2|D_2}(y_2 | d_2)}{\mathbb{P}_\alpha^{D_2|Y_1 D_1}(d_2 | y_1, d_1)} \\ &= \frac{\mathbb{P}_\alpha^{Y_{[2]}|D_2|D_1}(y_1, y_2, d_2 | d_1)}{\mathbb{P}_\alpha^{D_2|Y_1 D_1}(d_2 | y_1, d_1)} \end{aligned}$$

That is (at least in this case), the result of “division by a conditional probability” used in the fix operation is a comb. We speculate that the output of the fix operation is, in general, an n -comb, but we have not proven this.

4.5.2 Representation of models with data dependent inputs

If we want to specify a “see-do” model where the input D_i might depend on inputs and outputs with indices lower than i , it might be substantially easier to talk about the comb $\mathbb{P}_\alpha^{Y_i|D}$ than about the conditional probability $\mathbb{P}_\alpha^{Y_i|D}$. The latter will have to account for possible dependence between outputs Y_i and *future* inputs D_j , which may not be straightforward, while by construction specification of the comb only requires the dependence of Y_i on past inputs and outputs.

The definitions of IO contractibility (Section 4.3) don’t apply directly to the case of combs, because (for example)

$$\text{swap}_\rho \mathbb{P}_C^{Y_i|D} \text{swap}_{\rho-1} \neq \mathbb{P}_C^{Y_\rho|D_\rho}$$

We can generalise IO contractibility to a notion that applies to generic Markov kernels, and do so in Section 4.6. The downside of this is that it’s no longer easy to talk about what the transformations mean in terms of equalities of conditional distributions of variables – nor indeed, in terms of probability comb equalities, because unlike a conditional distribution, the product of a probability comb and a swap map is not necessarily a probability comb itself. In any case, Theorem 4.6.8 is an analogue of Theorem 4.3.22 for the case of a data-dependent model. There are two crucial differences between these theorems. First, while Theorem 4.3.22 constructs the hypothesis H as a function of the given variables, Theorem 4.6.8 extends the sample space to construct the corresponding hypothesis G . If the “given variables” are observable, this means that G is not necessarily able to be constructed from observables. This leads to an open question:

- Under what conditions is the hypothesis G (as defined in 4.6.8) equal to a function of the given variables?

Secondly, Theorem 4.6.8 is proved without the “auxiliary” variable W , and as a result it includes the additional assumption $H \perp\!\!\!\perp_{\mathbb{P}_C} (X, C)$.

4.6 IO contractible Markov kernels - definitions and explanation

The following definitions mirror the definitions Section 4.3, except they are stated in terms of kernel products instead of variables. This is so that they can be applied to combs, instead of limited to conditional probabilities.

Definition 4.6.1 (kernel locality). A Markov kernel $\mathbb{K} : X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ is *local* if for all $n \in \mathbb{N}$, $A_i \in \mathcal{Y}$, $(x_{[n]}, x_{[n]^c}) \in \mathbb{N}$ there exists $\mathbb{L} : X^n \rightarrow Y^n$ such that

$$\begin{array}{c}
 \begin{array}{ccc}
 & W & \\
 & \swarrow & \searrow \\
 D_{[n]} & \text{---} \boxed{\mathbb{P}_\alpha^{Y|WD}} & \text{---} Y_{[n]} \\
 \uparrow & & \uparrow \\
 D_{(n,\infty)} & & *
 \end{array}
 \end{array}
 \quad
 \begin{array}{ccc}
 & W & \\
 & \swarrow & \searrow \\
 D_{[n]} & \text{---} \boxed{\mathbb{P}_\alpha^{Y|WD_{[n]}}} & \text{---} Y_{[n]} \\
 \uparrow & & \uparrow \\
 D_{(n,\infty)} & & *
 \end{array}
 \\
 = \\
 \Longleftrightarrow \\
 \mathbb{K}(\bigtimes_{i \in [n]} A_i \times Y^{\mathbb{N}} | x_{[n]}, x_{[n]^c}) = \mathbb{L}(\bigtimes_{i \in [n]} A_i | x_{[n]})
 \end{array}$$

Definition 4.6.2 (kernel exchange commutativity). A Markov kernel $\mathbb{K} : X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ *commutes with exchange* if for all finite permutations $\rho : \mathbb{N} \rightarrow \mathbb{N}$, $A_i \in \mathcal{Y}$, $(x_{[n]}, x_{[n]^c}) \in \mathbb{N}$

$$\begin{array}{c}
 \mathbb{K} \text{swap}_{\rho, Y} = \text{swap}_{\rho, X} \mathbb{K} \\
 \Longleftrightarrow \\
 \mathbb{K}(\bigtimes_{i \in \mathbb{N}} A_{\rho(i)} | (x_i)_{i \in \mathbb{N}}) = \mathbb{K}(\bigtimes_{i \in \mathbb{N}} A_i | (x_{\rho(i)})_{i \in \mathbb{N}})
 \end{array}$$

IO contractibility is the conjunction of both assumptions.

Definition 4.6.3 (kernel IO contractibility). A Markov kernel $\mathbb{K} : X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ is *IO contractible* if it is local and commutes with exchange.

4.6.1 Representation of IO contractible Markov kernels

The main theorem is proved in this section. Much of the work parallels work already done in Section 4.3.

Theorem 4.6.4 is similar to Theorem 4.3.9, except it is stated in terms of transformations of a Markov kernel instead of in terms of conditional probabilities of variables.

Theorem 4.6.4 (Equality of equally sized contractions). *A Markov kernel $\mathbb{K} : X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ is IO contractible if and only if for every $n \in \mathbb{N}$ and every $A \subset \mathbb{N}$ there exists some $\mathbb{L} : X^n \rightarrow Y^n$ such that*

$$\mathbb{K} \text{marg}_A = \text{swap}_{[n] \leftrightarrow A} \mathbb{L} \otimes \text{del}_{X^{\mathbb{N}}}$$

Proof. Only if: By exchange commutativity

$$\text{swap}_{[n] \leftrightarrow A} \mathbb{K} = \mathbb{K} \text{swap}_{[n] \leftrightarrow A}$$

multiply both sides by $\text{swap}_{[n] \leftrightarrow A}$ on the right and, because $\text{swap}_{[n] \leftrightarrow A}$ is its own inverse,

$$\text{swap}_{[n] \leftrightarrow A} \mathbb{K} \text{swap}_{[n] \leftrightarrow A} = \mathbb{K}$$

so

$$\begin{aligned} \mathbb{K} \text{marg}_A &= \text{swap}_{[n] \leftrightarrow A} \mathbb{K} \text{swap}_{[n] \leftrightarrow A} \text{marg}_A \\ &= \text{swap}_{[n] \leftrightarrow A} \mathbb{K} \text{marg}_{[n]} \end{aligned}$$

By locality, there exists some $\mathbb{L} : X^n \rightarrow Y^n$ such that

$$\begin{aligned} \mathbb{K} \text{marg}_{[n]} &= \mathbb{K}(\text{id}_{[n]} \otimes \text{del}_{X^{\mathbb{N}}}) \\ &= \mathbb{L} \otimes \text{del}_{X^{\mathbb{N}}} \end{aligned}$$

If: Taking $A = [n]$ for all n establishes locality.

For exchange commutativity, note that for all $x \in X^{\mathbb{N}}$, $n \in \mathbb{N}$, we have

$$\begin{aligned} \text{swap}_{A \leftrightarrow [n]} \mathbb{K} \text{marg}_A &= \text{swap}_{A \leftrightarrow [n]} \mathbb{K} \text{swap}_{A \leftrightarrow [n]} (\text{id}_{[n]} \otimes \text{del}_{X^{\mathbb{N}}}) \\ &= \mathbb{K} \text{marg}_{[n]} \\ &= \mathbb{K}(\text{id}_{[n]} \otimes \text{del}_{X^{\mathbb{N}}}) \end{aligned}$$

Then by Lemma 4.3.6

$$\text{swap}_{A \leftrightarrow [n]} \mathbb{K} \text{swap}_{A \leftrightarrow [n]} = \mathbb{K}$$

Consider an arbitrary finite permutation $\rho : \mathbb{N} \rightarrow \mathbb{N}$. ρ can be decomposed into a finite set of cyclic permutations on disjoint orbits. Each cyclic permutation is simply the composition of some set of transpositions, and so ρ itself can be written as a composition of a sequence of transpositions. Thus for any finite $\rho : \mathbb{N} \rightarrow \mathbb{N}$

$$\text{swap}_{\rho} \mathbb{K} \text{swap}_{\rho} = \mathbb{K}$$

□

Theorem 4.6.5 is similar to Theorem 4.3.18, except the latter uses a variable Y^D explicitly defined on the sample space, while Theorem 4.6.5 simply says an appropriate probability distribution exists.

Theorem 4.6.5. *A Markov kernel $\mathbb{K} : X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ is IO contractible if and only if there exists a column exchangeable probability distribution $\mu \Delta(Y^{|\mathbb{X}| \times \mathbb{N}})$ such*

that

$$\begin{aligned} \mathbb{K} &= \begin{array}{c} \text{W} \\ \text{D} \end{array} \begin{array}{c} \boxed{\mathbb{P}_\alpha^{Y^D|W}} \\ \text{---} \end{array} \begin{array}{c} \boxed{\mathbb{F}_{lu}} \\ \text{---} \end{array} Y \\ &\iff \\ \mathbb{K}(A|(x_i)_{i \in \mathbb{N}}) &= \mu \Pi_{(x_i)_{i \in \mathbb{N}}}(A) \forall A \in \mathcal{Y}^{\mathbb{N}} \end{aligned}$$

Where $\Pi_{(d_i i)_{i \in \mathbb{N}}} : Y^{|X| \times \mathbb{N}} \rightarrow Y^{\mathbb{N}}$ is the function

$$(y_{ji})_{j,i \in X \times \mathbb{N}} \mapsto (y_{d_i i})_{i \in \mathbb{N}}$$

that projects the (x_i, i) indices of y for all $i \in \mathbb{N}$, and \mathbb{F}_{ev} is the Markov kernel associated with the evaluation map

$$\begin{aligned} ev : X^{\mathbb{N}} \times Y^{X \times \mathbb{N}} &\rightarrow Y \\ ((x_i)_{i \in \mathbb{N}}, (y_{ji})_{j,i \in X \times \mathbb{N}}) &\mapsto (y_{x_i i})_{i \in \mathbb{N}} \end{aligned}$$

Proof. Only if: Choose $e := (e_i)_{i \in \mathbb{N}}$ such that $e_{i+|X|j}$ is the i th element of X for all $i, j \in \mathbb{N}$.

Define

$$\mu\left(\bigtimes_{(i,j) \in X \times \mathbb{N}} A_{ij}\right) := \mathbb{K}\left(\bigtimes_{(i,j) \in X \times \mathbb{N}} A_{ij} | e\right) \forall A_{ij} \in \mathcal{Y}$$

Now consider any $x := (x_i)_{i \in \mathbb{N}} \in X^{\mathbb{N}}$. By definition of e , $e_{x_i i} = x_i$ for any $i, j \in \mathbb{N}$.

Define

$$\begin{aligned} \mathbb{Q} : X^{\mathbb{N}} &\rightarrow Y^{\mathbb{N}} \\ \mathbb{Q} &:= \begin{array}{c} \text{W} \\ \text{D} \end{array} \begin{array}{c} \boxed{\mathbb{P}_\alpha^{Y^D|W}} \\ \text{---} \end{array} \begin{array}{c} \boxed{\mathbb{F}_{lu}} \\ \text{---} \end{array} Y \end{aligned}$$

and consider some $A \subset \mathbb{N}$, $|A| = n$ and $B := (x_i, i)_{i \in A}$. Note that the subsequence of e indexed by B , $e_B := (e_{x_i i})_{i \in A} = x_A$. Thus given the swap map $\text{swap}_{A \leftrightarrow B} : \mathbb{N} \rightarrow \mathbb{N}$ that sends the first element of A to the first element of B and so forth, $\text{swap}_{A \leftrightarrow B}(e_B) = x_A$. For arbitrary $\{C_i \in \mathcal{Y} | i \in A\}$, define $C_A := \text{swap}_{[n] \leftrightarrow A}(\times_{i \in [n]} C_i \times Y^{\mathbb{N}})$. Then, for arbitrary $x \in X^{\mathbb{N}}$

$$\mathbb{Q}(C_A | x) = \mu(\text{ev}_x^{-1}(C_A)) \quad (4.18)$$

The argument of μ is

$$\begin{aligned} \text{ev}_x^{-1}(C_A) &= \{(y_{ji})_{j,i \in X \times \mathbb{N}} | (y_{x_i i})_{i \in \mathbb{N}} \in C_A\} \\ &= \bigtimes_{i \in \mathbb{N}} \bigtimes_{j \in X} D_{ji} \end{aligned}$$

where

$$D_{ji} = \begin{cases} C_i & (j, i) \in B \\ Y & \text{otherwise} \end{cases}$$

and so

$$\text{swap}_{A \leftrightarrow B}(\text{ev}_x^{-1}(C_A)) = C_A \quad (4.19)$$

Substituting Equation (4.19) into (4.18)

$$\begin{aligned} \mathbb{Q}(C_A|x) &= \mu \text{swap}_{A \leftrightarrow B}(C_A) \\ &= \mathbb{K} \text{swap}_{A \leftrightarrow B}(C_A|e) \\ &= \mathbb{K} \text{swap}_{A \leftrightarrow B}(C_A|e_B, \text{swap}_{B \leftrightarrow A}(x)_B^C) && \text{by locality} \\ &= \mathbb{K} \text{swap}_{A \leftrightarrow B}(C_A|\text{swap}_{B \leftrightarrow A}(x)) \\ &= \text{swap}_{B \leftrightarrow A} \mathbb{K} \text{swap}_{A \leftrightarrow B}(C_A|x) \\ &= \mathbb{K}(C_A|x) && \text{by commutativity of exchange} \end{aligned}$$

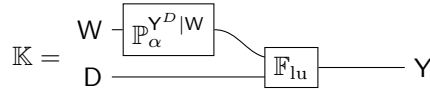
Because this holds for all x , $A \subset \mathbb{N}$, by Lemma 4.3.6

$$\mathbb{Q} = \mathbb{K}$$

Next we will show μ is column exchangeable. Consider any column swap $\text{swap}_c : X \times \mathbb{N} \rightarrow X \times \mathbb{N}$ that acts as the identity on the X component and a finite permutation on the \mathbb{N} component. From the definition of e , $\text{swap}_c(e) = e$. Thus by commutativity of exchange, for any $A \in \mathcal{Y}^{\mathbb{N}}$

$$\begin{aligned} \mathbb{K}(A|e) &= \text{swap}_c \mathbb{K} \text{swap}_c(A|e) \\ &= \mathbb{K} \text{swap}_c(A|\text{swap}_c(e)) \\ &= \mathbb{K} \text{swap}_c(A|e) \end{aligned}$$

If: Suppose



where μ is column exchangeable, and consider any two $x, x' \in X^{\mathbb{N}}$ such that some subsequences are equal $x_S = x'_T$ with $S, T \subset \mathbb{N}$ and $|S| = |T| = [n]$.

For any $\{A_i \in \mathcal{Y} | i \in S\}$, let $A_S = \text{swap}_{[n] \leftrightarrow S} \times_{i \in [n]} A_i \times Y^{\mathbb{N}}$, $A_T = \text{swap}_{S \leftrightarrow T}(A_S)$, $B = (x_i)_{i \in S}$ and $C = (x_i)_{i \in T} = (x_{\text{swap}_{S \leftrightarrow T}(i)})_{i \in S}$. By Equa-

tions (4.18) and (4.19)

$$\begin{aligned}
\mathbb{K}(A_S|x) &= \mu \text{swap}_{S \leftrightarrow B}(A_S) \\
&= \mu \text{swap}_{T \leftrightarrow C}(A_T) && \text{by column exchangeability of } \mu \\
&= \mathbb{K}(A_T | \text{swap}_{S \leftrightarrow T}(x)) \\
&= \text{swap}_{S \leftrightarrow T} \mathbb{K}(A_T | x) \\
&= \text{swap}_{S \leftrightarrow T} \mathbb{K} \text{swap}_{S \leftrightarrow T}(A_S | x)
\end{aligned}$$

so \mathbb{K} is IO contractible by Theorem 4.6.4. \square

Lemma 4.6.6 (Exchangeable table to response functions). *Given $\mathbb{K} : X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$, X and Y standard measurable, if*

$$\mathbb{K} = \begin{array}{c} \triangleleft \mu \\ \text{D} \longrightarrow \boxed{\mathbb{F}_{\text{lu}}} \longrightarrow Y \end{array}$$

for $\mu \in \Delta(Y^{X \times \mathbb{N}})$ column exchangeable, then defining $(H, \mathcal{H}) := \mathcal{M}_1(Y^{X \times \mathbb{N}})$ there is some $\mathbf{H} : Y^{X \times \mathbb{N}} \rightarrow H$ and $\mathbb{L} : H \times X \rightarrow Y$ such that

$$\mathbb{K} = \begin{array}{c} \triangleleft \mu \longrightarrow \boxed{\mathbb{F}_{\mathbf{H}}} \longrightarrow \bullet \longrightarrow \boxed{\begin{array}{c} \text{--- } H \\ \text{--- } X \text{ --- } \boxed{\mathbb{L}} \text{ --- } Y \\ i \in \mathbb{N} \end{array}} \end{array}$$

Proof. As a preliminary, we will show

$$\mathbb{F}_{\text{ev}} = \boxed{\begin{array}{c} Y^D \text{ --- } \\ D \text{ --- } \end{array} \longrightarrow \boxed{\mathbb{F}_{\text{lus}}} \longrightarrow Y} \quad (4.20)$$

$i \in \mathbb{N}$

where $\text{ev}_{Y^X \times X} : Y^X \times X \rightarrow Y$ is the single-shot evaluation function

$$(x, (y_i)_{i \in X}) \mapsto y_x$$

Recall that ev is the function

$$((x_i)_{i \in \mathbb{N}}, (y_{ji})_{j, i \in X \times \mathbb{N}}) \mapsto (y_{xi})_{i \in \mathbb{N}}$$

By definition, for any $\{A_i \in \mathcal{Y} | i \in \mathbb{N}\}$

$$\begin{aligned}
 \mathbb{F}_{\text{ev}}\left(\bigtimes_{i \in \mathbb{N}} A_i | (x_i)_{i \in \mathbb{N}}, (y_{ji})_{i \in X \times \mathbb{N}}\right) &= \delta_{(y_{x_i i})_{i \in \mathbb{N}}} \left(\bigtimes_{i \in \mathbb{N}} A_i\right) \\
 &= \prod_{i \in \mathbb{N}} \delta_{y_{x_i i}}(A_i) \\
 &= \prod_{i \in \mathbb{N}} \mathbb{F}_{\text{evs}}(A_i | x_i, (y_{ji})_{j \in X}) \\
 &= \left(\bigotimes_{i \in \mathbb{N}} \mathbb{F}_{\text{evs}}\right) \left(\bigtimes_{i \in \mathbb{N}} A_i | (x_i)_{i \in \mathbb{N}}, (y_{ji})_{j \in X \times \mathbb{N}}\right)
 \end{aligned}$$

which is what we wanted to show.

Define $\mathbb{M} : H \rightarrow Y^X$ by $\mathbb{M}(A|h) = h(A)$ for all $A \in \mathcal{Y}^X$, $h \in H$. By the column exchangeability of μ , from Kallenberg (2005a, Prop. 1.4) there is a directing random measure $\mathbb{H} : Y^{X \times \mathbb{N}} \rightarrow H$ such that

$$\begin{aligned}
 \mu(\mathbb{F}_{\mathbb{H}} \otimes \text{id}_{Y^{X \times \mathbb{N}}}) &= \text{Diagram 1} \\
 &\iff \\
 \mu\left(\bigtimes_{i \in \mathbb{N}} A_i \times B\right) &= \int_B \prod_{i \in \mathbb{N}} \mathbb{M}(A_i | h) \mu \mathbb{F}_{\mathbb{H}}(dh) \quad \forall A_i \in \mathcal{Y}^X
 \end{aligned}$$

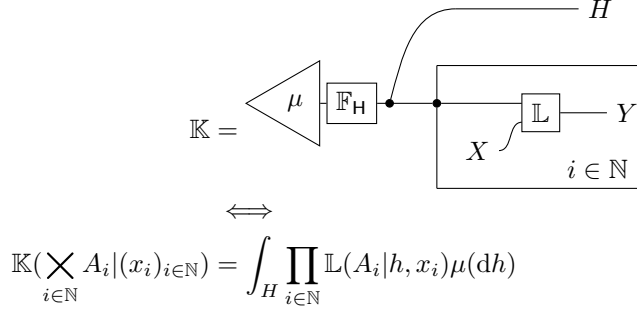
By Equations (4.21) and (4.20)

$$\begin{aligned}
 \mathbb{K} &= \text{Diagram 2} \\
 &:= \text{Diagram 3}
 \end{aligned}$$

Where we can connect the copied outputs of $\mu \mathbb{F}_{\mathbb{H}}$ to the inputs of each \mathbb{M} “inside the plate” as the plates in Equations (4.12) and (4.13) are equal in number and each connected wire represents a single copy of Y^D . \square

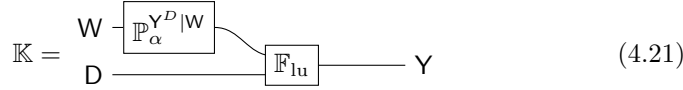
Theorem 4.6.7 is similar to Theorem 4.3.22, but it is stated without the use of variables.

Theorem 4.6.7. *Given a kernel $\mathbb{K} : X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$, let $(H, \mathcal{H}) := \mathcal{M}_1(Y^X)$ be the set of probability distributions on (Y^X, \mathcal{Y}^X) . \mathbb{K} is IO contractible if and only if there is some $\mu \in \Delta(H)$ and $\mathbb{L} : H \times X \rightarrow Y$ such that*



$$\mathbb{K}(\times_{i \in \mathbb{N}} A_i | (x_i)_{i \in \mathbb{N}}) = \int_H \prod_{i \in \mathbb{N}} \mathbb{L}(A_i | h, x_i) \mu(dh)$$

Proof. Only if: By Theorem 4.6.5, we can represent the conditional probability \mathbb{K} as



$$\mathbb{K} = \begin{array}{c} W \\ \text{---} \end{array} \begin{array}{c} \boxed{\mathbb{P}_\alpha^{Y^D | W}} \\ \text{---} \end{array} \begin{array}{c} \boxed{\mathbb{F}_{lu}} \\ \text{---} \end{array} Y \quad (4.21)$$

where $\nu \in \Delta(Y^{X \times \mathbb{N}})$ is column exchangeable.

Applying Lemma 4.6.6 yields the desired result.

If: By assumption, for any $\{A_i \in \mathcal{Y} | i \in \mathbb{N}\}$, $x := (x_i)_{i \in \mathbb{N}} \in X^{\mathbb{N}}$

$$\mathbb{K}(\times_{i \in \mathbb{N}} A_i | x) = \int_H \prod_{i \in \mathbb{N}} \mathbb{L}(A_i | h, x_i) \mu(dh)$$

Consider any $S, T \subset \mathbb{N}$ with $|S| = |T|$, and define $A_S := \times_{i \in \mathbb{N}} B_i$ where $B_i = Y$ if $i \notin S$, otherwise A_i is an arbitrary element of \mathcal{Y} . Define $A_T := \times_{i \in \mathbb{N}} B_{\text{swap}_{S \leftrightarrow T}(i)}$.

$$\begin{aligned} \mathbb{K}(A_S | x) &= \int_H \prod_{i \in S} \mathbb{L}(A_i | h, x_i) \mu(dh) \\ &= \int_H \prod_{i \in T} \mathbb{L}(A_i | h, x_{\text{swap}_{S \leftrightarrow T}(i)}) \mu(dh) \\ &= \text{swap}_{S \leftrightarrow T} \mathbb{K}(A_T | x) \\ &= \text{swap}_{S \leftrightarrow T} \mathbb{K} \text{swap}_{S \leftrightarrow T}(A_S | x) \end{aligned}$$

So by Theorem 4.3.9, \mathbb{K} is IO contractible. \square

Theorem 4.3.22 is the main result of this section. It shows that a IO contractible Markov kernel $X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ is representable as a “prior” $\mu \in \Delta(H)$ and a “parallel product” of Markov kernels $H \times X \rightarrow Y$. These will be the response conditionals when Theorem 4.3.22 is applied to probability set models. It is a data-dependent approximate analogue of Theorem 4.3.22.

Theorem 4.6.8. *Given a sequential input-output model (\mathbb{P}'_C, D', Y') on (Ω, \mathcal{F}) , then $\mathbb{P}'_C{}^{Y'D'}$ is IO contractible if and only if there is a latent extension \mathbb{P}_C of \mathbb{P}'_C to $(\Omega \times H, \mathcal{F} \otimes \mathcal{Y}^{D \times \mathbb{N}})$ with projection map $H : \Omega \times H \rightarrow H$ such that $Y_i \perp\!\!\!\perp_{\mathbb{P}'_C} (Y_{<i}, X_{<i}, C) | (X_i, H)$ and $\mathbb{P}_C^{Y_i | X_i H} = \mathbb{P}_C^{Y_j | X_j H}$ for all $i, j \in \mathbb{N}$ and $H \perp\!\!\!\perp_{\mathbb{P}_C} (X, C)$.*

Proof. If: By assumption, there is some $\mathbb{L} : H \times D \rightarrow Y$ such that

$$\mathbb{P}_C^{Y_i | HD_i} = \mathbb{L}$$

and $Y_i \perp\!\!\!\perp_{\mathbb{P}_C} (Y_{<i}, D_{<i}) | (D_i, H)$. Thus

$$\mathbb{P}_C^{Y_i | HD_i Y_{<i} D_{<i}} = \mathbb{L} \otimes \text{erase}_{Y^{i-1} \times D^{i-1}}$$

and so

$$\mathbb{P}_C^{YD} = \triangleleft \mathbb{P}_C^H \quad \begin{array}{|c|} \hline \begin{array}{c} \bullet \\ \text{---} \mathbb{L} \text{---} Y_i \\ \text{---} D_i \end{array} \\ \hline i \in \mathbb{N} \end{array} \quad (4.22)$$

and so by Theorem 4.6.7, \mathbb{P}_C^{YD} is IO contractible.

Only if: First, define the extension \mathbb{P}_C . From Theorem 4.6.7 and IO contractibility of $\mathbb{P}'_C{}^{Y'D'}$ there is some set H , $\mu \in \Delta(H)$ and $\mathbb{L} : H \times D \rightarrow Y$ such that

$$\mathbb{P}'_C{}^{Y'D'} = \triangleleft \mu \quad \begin{array}{|c|} \hline \begin{array}{c} \bullet \\ \text{---} \mathbb{L} \text{---} Y_i \\ \text{---} D_i \end{array} \\ \hline i \in \mathbb{N} \end{array}$$

thus, by the definition of the comb insert operation

$$\mathbb{P}'_C{}^{D'_{[n]} Y'_{[n]}} = \mathbb{P}'_C{}^{D_1} \odot \text{insert}(\mathbb{P}'_C{}^{D'_{[2,n]} Y'_{[n-1]}}, \mathbb{P}'_C{}^{Y'_{[n]} D'_{[n]}})$$

Let

$$\mathbb{P}_C^{Y_i | HD_i} = \mathbb{L} \quad (4.23)$$

and let $Y_i \perp\!\!\!\perp_{\mathbb{P}_C} (Y_{<i}, D_{<i}) | (D_i, H)$, and for all α set $\mathbb{P}_\alpha^{W|DY} = \mathbb{P}_\alpha^{W'|D'Y'}$ for all $W' : \Omega \rightarrow W$ and $\mathbb{P}_\alpha^{D_i | Y_{<i} D_{<i}} = \mathbb{P}_\alpha^{D'_i | Y'_{<i} D'_{<i}}$.

It remains to be shown that $\mathbb{P}_\alpha^{DY} = \mathbb{P}'_C{}^{DY}$.

By Equation (4.23) and $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{<i}, D_{<i}) | (D_i, H)$, it follows (for identical reasons as Equation (4.22)) that

$$\begin{aligned}
 \mathbb{P}_C^{Y_i D} &= \begin{array}{c} \triangleleft \mathbb{P}_C^H \\ \bullet \\ \boxed{\begin{array}{c} \text{---} \mathbb{L} \text{---} Y_i \\ D_i \text{---} \end{array}} \\ i \in \mathbb{N} \end{array} \\
 &= \begin{array}{c} \triangleleft \mu \\ \bullet \\ \boxed{\begin{array}{c} \text{---} \mathbb{L} \text{---} Y_i \\ D_i \text{---} \end{array}} \\ i \in \mathbb{N} \end{array} \\
 &= \mathbb{P}_C^{Y' D'}
 \end{aligned}$$

And so for all $n \in \mathbb{N}$

$$\begin{aligned}
 \mathbb{P}_\alpha^{D_{[n]} Y_{[n]}} &= \mathbb{P}_\alpha^{D_1} \odot \text{insert}(\mathbb{P}_\alpha^{D_{[2,n]} Y_{[n-1]}}, \mathbb{P}_C^{Y_{[n]} D_{[n]}}) \\
 &= \mathbb{P}_\alpha^{D_1} \odot \text{insert}(\mathbb{P}_\alpha^{D'_{[2,n]} Y'_{[n-1]}}, \mathbb{P}_C^{Y'_{[n]} D'_{[n]}}) \\
 &= \mathbb{P}_\alpha^{D'_{[n]} Y'_{[n]}}
 \end{aligned}$$

□

4.7 Discussion

The work in this chapter is motivated by the aim of better understanding the assumption of repeated response functions. We show that this assumption implies symmetries that are often unreasonable in typical causal inference problems. In particular, causal inference is often interested in drawing lessons from data generated in one context in order to exercise control in a context that is usually substantially different – not the least that, in the latter context, some aspects of the are under the decision maker’s control. However, the assumption of repeated response functions implies that this shift in context makes no difference at all in terms of what we can learn from the data in the long run (though, as we point out in Section 4.4 Example 3, the shift is allowed to make a difference in the short run).

For this reason, we don’t think that assuming repeatable response functions is a viable starting point for analysis of causal inference problems. This point is perhaps not news to many people who have engaged with this question in much depth. Instead, we want to consider weaker assumptions that are more broadly acceptable, and perhaps we could speculate that repeated response functions arise as a limiting case of an appropriately weaker assumption. In Chapter 5 that follows this one, we explore a few candidates for such weaker assumptions.

While this chapter also includes a discussion of data-dependent models (Theorem 4.6.8), this work has not yet offered any easy-to-interpret equivalences between repeated response functions and model symmetries. Notably, however,

combs play a key role in this analysis as well as the analysis of the seemingly unrelated question of identification of marginal graphical models. One might wonder if, given a more complete understanding of the theory of probability comes, they might come to be important tools in the causal modeller's toolbox.

Chapter 5

Other causal modelling frameworks

In this chapter, we examine the types of decision models that can be constructed from causal Bayesian networks and potential outcomes models. Neither of these popular approaches to causal inference yields a fully specified decision making model. Causal Bayesian networks are usually specified in a “rolled up” form, and certain judgements must be made about how this should be unrolled to a sequential model. Potential outcomes models, on the other hand, do not feature a native notion of “choices”, and a judgement must be made about what the relevant collection of choices in a potential outcomes model is.

5.1 What is a Causal Bayesian Network?

5.1.1 Definition of a Causal Bayesian Network

We follow the definition of a Causal Bayesian Network on Pearl (2009, page 23-24). There are a couple of technical differences: we require that interventional models are a measurable map from interventions to probability distributions, and we assume that there is a common sample space for every interventional distribution. There are also some non-technical differences: the notation is adapted for compatibility with the rest of the work in this thesis, and we separate the definition into two parts for clarity (Definitions 5.1.10 and 5.1.11).

An interventional model is a *Causal Bayesian Network* with respect to a directed acyclic graph if it satisfies a number of compatibility requirements. The following definitions are standard, and reproduced here for convenience. The definitions here are terse, readers should refer to Pearl (2009, chap. 1) for a more intuitive explanation.

Definition 5.1.1 (Directed graph). A directed graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$ is a set of nodes $\tilde{\mathcal{V}}$ and edges, which are ordered pairs of nodes $\tilde{\mathcal{E}} \subset \tilde{\mathcal{V}} \times \tilde{\mathcal{V}}$. Nodes are

written using the font $\tilde{\mathbf{V}}$.

The parents of a target node are all nodes with an edge ending at the target node.

Definition 5.1.2 (Parents). Given a directed graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$ and $\tilde{\mathbf{V}}_i \subset \tilde{\mathcal{V}}$, the parents of $\tilde{\mathbf{V}}_i$ are $\text{Pa}_{\tilde{\mathcal{G}}}(\tilde{\mathbf{V}}_i) := \{\tilde{\mathbf{V}}_j \mid (\tilde{\mathbf{V}}_j, \tilde{\mathbf{V}}_i) \in \tilde{\mathcal{E}}\}$.

A path is a sequence of edges such that the i th edge and the $i + 1$ th edge share exactly one node.

Definition 5.1.3 (Path). Given a directed graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$, a path is a sequence of edges $(E_i)_{i \in A}$ (where A is either $[n]$ or \mathbb{N}) such that for any i , E_i and E_{i+1} share exactly one node.

A directed path is a sequence of edges such that the end of the i th edge is the beginning of the $i + 1$ th edge.

Definition 5.1.4 (Directed path). Given a directed graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$, a directed path is a sequence of edges $(E_i)_{i \in A}$ (where A is either $[n]$ or \mathbb{N}) such that for any i , $E_i = (\tilde{\mathbf{V}}_k, \tilde{\mathbf{V}}_l)$ implies $E_{i+1} = (\tilde{\mathbf{V}}_l, \tilde{\mathbf{V}}_m)$ for some $\tilde{\mathbf{V}}_m \in \tilde{\mathcal{V}}$.

In an acyclic graph, directed paths never reach to the same node more than once.

Definition 5.1.5 (Directed acyclic graph). A directed graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$ is acyclic if, for every path, each node appears at most once. Directed acyclic graph is abbreviated to “DAG”.

D-separation is a key property of directed acyclic graphs for defining causal Bayesian networks. It is defined with respect to undirected paths.

Definition 5.1.6 (Blocked path). Given a DAG $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$, a path p is blocked by $\tilde{\mathbf{V}}_A \subset \tilde{\mathcal{V}}$ iff

1. $(\tilde{\mathbf{V}}_i, \tilde{\mathbf{V}}_j) \in p$ and $(\tilde{\mathbf{V}}_j, \tilde{\mathbf{V}}_k) \in p$ while $\tilde{\mathbf{V}}_j \in \tilde{\mathbf{V}}_A$
2. $(\tilde{\mathbf{V}}_j, \tilde{\mathbf{V}}_i) \in p$ and $(\tilde{\mathbf{V}}_j, \tilde{\mathbf{V}}_k) \in p$ while $\tilde{\mathbf{V}}_j \in \tilde{\mathbf{V}}_A$
3. $(\tilde{\mathbf{V}}_i, \tilde{\mathbf{V}}_j) \in p$ and $(\tilde{\mathbf{V}}_k, \tilde{\mathbf{V}}_j) \in p$ while $\tilde{\mathbf{V}}_j \cup \text{De}_{\tilde{\mathcal{G}}}(\tilde{\mathbf{V}}_j) \cap \tilde{\mathbf{V}}_A = \emptyset$

Definition 5.1.7 (d-separation). Given a DAG $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$, $\tilde{\mathbf{V}}_A$ is d -separated from $\tilde{\mathbf{V}}_B$ by $\tilde{\mathbf{V}}_C$ (all subsets of $\tilde{\mathcal{V}}$) if $\tilde{\mathbf{V}}_C$ blocks every path starting at $\tilde{\mathbf{V}}_A$ and ending at $\tilde{\mathbf{V}}_B$. This is written $\tilde{\mathbf{V}}_A \perp_{\tilde{\mathcal{G}}} \tilde{\mathbf{V}}_B \mid \tilde{\mathbf{V}}_C$.

Definition 5.1.8 (Variable-node association). Given a graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$ and a sequence of variables $(V_i)_{i \in A}$, if $|A| = |\tilde{\mathcal{V}}|$ we can associate a variable with each node of the graph with an invertible map $m : \{V_i | i \in A\} \rightarrow \tilde{\mathcal{V}}$. By convention, we give associated variables and nodes corresponding indices, and graphical operations are defined on variables through m , i.e. $\text{Pa}(V_i) := m(\text{Pa}(m^{-1}(V_i)))$.

Definition 5.1.9 (Compatibility). Given a measurable space (Ω, \mathcal{F}) , a Markov kernel $\mathbb{P} : C \rightarrow \Omega$ and a sequence of variables $(V_i)_{i \in A}$ with $V_i : \Omega \rightarrow V_i$ and a DAG \mathcal{G} with nodes $\{\tilde{V}_i\}_{i \in A}$ and the variable-node association $m : V_i \mapsto \tilde{V}_i$, \mathbb{P} is compatible with \mathcal{G} relative to m if for all $I, J, K \subset A$, $\tilde{V}_I \perp_{\mathcal{G}} \tilde{V}_J | \tilde{V}_K$ implies $V_I \perp_{\mathbb{P}}^c V_J | (V_K, C)$.

The following definition is reproduced from Pearl (2009) with the differences mentioned: notation has been matched to ours, the interventional model is assumed to be measurable and the interventional distributions defined on a common sample space.

Definition 5.1.10 (Interventional model). An interventional model is a tuple $(\mathbb{P}, C, \Omega, (V_i)_{i \in A})$ where (Ω, \mathcal{F}) is a measurable space, $V := (V_i)_{i \in A}$ a sequence of variables with $V_i : \Omega \rightarrow V_i$, V_i denumerable, where C the choice set

$$C := \{\text{do}_\emptyset\} \cup \{(\text{do}_B, v_B) | B \subset A, v_B \in \text{Range}(V_B)\}$$

That is, we take every subsequence V_B of V and add to C every element of the range of V_B , each labeled with the symbol do_B .

Definition 5.1.11 (Causal Bayesian network). Given an interventional model $(\mathbb{P}, C, \Omega, (V_i)_{i \in A})$ and a directed acyclic graph $\tilde{\mathcal{G}}$ with nodes $\tilde{\mathcal{V}}$, $(\mathbb{P}, C, \Omega, (V_i)_{i \in A}, \tilde{\mathcal{G}})$ is a *causal Bayesian network* with respect the node-variable association $m : \tilde{V}_i \mapsto V_i$ if:

1. \mathbb{P} is compatible with $\tilde{\mathcal{G}}$ with respect to m
2. $B \neq \emptyset \implies \mathbb{P}_{(\text{do}_B, v_B)}^{V_B} = \delta_{v_B}$
3. $\mathbb{P}_{(\text{do}_B, v_B)}^{V_i | \text{Pa}(V_i)} \stackrel{\mathbb{P}^{(\text{do}_B, v_B)}}{\cong} \mathbb{P}_{\text{do}_\emptyset}^{V_i | \text{Pa}(V_i)}$ for all $i \notin B$

We have a two comments to make about this definition. First, the sequence of variables $(V_i)_{i \in A}$ cannot be arbitrary – they must be “causally compatible” (see Section 1.4. For example, the sequence $V := (X, X)$ for some $X : \Omega \rightarrow X$ is a perfectly legitimate variable, but by condition (2) the intervention $\mathbb{P}_{\text{do}_{\{1,2\}}, (0,1)}$ is asked to assign probability 1 to the impossible event $(X \bowtie 0) \cap (X \bowtie 1)$. Second, condition (3) is subtly under-specified: $\mathbb{P}_{\text{do}_\emptyset}^{V_i | \text{Pa}(V_i)}$ is not necessarily $\mathbb{P}_{(\text{do}_B, v_B)}$ -almost surely unique. We could therefore either require condition (3) to hold for some version of $\mathbb{P}_{\text{do}_\emptyset}^{V_i | \text{Pa}(V_i)}$ for each intervention in C , or we could require it for a single version of $\mathbb{P}_{\text{do}_\emptyset}^{V_i | \text{Pa}(V_i)}$ uniformly over all interventions in C .

For continuously valued variables, the ability to pick a version of the conditional probability for each intervention leads to undesirable results. Suppose \tilde{V}_i is a parent of \tilde{V}_j , and the associated variable V_i is continuously valued and $\mathbb{P}_{\text{do}_\emptyset}^{V_i}(\{v_i\}) = 0$ for all singletons $v_i \in V_i$. Then for every intervention $\text{do}_{\{i\}}(v_i)$, we can choose a version of $\mathbb{P}_{\text{do}_\emptyset}^{V_j|V_i}$ that takes an arbitrary value at the point v_i (because this point has measure 0), so property (3) is satisfied trivially. Because of this, we will henceforth suppose that variables are discrete.

5.1.2 Unrolling a causal Bayesian network

Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ and an independent and identically distributed (IID) sequence $\mathbf{X} := (\mathbf{X}_i)_{i \in [n]}$, it is common to “roll up” the joint distribution $\mathbb{P}^{\mathbf{X}} \in \Delta(X^n)$ to a single representative distribution $\mathbb{P}^{\mathbf{X}_0} \in \Delta(X)$ and say something like “the \mathbf{X}_i are IID according to $\mathbb{P}^{\mathbf{X}_0}$ ”. Because of the IID assumption, the full joint distribution $\mathbb{P}^{\mathbf{X}} \in \Delta(X^n)$ can be unambiguously reconstructed from a statement like this.

A causal Bayesian network is similarly a rolled-up representation of a model of some sequence of variables. Unlike an IID sequence, it isn’t completely unambiguous how to unroll it. We propose the following method: first, posit a sequence of variables $\mathbf{V} := (\mathbf{V}_{ij})_{i \in A, j \in [n]}$, and extend the set C to be the set of sequences of interventions

$$\{(\text{do}_{B_j j}(v_B))_{j \in [n]} | \forall j : B_j \subset A, v_{B_j} \in \text{Range}(\mathbf{V}_{B_j j})\}$$

i.e. C now consists of all sequences of separate interventions to each subsequence of variables $\mathbf{V}_{Aj} := (\mathbf{V}_{ij})_{i \in A}$, understood to refer to variables arising from a particular iteration of the decision procedure.

Given a graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$, we now have a collection of variable-node association maps $m_j : \{\mathbf{V}_{ij} | i \in A\} \rightarrow \tilde{\mathcal{V}}$ such that $m_j(\mathbf{V}_{ij}) = \tilde{V}_i$.

We now need to specify how variables in an unrolled causal Bayesian network are distributed, given some sequence of interventions. By analogy with the original case of IID variables, we conclude that the $\mathbf{V}_{Aj} := (\mathbf{V}_{ij})_{i \in A}$ are mutually independent given any particular sequence of interventions. Furthermore, Definition 5.1.11 constrains the distribution of each variable given a particular sequence of interventions from C . For a sequence of interventions $\alpha \in C$, let $\pi_j(\alpha)$ be the j th intervention in the sequence. We might posit the following analogue of condition (3):

$$3' \quad \pi_j(\alpha) = (\text{do}_{B_j}, v_{B_j}) \text{ implies } \mathbb{P}_\alpha^{V_{ij} | \text{Pa}(V_{ij})} \stackrel{\mathbb{P}_\alpha}{\cong} \mathbb{P}_{\text{do}_\emptyset^n}^{V_{i1} | \text{Pa}(V_{i1})} \text{ for all } i \notin B$$

Where do_\emptyset^n is a sequence of n do_\emptyset interventions. This is a combination of an assumption that variables in the sequence are conditionally identically distributed given appropriate interventions and condition (3) from Definition 5.1.11. However, it’s not quite satisfactory. Take $B := \text{Pa}(V_{i1})$, and suppose $\mathbb{P}_{\text{do}_\emptyset^n}^{V_{B1}}(\{x\}) = 0$. Then

(3') would be satisfied by a model for which

$$\begin{aligned}\mathbb{P}_{(\text{do}_{B_1}, x, \text{do}_{B_2}, x)}^{\mathbf{V}_{i1}|\mathbf{B}}(U|x) &= \delta_0(U) \\ \mathbb{P}_{(\text{do}_{B_1}, x, \text{do}_{B_2}, x)}^{\mathbf{V}_{i2}|\mathbf{B}}(U|x) &= \delta_1(U)\end{aligned}$$

that is, if the empty intervention is unsupported over some element of the range of a variable, then (3') allows models that assign different consequences to repetitions of the same intervention on this variable, if those intervention forces the variable into the region that originally had no support.

We propose instead the restricted assumption of identical response functions: for any pair \mathbf{V}_{ij} and \mathbf{V}_{ik} , unless i is intervened on by $\pi_j(\alpha)$ and not intervened on by $\pi_k(\alpha)$, then then the conditional probability of \mathbf{V}_{ij} given its parents is equal to the conditional probability of \mathbf{V}_{ik} given its parents. This is condition [3*].

In order to be able to “roll up” a sequence of interventions, we also require that the response to the j th intervention does not depend on any of the interventions other than the j th. If this were not the case, then even if the restricted assumption of identical response functions were satisfied, different sequences of interventions would “roll up” to different interventional models. Condition 4* is the formalisation of this requirement. In Chapter 4, we showed that conditionally independent and identical response functions allow for the estimation of conditional probabilities from previous data, but noted that this did not necessarily imply that estimating conditional probabilities under a particular fixed choice was sufficient for decision making. Condition 4* is the assumption that, once we have the interventional conditional probabilities for any sequence of interventions, nothing else is needed.

Condition 5* is the requirement that observations are mutually independent.

Definition 5.1.12 (Unrolled causal Bayesian network). Given an interventional model $(\mathbb{P}, C, \Omega, (\mathbf{V}_{ij})_{i \in A, j \in [n]})$ and a directed acyclic graph $\tilde{\mathcal{G}}$ with nodes $\tilde{\mathcal{V}}$, $(\mathbb{P}, C, \Omega, (\mathbf{V}_i)_{i \in A}, \tilde{\mathcal{G}})$ is an *unrolled causal Bayesian network* with respect the node-variable association maps $m_j : \tilde{\mathbf{V}}_{ij} \mapsto \mathbf{V}_i$ if, for all $j, k \in [n]$:

- 1* $\mathbb{P}^{\mathbf{V}_{Aj}}$ is compatible with $\tilde{\mathcal{G}}$ with respect to m_j for all $j \in [n]$
- 2* $\pi_j(\alpha) = (\text{do}_{B_j}, v_{B_j})$ and $B_j \neq \emptyset$ implies $\mathbb{P}_{(\text{do}_{B_j}, v_{B_j})}^{\mathbf{V}_{Bj}} = \delta_{v_{B_j}}$
- 3* If $\pi_j(\alpha) = (\text{do}_{B_j}, v_{B_j})$, $\pi_k(\alpha) = (\text{do}_{B_k}, v_{B_k})$ and $i \notin B_j \cup B_k$ then $\mathbb{P}_{\alpha}^{\mathbf{V}_{ij}|\text{Pa}(\mathbf{V}_{ij})} \stackrel{\mathbb{P}_{\alpha}}{\cong} \mathbb{P}_{\text{do}_{\emptyset}}^{\mathbf{V}_{ik}|\text{Pa}(\mathbf{V}_{ik})}$
- 4* $\pi_j(\alpha) = \pi_j(\alpha')$ implies $\mathbb{P}_{\alpha}^{\mathbf{V}_{Aj}} = \mathbb{P}_{\alpha'}^{\mathbf{V}_{Aj}}$
- 5* $\mathbf{V}_{Aj} \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathbf{V}_{A[n] \setminus \{j\}} | \mathbf{C}$

5.1.3 Uncertainty in an unrolled causal Bayesian network

Condition 3* of Definition 5.1.12 establishes that, depending on the precise sequence of interventions chosen, certain conditionals are identical. In Chapter 4, we considered conditionals (or “response functions”) that were identical *conditional on some hypothesis* H . Problems addressed with causal Bayesian networks are also usually problems where these conditional distributions are initially unknown (and, in some cases, they may remain unknown even after examining an arbitrarily large amount of data). We propose to use the same method to represent uncertainty over conditional distributions in a causal Bayesian network. In particular, an *uncertain* unrolled causal Bayesian network is an interventional model $(\mathbb{P}, C, \Omega, (V_{ij})_{i \in A, j \in [n]})$ with some variable H such that, conditional on any $h \in H$, the result is an unrolled causal Bayesian network.

Definition 5.1.13 (Uncertain unrolled causal Bayesian network). Given an interventional model $(\mathbb{P}, C, \Omega, (V_{ij})_{i \in A, j \in [n]})$ and a directed acyclic graph $\tilde{\mathcal{G}}$, $(\mathbb{P}, C, \Omega, (V_{ij})_{i \in A, j \in [n]}, H, \tilde{\mathcal{G}})$ is an *uncertain unrolled causal Bayesian network* with respect to some variable $H : \Omega \rightarrow H$ if for each $h \in H$, defining $\mathbb{P}_{\cdot, h} := \alpha \mapsto \mathbb{P}_{\alpha}^{\text{id}_{\Omega}|H}(\cdot|h)$, $(\mathbb{P}_{\cdot, h}, C, \Omega, (V_{ij})_{i \in A, j \in [n]}, \tilde{\mathcal{G}})$ is an unrolled causal Bayesian network.

Recalling the discussion in Section 3.1.2, Definition 5.1.13 associates each intervention with a unique probability distribution. One could imagine therefore calling uncertain unrolled causal Bayesian networks “Bayesian causal Bayesian networks”, although this is obviously a bit of a confusing name.

An uncertain unrolled causal Bayesian network is *almost* a conditionally independent and identical response function model. Due to 3*, such a model features conditionally independent and identical response functions wherever α consists of a sequence of interventions none of which target i . This leads us to the key result of this section: considering a subset of the interventions in C , an uncertain unrolled causal Bayesian network is IO contractible (with respect to some parameters) by application of Theorem 4.3.22.

Theorem 5.1.14 (IO contractibility of CBNs). *Given an uncertain unrolled causal Bayesian network $(\mathbb{P}, C, \Omega, (V_{ij})_{i \in A, j \in [n]}, H, \tilde{\mathcal{G}})$, take $C' \subset C$ to be sequences of interventions that, for some $j \in [n]$, do not target a particular V_{ij} for any $i \in A$. Then $V_{i[n]} \perp\!\!\!\perp_{\mathbb{P}_{C'}}^e C | (H, \text{Pa}(V_{i[n]}))$ and $\mathbb{P}_C^{V_{i[n]} | \text{HPa}(V_{i[n]})}$ is IO contractible over H .*

Proof. First we will prove $V_{i[n]} \perp\!\!\!\perp_{\mathbb{P}_{C'}}^e C | (H, \text{Pa}(V_{i[n]}))$. This is equivalent to the claim that $\mathbb{P}_{\alpha}^{V_{i[n]} | \text{HPa}(V_{i[n]})}$ is the same as $\mathbb{P}_{\alpha'}^{V_{i[n]} | \text{HPa}(V_{i[n]})}$ for any α, α' . By assumption 5* of Definition 5.1.12, for each $h \in H$

$$V_{Aj} \perp\!\!\!\perp_{\mathbb{P}_{\alpha, h}}^e V_{A[n] \setminus \{j\}} | C$$

which implies

$$\begin{aligned} V_{Aj} &\perp\!\!\!\perp_{\mathbb{P}_{\alpha}}^e V_{A[n] \setminus \{j\}} | (C, H) \\ \implies V_{ij} &\perp\!\!\!\perp_{\mathbb{P}_{\alpha}}^e V_{A[n] \setminus \{j\}} | (H, \text{Pa}(V_{i[n]}), C) \end{aligned} \tag{5.1}$$

thus it is sufficient to show that, for any $\alpha, \alpha' \in C'$ and $j \in [n]$

$$\mathbb{P}_\alpha^{V_{ij}|\text{HPa}(V_{ij})} = \mathbb{P}_{\alpha'}^{V_{ij}|\text{HPa}(V_{ij})}$$

By assumption, if $\pi_j(\alpha) =: (\text{do}_{B_j}, v_{B_j})$ and $\pi_j(\alpha') =: (\text{do}_{B'_j}, v'_{B'_j})$, $i \notin B_j \cup B'_j$, and similarly replacing the j s with k s for any $k \in [n]$. Define α'' such that, for some k , $\pi_k(\alpha'') = \pi_k(\alpha')$ and $\pi_j(\alpha'') = \pi_j(\alpha)$. Then by 4*, for all $h \in H$

$$\begin{aligned} \mathbb{P}_\alpha^{V_{ij}|\text{HPa}(V_{ij})}(A|h, y) &= \mathbb{P}_{\alpha'}^{V_{ij}|\text{HPa}(V_{ij})}(A|h, y) \\ &= \mathbb{P}_{\alpha''}^{V_{ik}|\text{HPa}(V_{ik})}(A|h, y) && \text{by 3*} \\ &= \mathbb{P}_{\alpha'}^{V_{ik}|\text{HPa}(V_{ik})}(A|h, y) && \text{by 4*} \\ &= \mathbb{P}_{\alpha'}^{V_{ij}|\text{HPa}(V_{ij})}(A|h, y) && \text{by 3*} \\ \implies \mathbb{P}_\alpha^{V_{ij}|\text{HPa}(V_{ij})} &= \mathbb{P}_{\alpha'}^{V_{ij}|\text{HPa}(V_{ij})} \end{aligned}$$

Next, IO contractibility of $\mathbb{P}_C^{V_{[n]}|\text{HPa}(V_{[n]})}$ over H . By Eq. (5.1)

$$V_{ij} \perp\!\!\!\perp_{\mathbb{P}_\alpha}^e (V_{i[1,j]}, \text{Pa}(V_{i[1,j]})) | (H, \text{Pa}(V_{i[n]}), C)$$

furthermore, by 3* and the assumption that no intervention $\alpha \in C'$ targets V_{ij} for any j , for any $\alpha \in C'$

$$\mathbb{P}_\alpha^{V_{ij}|\text{HPa}(V_{ij})}(A|h, y) = \mathbb{P}_\alpha^{V_{ik}|\text{HPa}(V_{ik})}(A|h, y)$$

thus \mathbb{P}_C has independent and identical response functions conditional on H and by Theorem 4.3.22, $\mathbb{P}_C^{V_{[n]}|\text{HPa}(V_{[n]})}$ is IO contractible over H . \square

5.1.4 Probabilistic Graphical Models

Lattimore and Rohde (2019b,a) have previously published work in which they demonstrated how to “unroll” causal Bayesian networks into what they call “Probabilistic Graphical Models”. Their work goes into more detail than this thesis on how identifiability results transfer from causal Bayesian networks to their unrolled forms.

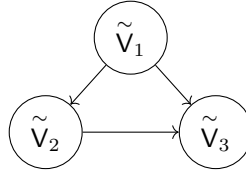
Illustrating the fact that some choices must be made in order to work out what kind of rolled-up model corresponds to a given causal Bayesian network, Rohde and Lattimore consider an unrolling where the empty intervention is always made in conjunction with a particular intervention on a particular node in the DAG. Where we explicitly write down a model of an entire sequence of observations, Probabilistic Graphical Models can be assumed to represent a sequence of an arbitrary number of empty interventions in conjunction with an arbitrary number of particular interventions on particular nodes in the DAG. Such compact representations are of course very useful when the extra details are redundant. The difference underscores the approach taken to causal modelling in this thesis – we proceed cautiously, aiming to explicitly represent all relevant

assumptions that go into building a particular type of causal model, and approach that can lead to relatively verbose model definitions and representations.

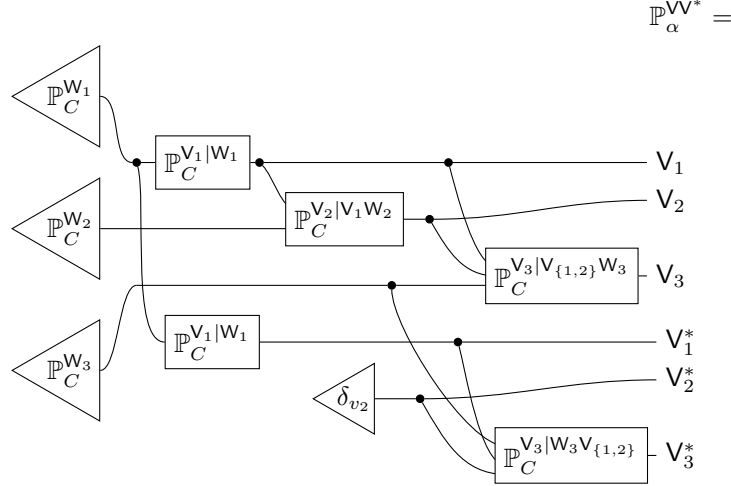
Precisely, a probabilistic graphical model is a map \mathbb{P} from the set of single-node interventions C to probability distributions \mathbb{P}_α defined on (Ω, \mathcal{F}) . A probabilistic graphical model is typically associated with a causal Bayesian network $(\mathbb{Q}, C, \Omega', (\mathbf{V}_i)_{i \in A}, \tilde{\mathcal{G}})$ where, for each $\mathbf{V}_i : \Omega \rightarrow V_i$ in the original causal Bayesian network, two variables \mathbf{V}_i and \mathbf{V}_i^* are defined on (Ω, \mathcal{F}) . The probabilistic graphical model also adds a “parameter” \mathbf{W}_i for each variable pair $(\mathbf{V}_i, \mathbf{V}_i^*)$ such that, taking C' to be interventions not targeting \mathbf{V}_i^* , for any $\alpha \in C'$, $\mathbb{P}_\alpha^{\mathbf{V}_i | \mathbf{W}_i \text{Pa}(\mathbf{V}_i)} = \mathbb{P}_\alpha^{\mathbf{V}_i^* | \mathbf{W}_i \text{Pa}(\mathbf{V}_i^*)}$ and $\mathbf{V}_i \perp\!\!\!\perp_{\mathbb{P}_{C'}}^e (\mathbf{V}_A^*, \mathbf{C}) | (\mathbf{W}_i)$ (where parents are assessed relative to the graph $\tilde{\mathcal{G}}$). This should look familiar - it is specifying, in a very similar manner to Theorem 5.1.14, that a Probabilistic Graphical Model constructed from a causal Bayesian network $(\mathbb{Q}, C, \Omega', (\mathbf{V}_i)_{i \in A}, \tilde{\mathcal{G}})$ features independent and identical response functions for each node given its parents conditional on the parameter \mathbf{W}_i .

A depiction of probabilistic graphical models and uncertain unrolled causal Bayesian networks using string diagrams gives some intuition regarding the structure of these different types of models, as well as some of the “off-page” assumptions of ordinary causal Bayesian networks.

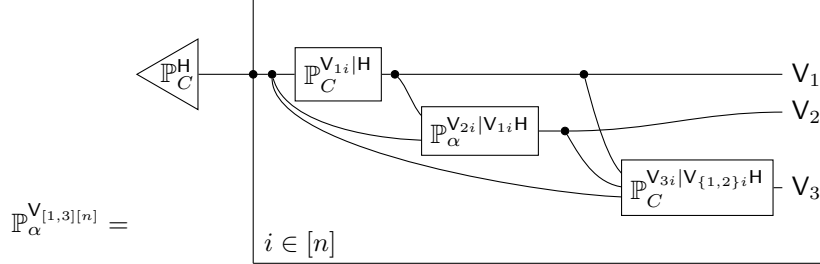
Here is the original graph $\tilde{\mathcal{G}}$ associated with $(\mathbb{Q}, C, \Omega', (\mathbf{V}_i)_{i \in A}, \tilde{\mathcal{G}})$:



Here is the probabilistic graphical model associated with the intervention (do_2, v_2)



and here is the uncertain unrolled CBN associated with the restricted set of interventions C' that consists of, for each element of the sequence, either the empty intervention or some intervention targeting V_2



where

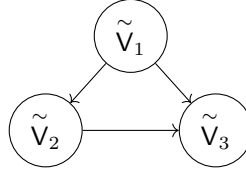
$$\mathbb{P}_\alpha^{V_{2i}|V_{1i}H} = \begin{cases} \delta_v & \pi_i(\alpha) = (\text{do}_2, v) \\ \mathbb{P}_{\text{mathrm do}_\emptyset}^{V_{2i}|V_{1i}H} & \text{otherwise} \end{cases}$$

5.1.5 Hidden confounders and precedents

One of the particularly interesting questions in causal inference is how to infer consequences of actions from observational data. A particular question of interest for problems of this type is the question of what kinds of inductive assumptions are applicable to this problem.

A common assumption in the causal Bayesian network tradition applicable to this kind of problem is the assumption of *hidden confounders*. Suppose we have

an uncertain unrolled causal Bayesian network $(\mathbb{P}, C, \Omega, (V_{ij})_{i \in [3], j \in [n]}, H, \tilde{\mathcal{G}})$ where the graph G is as follows:



and we consider the subset $C' \subset C$ of interventions that are either empty or target V_2 only. We note that Theorem 5.1.14 implies that $\mathbb{P}_C^{V_{3[n]}|HV_{1[n]}V_{2[n]}}$ is IO contractible, but not $\mathbb{P}_C^{V_{3[n]}|HV_{2[n]}}$. We can specify somewhat informally that $V_{1[n]}$ is not observed – that is, it is not associated with a measurement procedure.

The assumption of a hidden confounder often implies that, for any choice, the consequences of “interventions” have been anticipated by *some* fraction of the observations. Specifically, the IO contractibility of $\mathbb{P}_C^{V_{3[n]}|HV_{1[n]}V_{2[n]}}$ implies that $\mathbb{P}_C^{V_{3[n]}|HV_{2[n]}}$ is unchanged by swaps that leave $V_{1[n]}$ unchanged.

Theorem 5.1.15. *Given $(\mathbb{P}, C, \Omega, (V_{ij})_{i \in [3], j \in [n]}, H, \tilde{\mathcal{G}})$ with $\mathbb{P}_C^{V_{3[n]}|HV_{1[n]}V_{2[n]}}$ IO contractible over H , V_i discrete for all $i \in [3]$ and $\mathbb{P}_C^{V_{[2]i}|H}(v_1, v_2|h) > 0$ for all v_1, v_2, h , let $Q : \Omega \rightarrow [n]^n$ be a permutation of $[n]$ such that $V_{1[n]} = V_{1Q([n])}$. Then*

$$\mathbb{P}_C^{V_{3[n]}|HV_{2[n]}} = \mathbb{P}_C^{V_{3Q([n])}|HV_{2Q([n])}}$$

Proof. By IO contractibility of $\mathbb{P}_C^{V_{3[n]}|HV_{1[n]}V_{2[n]}}$ over H

$$\begin{aligned} \mathbb{P}_C^{V_{3[n]}|HV_{1[n]}V_{2[n]}} &= \mathbb{P}_C^{V_{3Q([n])}|HV_{1Q([n])}V_{2Q([n])}} \\ &= \mathbb{P}_C^{V_{3Q([n])}|HV_{1[n]}V_{2Q([n])}} \end{aligned}$$

on taking any of the decision maker's actual choices at each point in time (that is, they didn't just have the *option* to preempt the decision make, but they actually did preempt them). For convenience, we index the variables under the control of hypothetical decision maker analogue with $i = \dots, -2, -1$ and the variables under the control of the decision maker with the single index $i = 0$. This reflects the intuitive fact that the passive observations/hypothetical decision maker's choices usually come before the consequences of the real decision maker's actions, though in the present analysis the reversed indexing does not play any important role.

To formalise this, we say there is a variable D that is like an "action" at the $i = 0$ index in the sense that a choice α that leads to a distribution of actions D_0 identical to some mixture of other choices α' and α'' induces consequences equal to the same mixture of α' and α'' . We require that the model is IO contractible over the sequence of actions $(D_i)_{-i \in \{0\} \cup \mathbb{N}}$, and that each action has positive probability in the "historical" indices $i < 0$. The reason that this assumption is weaker than IO contractibility is that D_i for $i < 0$ is unobserved – that is, rules for choosing the action D_0 cannot depend on D_i for $i < 0$.

Definition 5.1.16 (Preemption). Given a probability set \mathbb{P}_C on (Ω, \mathcal{F}) and variables $Y := (Y_i)_{-i \in \{0\} \cup \mathbb{N}}$ and $D := (D_i)_{-i \in \{0\} \cup \mathbb{N}}$, D discrete, we say (\mathbb{P}_C, D, Y) is preempted if for directing random measure H , $\mathbb{P}_\alpha^{Y_i | HD_i}$ are independent and identical responses for all i ,

$$\begin{aligned} \mathbb{P}_\alpha^{D_0 | Y_{-N}} &= a\mathbb{P}_{\alpha'}^{D_0 | Y_{-N}} + b\mathbb{P}_{\alpha''}^{D_0 | Y_{-N}} \\ \implies \mathbb{P}_\alpha &= a\mathbb{P}_{\alpha'} + b\mathbb{P}_{\alpha''} \end{aligned}$$

as well as $\mathbb{P}_\alpha^{D_{-N}}$ is exchangeable and, defining G to be the directing random measure of $(P_C, *, D)$, $\mathbb{P}_\alpha^{D_i | G}(\alpha | g) > 0$ for all $\alpha \in C$, almost all $g \in G$, $i < 0$.

The assumption of preemption as given in Definition 5.1.16 can, under some conditions, yield nontrivial conclusions. Theorem 5.1.18 comes with a lot of complicated conditions, so it is worth explaining with an example first.

Suppose we have a collection of doctors Z_i who each see a series of patients and offer a treatment X_i and report their results Y_i . Each doctor may decide whether or not to prescribe based on any number of unobserved factors, and may offer additional unrecorded treatments, vary in their bedside manner and so forth, and there may be stochastic variation in any of these. The decision maker is also a doctor, and is reviewing the data contained in the sequences $(X_i, Y_i, Z_i)_{i \in [n]}$. The decision maker supposes that whatever overall treatment plan they adopt (which could also depend on features not listed in this set of variables), the same thing has probably been done at least sometimes by some of these prior doctors – that is, their treatment protocol is preempted. They also assume that the doctor's identity has no bearing on outcomes over and above the treatment protocol, they assume that doctors don't all select the same mixture of treatment protocols. It then stands to reason that the doctors who choose different treatment plans will see slightly different results *if the*

different treatment plans actually lead to different results. Conversely, if there is *no* variation in results after conditioning on whether patients were treated this suggests that whether or not treatment occurred is the *only* important feature of any treatment plan.

One way that this story could fail is if the doctors all knew exactly the long-run probabilistic outcomes of different treatment plans and coordinated with one another, they could (in principle) each pick different mixtures of treatment plans just so that the variation in outcomes is masked – that is, for example, doctor 1 picks a medium effectiveness plan 100% of the time, while doctor 2 picks a highly effective plan 50% of the time and a low effectiveness plan 50% of the time leading to the same distribution over outcomes. The conditions in Theorem 5.1.18 requiring domination by the uniform measure on $[0, 1]$ are assumptions that this kind of thing does not happen, either because the doctors don't coordinate or because, even if they did coordinate, they would not know the long-run averages of outcomes associated with each plan precisely enough to completely mask the variation.

Nxample 5.1.17 (Matrix notation). Given a sequential input-output model (\mathbb{P}_C, D, Y) with D, Y discrete, the directing random measure H takes values in the set of Markov kernels $D \rightarrow Y$, which can be identified with a subset of matrices in $\mathbb{R}^{|D| \times |Y|}$. We can therefore refer to elements of H as matrices $(h_d^y)_{d \in D, y \in Y}$ with $\sum_{y \in Y} h_d^y = 1$ for all d , and $h_d^y \stackrel{\mathbb{P}_\alpha}{\cong} \mathbb{P}_\alpha^{Y|HD}(y|h, d)$. We can also define H_d^y as the d, y -th projection of H .

Theorem 5.1.18. *Suppose we have a probability set \mathbb{P}_C on (Ω, \mathcal{F}) with variables $Y := (Y_i)_{-i \in \{0\} \cup \mathbb{N}}$, $D := (D_i)_{-i \in \{0\} \cup \mathbb{N}}$, $X := (X_i)_{-i \in \{0\} \cup \mathbb{N}}$ and $Z_i := (Z_i)_{-i \in \mathbb{N}}$, with D, X, Y, Z discrete. Suppose further that $(\mathbb{P}_C, (D, Z), (X, Y))$ is preempted. Let G be the directing random measure of (\mathbb{P}_C, Z, D) and H the directing random measure of $(\mathbb{P}_C, (D, Z), (X, Y))$. Suppose for all i , $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e Z_i | (H, X_i, C)$. Finally suppose for each $h \in H$ and $d \in D$ and $z, z' \in Z$,*

$$\mathbb{P}_\alpha^{G_z^d | HG_{z'}^d}(\cdot | h, g_{z'}^d) \ll U_{[0,1]} \quad (5.2)$$

where $U_{[0,1]}$ is the uniform probability measure on $([0, 1], \mathcal{B}([0, 1]))$ – that is, the Lebesgue measure on $[0, 1]$ restricted to the Borel sets.

Then $\mathbb{P}_\alpha^{Y_i | HX_i}$ are independent and identical responses for all $-i \in \{0\} \cup \mathbb{N}$.

Proof. First, define matrices k and l by

$$\begin{aligned} \mathbb{P}_\alpha^{Y_i | HD_i Z_i X_i}(y|h, d, z, x) &\stackrel{\mathbb{P}_\alpha}{\cong} k_{d z x}^y \\ \mathbb{P}_\alpha^{X_i | HD_i Z_i}(x|h, d, z) \mathbb{P}_\alpha &\cong l_{d z}^x \end{aligned}$$

noting that both k and l are almost surely deterministic functions of h by

independence lemma

The assumption $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e Z_i | (H, X_i, C)$ implies, for \mathbb{P}_α -almost all $k, l, \alpha, z, z', x, y$

$$\sum_{d \in D} k_{d zx}^y \frac{l_{dz}^x g_z^d}{\sum_{d' \in D} l_{d' z}^x g_z^{d'}} = \sum_{d \in D} k_{d z' x}^y \frac{l_{dz'}^x g_{z'}^d}{\sum_{d' \in D} l_{d' z'}^x g_{z'}^{d'}} \quad (5.3)$$

Fixing k, l and g_z^d , Eq. (5.3) defines a polynomial constraints on $g_{z'}^d$. We will show that, unless $k_{d zx}^y = k_{d' zx}^y$ for all d, d' and z then this constraint is nontrivial for some z' . For arbitrary d , without loss of generality, assume $k_{d z' x}^y > k_{d' z' x}^y$ for some $d' <$.

Then either $l_{dz'}^x = l_{d' z'}^x$, $l_{dz'}^x < l_{d' z'}^x$, or $l_{dz'}^x = l_{d' z'}^x$. Consider the first case, and take g' such that $g' d_{z'} = g_{z'}^d - \epsilon$ and $g' d_{z'}^< = g_{z'}^{d'} + \epsilon$ and equal to g otherwise. There is almost surely some ϵ such that g' is a Markov kernel, as $g_{z'}^d > 0$ almost surely. Then

$$\begin{aligned} \frac{l_{dz}^x g_z^d}{\sum_{d' \in D} l_{d' z}^x g_z^{d'}} &> \frac{l_{dz}^x g_z^{d'}}{\sum_{d' \in D} l_{d' z}^x g_z^{d'}} \\ \frac{l_{d' z}^x g_z^{d'}}{\sum_{d' \in D} l_{d' z}^x g_z^{d'}} &< \frac{l_{d' z}^x g_z^{d'}}{\sum_{d' \in D} l_{d' z}^x g_z^{d'}} \end{aligned}$$

because by assumption the denominator remains the same. But then

$$\sum_{d \in D} k_{d zx}^y \frac{l_{dz}^x g_z^d}{\sum_{d' \in D} l_{d' z}^x g_z^{d'}} > \sum_{d \in D} k_{d z' x}^y \frac{l_{dz'}^x g_{z'}^d}{\sum_{d' \in D} l_{d' z'}^x g_{z'}^{d'}} \quad (5.4)$$

because on the left side a larger term in the sum receives more weight, a smaller term receives less weight and all other terms are weighted equally.

Consider $l_{dz'}^x > l_{d' z'}^x$. Then we still have

$$\begin{aligned} \frac{l_{dz}^x g_z^d}{\sum_{d' \in D} l_{d' z}^x g_z^{d'}} &> \frac{l_{dz}^x g_z^{d'}}{\sum_{d' \in D} l_{d' z}^x g_z^{d'}} \\ \frac{l_{d' z}^x g_z^{d'}}{\sum_{d' \in D} l_{d' z}^x g_z^{d'}} &< \frac{l_{d' z}^x g_z^{d'}}{\sum_{d' \in D} l_{d' z}^x g_z^{d'}} \end{aligned}$$

For the first inequality, both the numerator and the denominator shrink. For the second, note that

$$\begin{aligned} \frac{l_{d' z}^x g_z^{d'}}{\sum_{d' \in D} l_{d' z}^x g_z^{d'}} &= \frac{l_{d' z}^x (g_z^{d'} + \epsilon)}{(1 + \frac{\epsilon}{g_z^{d'}}) \sum_{d' \in D} l_{d' z}^x g_z^{d'}} \\ &< \frac{l_{d' z}^x (g_z^{d'} + \epsilon)}{\sum_{d' \in D} l_{d' z}^x g_z^{d'}} \end{aligned}$$

and so the conclusion in Eq. (5.4) follows for the same reasons. Finally considering $l_{dz'}^x < l_{d' z'}^x$, analogous reasoning implies Eq. (5.4) once again.

Thus, unless $k_{d zx}^y = k_{d' zx}^y$ for all d, d' and z , Eq. (5.3) implies a nontrivial constraint on $g_{z'}^d$ for some z' . Thus, for some z', x, d, d', y the set of solutions

$A := \{g_{z'}^d | \text{Eq. (5.3) is satisfied} \wedge k_{d'zx}^y \neq k_{d'zx}^y\}$ has Lebesgue measure 0 in the set $[0, 1]^D$ (Okamoto, 1973), and so finally

$$\mathbb{P}_{\alpha}^{\mathbb{G}_{z'}^d | \text{HG}_{z'}^d}(A|h, g_z^d) = 0$$

by the assumption that this probability is dominated by the Lebesgue measure. On the other hand, by assumption the set $B := \{g_{z'}^d | \text{Eq. (5.3) is satisfied}\}$ has measure 1. Thus we conclude that $k_{d'zx}^y = k_{d'zx}^y$ with probability 1. That is, $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e D_i | (H, Z_i, X_i, C)$. Thus, by contraction with $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e Z_i | (H, X_i, C)$, we have $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (Z_i, D_i) | (H, X_i, C)$, which implies $\mathbb{P}_{\alpha}^{Y_i | \text{HX}_i}$ are independent and identical response functions. \square

We will now specify the medical example in more detail

Example 5.1.19. Let $Y := (Y_i)_{-i \in \{0\} \cup \mathbb{N}}$ be associated with patient outcomes, $D := (D_i)_{-i \in \{0\} \cup \mathbb{N}}$ with treatment plans (including, for example, what assessments are made, what other treatments are used and so forth), some of which the decision maker can consider, $X := (X_i)_{-i \in \{0\} \cup \mathbb{N}}$ to be the recommendation of a particular treatment and $Z_i := (Z_i)_{-i \in \mathbb{N}}$ to be other doctor's identifiers in the dataset.

Assume D_i screens off Z_i from X_i and Y_i – that is, the latent treatment plan is sufficiently detailed to screen off the relevance of the doctor's identity (this is a causal assumption), and the stochastic response to treatment plans is identical for each patient with positive support in the past data for each plan the decision maker is considering. That is, assume $(\mathbb{P}_C, D, (X, Y))$ is preempted. By the assumption that D_i screens off Z_i , we can also conclude that $(\mathbb{P}_C, (D, Z), (X, Y))$ is preempted.

Suppose that each doctor makes choices \mathbb{G}_z^d by some deterministic function of their beliefs of what every other doctor does $\mathbb{G}_{z'}^e$, and of the effect of treatment plans H , but they estimate both $\mathbb{G}_{z'}^e$ and H with continuously distributed noise. Then their beliefs, and hence (by supposition) their choices end up dominated by the uniform measure on $[0, 1]$.

If the decision maker is then told by an oracle that for all i , $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e Z_i | (H, X_i, C)$, she may then conclude that $\mathbb{P}_{\alpha}^{Y_i | \text{HX}_i}$ are independent and identical responses for all $-i \in \{0\} \cup \mathbb{N}$.

This story makes a number of assumptions with a causal character. First, the assumption of preemption, which we've acknowledged isn't perfectly worked out, is taken here quite literally – we are not just supposing that the problem can be posed “as if” the consequences of our choices had been previously realised, but we are actually taking the D_i s to literally stand for unobserved choices that actual doctors make. Secondly, the assumption that the unobserved D_i s screen off the doctor's identities is reminiscent of the idea from the causal graphical models tradition that variables have complete sets of causal parents, some of which may be unobserved, but all of which together screen that variable off from other nondescendant variables. Note that this assumption isn't required by Theorem 5.1.18, but it supports the particular story being told here.

Finally, the idea that the long run frequencies of each doctor’s choices are generic conditioned on the other doctor’s choices and the long run input-output relationship seems closely related to the idea of “independent causal mechanisms”. This comes up in two forms in the literature. First, it is used to justify the assumption of *causal faithfulness*: here, it is shown that *if* one makes an assumption that conditional probabilities in a causal model are generic with respect to one another in a manner similar to Equation (5.2), then causal faithfulness holds with probability 1 (Meek, 1995). However, it’s been noted that conditional probabilities routinely do line up in “non-generic” ways in an anti-causal direction.

Interestingly, Theorem 5.1.18 itself doesn’t depend on a notion of causal direction, and merely shows that a conclusion of independent and identical response functions follows from an assumption of preemption and an assumption of “generic mechanism association”. Example 5.1.19 shows one way that this generic association could be argued for.

Note that in that example, there is no reason not to expect that each doctor doesn’t select a mixture of treatment plans that is *close* to having support at a special singleton – in fact, it is assumed that doctors try to take into account the response of patient outcomes to treatment plans and the actions of other doctors, and that they simply fail to do this perfectly. We also cheat by having an oracle tell the decision maker that the key conditional independence holds. In particular, we ask the decision maker to conclude something precise about \mathbf{H} (namely, the key conditional independence) while also assuming that none of the other doctors are able to do this.

There is a substantial literature that aims to draw causal conclusions from observational data by first applying a graph learning algorithm to a sequence of observational data, and then using the graph obtained as a DAG for a causal Bayesian network. Earlier examples treat the graph learning problem as a discrete optimization problem and include the PC algorithm and the Causal Inference Algorithm Spirtes et al. (2000, Ch. 5& 6) and Greedy Equivalent Search Chickering (2003, 2002). More recent examples pose graph learning as a continuous optimization problem Zheng et al. (2018); Ng et al. (2019). Underpinning all of these approaches are a number of key assumptions, which include the assumption of *faithfulness* – that missing edges in the learned graph correspond to missing edges in the appropriate causal DAG – and often also the assumption of *causal sufficiency*, which is the assumption that there are “no relevant unobserved variables”. Together, these assumptions imply that certain conditional independences in the observational data sequence imply the same conditional independences in the data produced under intervention. One open question we raise is: can this implication also be understood as a special case of the interventional data being preempted by the observational data?

On unobservability

The fact that we are offering the assumption that covariates are unobservable as an informal assumption is due to the fact that we are limiting our attention to

data-independent models (recall Definition 4.3.16). In these models, actions never depend on the available data, and choosing some action based on observations must happen outside the model. If we were considering some data-dependent variation of a causal Bayesian network, the fact that $V_{1[n]}$ is unobserved would have formal implications for our model. For example, if V_{1i} is unobserved for all i while V_{2i} is directly controlled for all i , then we should require that, under every choice α , V_{2i} is independent of $V_{1[1,i]}$ given $V_{\{2,3\}[1,i]}$ – that is, there is no choice that induces the controlled variable V_{2i} to be dependent on the history of unobserved variables $V_{1[1,i]}$, given the history of the observed variables.

5.2 What is a Potential Outcomes model?

Potential outcomes is another popular framework for modelling causal problems. There are two key differences between the potential outcomes approach and the causal Bayesian network approach: potential outcomes models are “unrolled by default” and they feature no notion of “intervention”. A third difference relates to the possibility of expressing “counterfactual” statements, although this difference seems to be contingent on the particular manner we use to unroll a causal Bayesian network – see Section ??, and recall from Section 5.1.2 that we had to make some choices in our construction of unrolled causal Bayesian networks, Definition 5.1.12.

Thus, to formulate a decision making model from a potential outcomes model, we do have to make a judgement about what the “choices” are (while CBNs provide the notion of “intervention” for this role), while we do not need to make any judgements about how to unroll a potential outcomes model, because this is already given. For the following, we rely on Rubin (2005) for the definition of a potential outcomes model.

Our definition of potential outcomes has a lot in common with the tabulated conditional distribution (Definition 4.3.12). However, it is different: in particular, $\mathbb{P}_\alpha^{Y_i|Y_i^D D_i}(y^d|y^D, d) = 1$, which is usually false for Definition 4.3.12.

Definition 5.2.1 (Potential outcomes). Given $(\mathbb{P}_C, \Omega, \mathcal{F})$ and, for some i , variables $D_i : \Omega \rightarrow D$ (D denumerable), $Y_i : \Omega \rightarrow Y$ and $Y_i^D : \Omega \rightarrow Y^D$, Y_i^D is a vector of *potential outcomes* with respect to D_i for all α

$$\mathbb{P}_\alpha^{Y_i|Y_i^D D_i} = \begin{array}{c} Y_i^D \\ \text{---} \quad \text{---} \quad \text{---} \\ D_i \text{---} \end{array} \boxed{\mathbb{F}_{\text{lus}}} \text{---} Y_i$$

Where \mathbb{F}_{lus} is the Markov kernel associated with the single-shot lookup map

$$\begin{aligned} \text{lus} : Y^D \times D &\rightarrow Y \\ (d, (y_i)_{i \in D}) &\mapsto y_d \end{aligned}$$

Note that $|D|$ copies of Y_i ($Y_i, Y_i, Y_i, \dots, Y_i$) always satisfies Definition 5.2.1. This definition is not the sole constraint on potential outcomes, but the additional

constraints come from what we want them to model, and are therefore not able to be formally stated.

A “potential outcomes model” is simply a probability map with potential outcomes. Traditionally, potential outcomes models did not feature any choices. That is, a “traditional” potential outcomes model is a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ rather than a probability function. We extend this to probability functions in what we think is the obvious way - to replace the probability space with a probability function space.

Definition 5.2.2 (Potential outcomes model). $(\mathbb{P}_C, \Omega, \mathcal{F})$ is a potential outcomes model with respect to $\mathbf{Y}^D := (\mathbf{Y}_i^D)_{i \in A}$, $\mathbf{Y} := (\mathbf{Y}_i)_{i \in A}$ and $(\mathbf{D}_i)_{i \in A}$ if \mathbf{Y}_i^D is a vector of potential outcomes with respect to \mathbf{D}_i and \mathbf{Y}_i for all $i \in A$.

Theorem 5.2.3. A potential outcomes model $(\mathbb{P}_C, \Omega, \mathcal{F})$ with respect to $\mathbf{D}_i : \Omega \rightarrow D$, $\mathbf{Y}_i : \Omega \rightarrow Y$ and $\mathbf{Y}_i^D : \Omega \rightarrow Y^D$, \mathbf{Y}_i^D has $\mathbf{Y} \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathbf{C} | (\mathbf{D}, \mathbf{Y}^D)$ and $\mathbb{P}_C^{\mathbf{Y} | \mathbf{Y}^D \mathbf{D}}$ is IO contractible (with respect to $*$).

Proof. IO contractibility of $\mathbb{P}_C^{\mathbf{Y} | \mathbf{Y}^D \mathbf{D}}$ follows from the fact that \mathbf{Y}_i is deterministic given \mathbf{Y}_i^D and \mathbf{D}_i , and thus $\mathbf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (\mathbf{D}_{\{i\}^c}, \mathbf{Y}_{\{i\}^c}, \mathbf{C}) | (\mathbf{Y}_i^D, \mathbf{D}_i)$. Furthermore, for all i, j

$$\mathbb{P}_C^{\mathbf{Y}_i | \mathbf{Y}_i^D \mathbf{D}_i} = \mathbb{P}_C^{\mathbf{Y}_j | \mathbf{Y}_j^D \mathbf{D}_j}$$

hence the $\mathbb{P}_C^{\mathbf{Y}_i | \mathbf{Y}_i^D \mathbf{D}_i}$ are independent and identical response functions conditional on $*$.

From Definition 5.2.1, $\mathbb{P}_\alpha^{\mathbf{Y}_i | \mathbf{Y}_i^D \mathbf{D}_i}$ is the same for all $\alpha \in C$, and by the argument above,

$$\mathbb{P}_C^{\mathbf{Y}_i | \mathbf{Y}_i^D \mathbf{D}_i \mathbf{Y}_{\{i\}^c}^D \mathbf{C}^{\mathbf{D}_{\{i\}^c}}} = \mathbb{P}_C^{\mathbf{Y}_i | \mathbf{Y}_i^D \mathbf{D}_i} \otimes \text{del}_{Y^D \times A \setminus \{i\} \times D \setminus A}$$

hence

$$\mathbb{P}_C^{\mathbf{Y} | \mathbf{Y}^D \mathbf{D}} = \bigotimes_{i \in A} \mathbb{P}_C^{\mathbf{Y}_i | \mathbf{Y}_i^D \mathbf{D}_i}$$

hence $\mathbf{Y} \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathbf{C} | (\mathbf{D}, \mathbf{Y}^D)$. □

A key theorem of potential outcomes is that, if \mathbf{D} is “strongly ignorable” with respect to \mathbf{Y}^D , then the average treatment effect is identified. “Strong ignorability” here means that the probability $\mathbb{P}_\alpha^{\mathbf{D}_i}(d) > 0$ for each d and for each choice α the inputs \mathbf{D} are independent of the potential outcomes \mathbf{Y}^D given the covariates and the choice. We reproduce this theorem in terms of IO contractibility. Note that Theorem 5.2.4 applies to potential outcomes models with sets of choices, rather than simply to single probability distributions.

Theorem 5.2.4 (Potential outcomes identifiability). If $(\mathbb{P}_C, \Omega, \mathcal{F})$ is a potential outcomes model with respect to $\mathbf{Y}^D := (\mathbf{Y}_i^D)_{i \in \mathbb{N}}$, $\mathbf{Y} := (\mathbf{Y}_i)_{i \in \mathbb{N}}$ and $(\mathbf{D}_i)_{i \in \mathbb{N}}$, each

value of D occurs infinitely often with probability 1, there is some $\mathbf{X} := (X_i)_{i \in \mathbb{N}}$ such that $\mathbb{P}_\alpha^{Y^D|\mathbf{X}}$ is exchangeable for all α and $D \perp\!\!\!\perp_{\mathbb{P}_C}^e Y^D | (X, Y, C)$ and for each α \mathbb{P}_α^D is absolutely continuous with respect to some exchangeable distribution in $\Delta(D^\mathbb{N})$, then there is some W such that for all α $\mathbb{P}_\alpha^{Y|W\mathbf{X}D}$ is IO contractible over W .

Proof. By exchangeability of $\mathbb{P}_\alpha^{Y^D|\mathbf{X}}$, $\mathbb{P}_\alpha^{Y^D|\mathbf{X}}$ commutes with exchange. Because Y is deterministic given D and Y^D , $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e (X, C) | (Y^D, D)$. Thus, for some finite permutation ρ , by IO contractibility of $\mathbb{P}_C^{Y|Y^D D}$

$$\begin{aligned}
 \mathbb{P}_\alpha^{Y|\mathbf{X}D} &= \begin{array}{c} X \text{ --- } \boxed{\mathbb{P}_\alpha^{Y^D|\mathbf{X}}} \\ D \text{ --- } \boxed{\mathbb{P}_C^{Y|Y^D D}} \end{array} \text{ --- } Y \\
 &= \begin{array}{c} X \text{ --- } \boxed{\mathbb{P}_\alpha^{Y^D|\mathbf{X}}} \\ D \text{ --- } \boxed{\text{swap}_\rho} \end{array} \begin{array}{c} \boxed{\text{swap}_\rho} \\ \boxed{\mathbb{P}_C^{Y|Y^D D}} \end{array} \text{ --- } \boxed{\text{swap}_{\rho^{-1}}} \text{ --- } Y \\
 &= \begin{array}{c} X \text{ --- } \boxed{\text{swap}_\rho} \\ D \text{ --- } \boxed{\text{swap}_\rho} \end{array} \begin{array}{c} \boxed{\mathbb{P}_\alpha^{Y^D|\mathbf{X}}} \\ \boxed{\mathbb{P}_C^{Y|Y^D D}} \end{array} \text{ --- } \boxed{\text{swap}_{\rho^{-1}}} \text{ --- } Y
 \end{aligned}$$

IO contractibility of $\mathbb{P}_\alpha^{Y|W\mathbf{X}D}$ over some W follows from Theorem 4.3.25. \square

5.3 Individual-level response functions

Exchangeability of potential outcomes, a key assumption in Theorem 5.2.4, is hard to explain in terms of symmetries we actually expect to see in the world. Given some experiment producing a sequence of pairs $(D_i, Y_i)_{i \in \mathbb{N}}$, say where D_i s are treatment administrations and Y_i s are health outcomes, there's no obvious generic way to design a related experiment whose model is the same as the original except with potential outcomes Y_i^D interchanged. This is in sharp contrast to the assumption of exchangeability of observed outcomes - say, instead of the potential outcomes being exchangeable, we hold that the pairs (D_i, Y_i) are exchangeable in the original experiment. Then we commit ourselves to the proposition that an alternative experiment which proceeds exactly as the first except, before being “committed to memory”, the experimental results are interchanged should be modeled exactly as the first.

One could propose that exchangeability of potential outcomes in our example experiment corresponds to an *exchangeability of patients*; perhaps, if we believe the model should be unchanged after we swap the order in which patients are seen, then we should accept that the model has exchangeable potential outcomes. First, note that this isn't a generic transformation like swapping labels - it depends on the experiment featuring a sequence of patients who can be swapped. Secondly, this proposition depends on some assumption that ties patients to potential outcomes. For example, if each patient were assumed to have a fixed

but unknown vector of potential outcomes that is unchanged by the swapping operation, then swapping patients does indeed correspond to swapping potential outcomes.

We formalise the idea of “potential outcomes attached to individuals” as *individual-level response functions*. We offer a formal definition of the assumption of individual-level response functions, but like exchangeability of potential outcomes it is difficult to understand fundamentally what this assumption entails, or what it might be motivated by. Nevertheless, it does allow us to separate the assumption of “exchangeable potential outcomes” into the assumption of individual level response functions and the assumption of exchangeability of individuals. We also use this notion to prove Theorem 5.3.8. At a high level, it plays a similar role to Theorem 5.2.4: it seems to justify causal identifiability in certain kinds of controlled experiments. However, the content of the two theorems is very different. While Theorem 5.2.4 concerns independence of the inputs and potential outcomes along with exchangeability of the potential outcomes, Theorem 5.3.8 says (informally) if:

- There are individual-level response functions
- Individuals are exchangeable
- Inputs are deterministically controlled by the choice
- There is only one choice for each value of the inputs

then the model is also IO contractible with respect to the inputs and the outputs only (ignoring the individual identifiers). In our view, this comes closer to a set of assumptions that are directly applicable to a controlled experiment than those in Theorem 5.2.4, and reflects Kasy (2016)’s dictum that, for the identifiability of causal effects, a “controlled experiment” is sufficient.

5.3.1 References to individual-level IO contractibility

The role of individuals has often been mentioned in literature on causal inference. For example, Greenland and Robins (1986) explain

Equivalence of response type may be thought of in terms of exchangeability of individuals: if the exposure states of the two individuals had been exchanged, the same data distribution would have resulted.

Here, the idea of “exchangeable individuals” plays a role in the author’s reasoning about model construction, but “individuals” are not actually referenced by the resulting model, and “exchanging individuals” does not correspond to a model transformation.

Dawid (2020) suggests (with some qualifications) that “post-treatment exchangeability” for a decision problem regarding taking aspirin to treat a headache may be acceptable if the data are from

A group of individuals whom I can regard, in an intuitive sense, as similar to myself, with headaches similar to my own.

As in the previous work, the similarity of individuals involved in an experiment is raised when justifying particular model constructions, but the individuals are not referenced by the model.

Pearl (2009, pg. 98) writes

Although the term unit in the potential-outcome literature normally stands for the identity of a specific individual in a population, a unit may also be thought of as the set of attributes that characterize that individual, the experimental conditions under study, the time of day, and so on all of which are represented as components of the vector u in structural modeling.

Once again, the idea of an individual (or a particular set of conditions) is raised in the context of explaining modelling choices. Unlike the previous authors, Pearl introduces a vector u to stand for the “unit”. However, he subsequently assumes that u is a sequence of *independent samples* from some distribution. This seems to contradict an important feature of “individuals” or “units”: individuals are typically supposed to be unique, a property that will usually not be satisfied by independently sampling from some distribution (at least, as long as the distribution is discrete).

Finally, Rubin (2005) writes:

Here there are N units, which are physical objects at particular points in time (e.g., plots of land, individual people, one person at repeated points in time).

Note that Rubin’s explanation of *units* guarantees that they are unique: they are particular things at particular times. These units are associated with input-output functions (the *potential outcomes*), which are later assumed to be exchangeable:

the indexing of the units is, by definition, a random permutation of $1, \dots, N$, and thus any distribution on the science must be row-exchangeable

Our proposition is: can the intuition that unique individuals are an important for the motivation for causal models, be captured by considering models that feature “unique identifier” variables referencing these unique individuals?

5.3.2 Unique identifiers

A sequence of *unique identifiers* is a vector of finite or infinite length such that no two coordinates are equal. We are interested in models that assign positive probability to any particular coordinate having any particular value. This is straightforward in the finite case. In the infinite case, note that a vector of $|\mathbb{N}|$ unique values with an arbitrary entry k in the j th coordinate can be obtained by starting with $(i)_{i \in \mathbb{N}}$ and then transposing j with k . More generally, we consider infinite length sequences of unique identifiers to be elements of the set of finite permutations $\mathbb{N} \rightarrow \mathbb{N}$.

Definition 5.3.1 (Measurable space of unique identifiers). The measurable space of unique identifiers (I, \mathcal{I}) is the set I of finite permutations $\mathbb{N} \rightarrow \mathbb{N}$ with the discrete σ -algebra \mathcal{I} .

The set I is countable, as it is the countable union of finite subsets (i.e. the permutations that leave all but the first n numbers unchanged for all n).

Definition 5.3.2 (Unique identifier). Given a sample space (Ω, \mathcal{F}) , a *sequence of unique identifiers* $\mathcal{I} : \Omega \rightarrow I$ is a variable taking values in I .

The values of each coordinate of sequence of unique identifiers is just called an identifier (for obvious reasons, we don't call it an identity).

Definition 5.3.3 (Identification). Given \mathbf{l} , define the i -th *identifier* $\mathbf{l}_i := \text{ev}(i, \mathbf{l})$, where $\text{ev} : \mathbb{N} \times I \rightarrow \mathbb{N}$ is the evaluation map $(i, f) \mapsto f(i)$.

For *any* sample space (Ω, \mathcal{F}) we can define a trivial \mathcal{I} that maps every $\omega \in \Omega$ to $(1, 2, 3, \dots) =: (\mathbb{N})$. In this case, the identifiers are all known by the modeller at the outset. Using this sequence of identifiers renders exchange commutativity trivial.

Example 5.3.4. Given a sequential input-output model $(\mathbb{P}_C, (D, \mathbf{l}), Y)$ where \mathbf{l} is the identifier variable $\omega \mapsto (\mathbb{N})$, $\mathbb{P}_\alpha^{Y|\mathbf{D}\mathbf{l}}$ commutes with exchange.

This is because for any permutation $\rho : \mathbb{N} \rightarrow \mathbb{N}$ except the identity, $\mathbb{P}_\alpha^{Y|\mathbf{D}\mathbf{l}}$ and $\text{swap}_\rho \mathbb{P}_\alpha^{Y|\mathbf{D}\mathbf{l}}$ will have no common support; the first will be supported on $\mathbf{l} \bowtie (\mathbb{N})$ only, and the second only on $\mathbf{l} \bowtie \rho(\mathbb{N})$.

We are particularly interested in models where exchange commutativity is not trivial, so we focus on the case where each identifier \mathbf{l}_i has some non-zero probability of taking any value in \mathbb{N} .

5.3.3 Identification with individual-level response functions

The key result of this section is Theorem 5.3.8. A key assumption for this theorem is the assumption of “individual-level response functions”. That is, the assumption that, given a sequential input-output model $(\mathbb{P}_C, (D, \mathbf{l}), Y)$, \mathbb{P}_C features independent and identical response functions conditional on some variable J (unlike H from Definition 4.3.14, J is not necessarily a function of the inputs and outputs). We also assume that each individual identifier \mathbf{l}_i has positive probability of taking on any particular value.

This assumption is somewhat difficult to understand. If we imagine that the identifiers \mathbf{l}_i are, for example, patient names in some medical experiment, it rules out certain possibilities. For example, this assumption is incompatible with the idea that, if Tina is first in line, then she will be deterministically cured by the treatment while if she is last in line she will be deterministically not-cured by it. In fact, it says precisely that, conditional on J , according to the model Tina will respond in the same stochastic way no matter where in line she is, no matter

which other patients have been seen and no matter what their treatments or outcomes were. Unlike the “directing random measure” from Definition 4.3.14, J is not some long-term limit of observations. Rather, it seems to be just a choice we could make for how we parametrize our model.

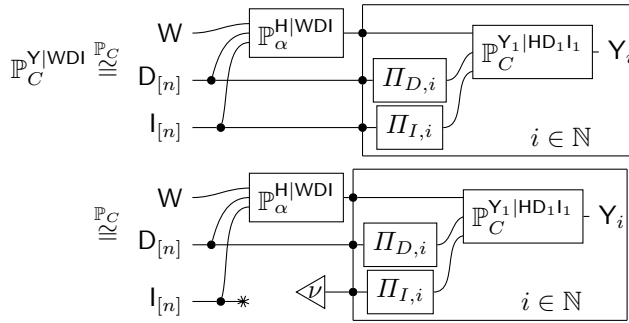
Before proving Theorem 5.3.8, we prove a number of lemmas and a preliminary theorem. Lemma 5.3.5 and Theorem ?? do *not* require that l be a sequence of unique identifiers, they hold just as well if it is a sequence of non-unique labels; that is, if $l_i \bowtie l_j$ had positive measure for some $i \neq j$. The reason why we are interested mainly in the case where l is a sequence of unique identifiers is that the assumption $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e l|C$ is substantially more limiting in the case that l is a non-unique sequence of labels. In particular, it implies that the conditional probability of (Y_1, Y_2) given $(l_1 = 1, l_2 = 1)$ is exactly the same as the conditional probability of (Y_1, Y_2) given $(l_i = 1, l_2 = 2)$; observations associated with equal labels are no more relevant than observations associated with different labels.

In the following, it is helpful to assume that each sub-experiment has a “unique identifier” l_i , with the sequence of all sub-experiment labels given by l . With this, if $\mathbb{P}_C^{Y|Dl}$ is assumed IO contractible, then it’s possible to talk about the individual response functions $\mathbb{P}_C^{Y_i|l_iHD_i}$. These plays a role very similar to the i th vector of potential outcomes Y_i^D . Because l_i is unique (i.e. never equal to l_j for $j \neq i$), only one observation of any individual is ever given, just like only one potential outcome is ever observed.

Theorem 5.3.8 can also be extended to the case where D is a function of the choice α and a “random signal” R , as in Theorem 5.3.9.

Lemma 5.3.5. *Given sequential input-output model $(\mathbb{P}_C, (D, l), Y)$ with $\mathbb{P}_\alpha^{Y|WDl}$ IO contractible over W , if $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e (l, C)|(W, D)$ and for any $j \in I$, $\sum_{\alpha \in C} \mathbb{P}_\alpha^{l_i}(j) > 0$, then there is some W' such that $\mathbb{P}_\alpha^{Y|W'D}$ is also IO contractible over W .*

Proof. Fix arbitrary $\nu \in \Delta(I^\mathbb{N})$ such that $\sum_{\alpha \in C} \mathbb{P}_\alpha^{l_i} \gg \nu$. By assumption of IO contractibility of $\mathbb{P}_C^{Y|WDl}$ and Theorem 4.3.22



Where $\Pi_{D,i} : D^\mathbb{N} \rightarrow D$ projects the i th coordinate, and similarly for $\Pi_{Y,i}$.

In particular, for any $i \in \mathbb{N}$, $j \in I$, this holds for some ν such that $\nu(\Pi_{Y,i}^{-1}(j)) = 1$ and by extension for any finite $A \subset \mathbb{N}$ we can find ν such

that $\nu(\Pi_{Y,i}^{-1}(j)) = 1$ for all $i \in A, j \in I$. Thus, for any $n \in \mathbb{N}$

$$\mathbb{P}_C^{Y_{[n]}|WD_{[n]}I_{[n]}} \stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} \begin{array}{c} W \\ D_{[n]} \\ I_{[n]} \end{array} \begin{array}{c} \mathbb{P}_\alpha^{H|WDI} \\ \Pi_{D,i} \\ \delta_j \end{array} \begin{array}{c} \mathbb{P}_C^{Y_i|HD_1I_1} \\ Y_i \end{array} \\ i \in [n] \end{array} \quad (5.5)$$

$$\stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} \begin{array}{c} W \\ D_{[n]} \end{array} \begin{array}{c} \mathbb{P}_\alpha^{H|WD} \\ \Pi_{D,i} \end{array} \begin{array}{c} \mathbb{P}_C^{Y_i|HD_1} \\ Y_i \end{array} \\ i \in [n] \end{array} \quad (5.6)$$

where Equation (5.5) follows from Theorem 2.3.6 and Equation (5.6) follows from the fact that Equation (5.5) holds for arbitrary $j \in I$.

Thus by Lemma 4.3.6

$$\mathbb{P}_C^{Y|WD} = \begin{array}{c} W \\ D_{[n]} \end{array} \begin{array}{c} \mathbb{P}_\alpha^{H|WD} \\ \Pi_{D,i} \end{array} \begin{array}{c} \mathbb{P}_C^{Y_i|HD_1I_1} \\ Y_i \end{array} \\ i \in \mathbb{N}$$

Applying Theorem 4.3.22, $\mathbb{P}_C^{Y|WD}$ is IO contractible over W . \square

Theorem 5.3.6. *Given a sequential input-output model $(\mathbb{P}_C, (D, I), Y)$ on (Ω, \mathcal{F}) with Y standard measurable and C countable, if there is some J such that for each α*

$$\begin{aligned} \mathbb{P}_\alpha^{Y_i|J_iD_i} &= \mathbb{P}_\alpha^{Y_i|J_iD_i} \\ Y_i &\perp\!\!\!\perp_{\mathbb{P}_C}^e (I_{\{i\}^c}, D_{\{i\}^c}) | (J, I_i, D_i) \end{aligned} \quad \forall i, j \in \mathbb{N}$$

and

$$\begin{aligned} Y &\perp\!\!\!\perp_{\mathbb{P}_C}^e I | C \\ YI &\perp\!\!\!\perp_{\mathbb{P}_C}^e D | C \\ YI &\perp\!\!\!\perp_{\mathbb{P}_C}^e C | D \\ \forall i, j \in \mathbb{N} : &\sum_{\alpha \in C} \mathbb{P}_\alpha^{I_i}(j) > 0 \end{aligned}$$

then $\mathbb{P}_C^{Y|JD}$ is IO contractible over J .

Proof. For any $\alpha \in C$

$$\begin{aligned} \mathbb{P}_\alpha^{YJ|I} &= \begin{array}{c} I \text{ --- } \bullet \text{ --- } \boxed{\mathbb{P}_\alpha^{D|I}} \text{ --- } \boxed{\mathbb{P}_C^{YJ|ID}} \text{ --- } (Y, J) \end{array} \\ &= \begin{array}{c} I \text{ --- } \boxed{\mathbb{P}_C^{YJ|ID}} \text{ --- } (Y, J) \\ \nearrow \boxed{\mathbb{P}_\alpha^D} \end{array} \end{aligned}$$

Define \mathbb{Q} by $\alpha \mapsto \mathbb{P}_\alpha$ and $\mathbb{Q}^{\cdot|C}$ by $\alpha \mapsto \mathbb{P}_\alpha^*$ and \mathbb{Q}^C is an arbitrary distribution in $\Delta(C)$ with full support. Note that the support of \mathbb{Q}^{IDYJ} is the union of the support of \mathbb{P}_α^{IDYJ} for all α . Then

$$\mathbb{Q}^{YJ|IC} \stackrel{\mathbb{Q}}{\cong} \begin{array}{c} I \text{ --- } \boxed{\mathbb{P}_C^{YJ|ID}} \text{ --- } (Y, J) \\ C \text{ --- } \boxed{\mathbb{Q}^{D|C}} \end{array}$$

By assumption $YI \perp\!\!\!\perp_{\mathbb{P}_C}^e D|C$, it is also the case that

$$\begin{aligned} \mathbb{Q}^{Y|ID} &\stackrel{\mathbb{Q}}{\cong} \begin{array}{c} I \text{ --- } \bullet \text{ --- } \boxed{\mathbb{Q}^{Y|IC}} \text{ --- } Y \\ D \text{ --- } \boxed{\mathbb{Q}^{C|ID}} \end{array} \\ &\stackrel{\mathbb{Q}}{\cong} \begin{array}{c} I \text{ --- } \boxed{\mathbb{Q}^{YJ|IC}} \text{ --- } (Y, J) \\ D \text{ --- } \boxed{\mathbb{Q}^{C|D}} \end{array} \\ &\stackrel{\mathbb{Q}}{\cong} \begin{array}{c} I \text{ --- } \boxed{\mathbb{P}_C^{YJ|ID}} \text{ --- } (Y, J) \\ D \text{ --- } \boxed{\mathbb{Q}^{C|D}} \text{ --- } \boxed{\mathbb{Q}^{D|C}} \end{array} \end{aligned}$$

But

$$\begin{aligned} \mathbb{Q}^{Y|ID} &= \sum_{\alpha \in C} \mathbb{P}_\alpha^{Y|ID} \mathbb{Q}^C(\alpha) \\ &= \mathbb{P}_C^{Y|ID} \\ \Rightarrow \begin{array}{c} I \text{ --- } \boxed{\mathbb{P}_C^{YJ|ID}} \text{ --- } (Y, J) \\ D \text{ --- } \boxed{\mathbb{Q}^{C|D}} \text{ --- } \boxed{\mathbb{Q}^{D|C}} \end{array} &= \mathbb{P}_C^{Y|ID} \end{aligned}$$

Furthermore, by assumption $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e I|C$, so there is some $\mathbb{K} : C \rightarrow Y \times W$

such that

$$\begin{aligned}
 \mathbb{Q}^{YJ|IC} &\stackrel{\mathbb{Q}}{\cong} \begin{array}{c} I \text{ --- } * \\ D \text{ --- } \end{array} \boxed{K} \text{ --- } (Y, J) \\
 \Rightarrow \mathbb{P}_C^{YJ|ID} &= \begin{array}{c} I \text{ --- } \boxed{F_\rho} \\ D \text{ --- } \boxed{Q^{C|D}} \text{ --- } \boxed{Q^{D|C}} \end{array} \boxed{\mathbb{P}_C^{YJ|ID}} \text{ --- } (Y, J) \\
 &= \begin{array}{c} I \text{ --- } * \\ D \text{ --- } \end{array} \boxed{\mathbb{P}_C^{YJ|C}} \text{ --- } (Y, J)
 \end{aligned}$$

Then by Lemma 5.3.5, $\mathbb{P}_C^{YJ|ID}$ is IO contractible over J . \square

Lemma 5.3.7 is used to apply Theorem 5.3.6 to models where I is a sequence of unique identifiers. Only in this case, exchangeability of the unique identifiers implies the identifiers are independent of the outcomes Y .

Lemma 5.3.7. *Given any probability set \mathbb{P}_C where $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e C|(D, I)$ and $I : \Omega \rightarrow I$ is an infinite sequence of unique identifiers, if for each finite permutation $\rho : \mathbb{N} \rightarrow \mathbb{N}$*

$$\mathbb{P}_\alpha^{Y|I} = (\text{swap}_{\rho(I)} \otimes \text{Id}_X) \mathbb{P}_\alpha^{Y|I}$$

then $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e I|C$.

Proof. By definition of the set I of finite permutations, for every $\rho \in I$, $B \in \mathcal{Y}^\mathbb{N}$, $d \in D^\mathbb{N}$ there is a finite permutation $\rho^{-1} \in I$ such that $\rho \circ \rho^{-1} = \text{id}_\mathbb{N}$. Then

$$\begin{aligned}
 \mathbb{P}_\alpha^{Y|I}(B|\rho) &= (\mathbb{F}_{\rho^{-1}} \otimes \text{Id}_X) \mathbb{P}_\alpha^{Y|I}(B|\rho) \\
 &= \mathbb{P}_\alpha^{Y|I}(B|\text{id}_\mathbb{N})
 \end{aligned}$$

Therefore

$$\mathbb{P}_\alpha^{Y|I} \stackrel{\mathbb{P}_C}{\cong} \text{erase}_I \otimes \mathbb{P}_\alpha^Y$$

\square

Theorem 5.3.8 presents a set of sufficient conditions for $\mathbb{P}_C^{Y_i|HD_i}$ to be conditionally independent and identical response functions with respect to the standard directing random measure H :

1. There exist variables I representing “unique identifiers” which satisfy the assumption that $\mathbb{P}_C^{Y_i|JD_i I_i}$ are a sequence of independent and identical response functions for some J
2. The identifiers I can be swapped without altering the model of the consequences Y

3. The inputs D and the choice C are substitutable with respect to Y and I :
 $YI \perp\!\!\!\perp_{\mathbb{P}_C}^e C|D$ and $YI \perp\!\!\!\perp_{\mathbb{P}_C}^e D|C$

Theorem 5.3.8. *Given a sequential input-output model $(\mathbb{P}_C, (D, I), Y)$, on (Ω, \mathcal{F}) with Y standard measurable and C and D countable, I an infinite sequence of unique identifiers, if there is some J such that*

$$\begin{aligned} \mathbb{P}_\alpha^{Y|I} &= (\text{swap}_{\rho(I)} \otimes \text{Id}_X) \mathbb{P}_\alpha^{Y|I} & \forall \text{ finite permutations } \rho \\ YI &\perp\!\!\!\perp_{\mathbb{P}_C}^e D|C \\ YI &\perp\!\!\!\perp_{\mathbb{P}_C}^e C|D \\ \forall i, j \in \mathbb{N} : &\sum_{\alpha \in C} \mathbb{P}_\alpha^{I_i}(j) > 0 \end{aligned}$$

and for each α

$$\begin{aligned} \mathbb{P}_\alpha^{Y_i|J_i D_i} &= \mathbb{P}_\alpha^{Y_i|J_i D_i} \\ Y_i &\perp\!\!\!\perp_{\mathbb{P}_C}^e (I_{\{i\}^C}, D_{\{i\}^C})|(J, I_i, D_i) \end{aligned} \quad \forall i, j \in \mathbb{N}$$

then $\mathbb{P}_C^{Y|HD}$ is also IO contractible over the directing random measure H .

Proof. Apply lemma 5.3.7 to get $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e I|C$, then apply Theorem 5.3.6 for $\mathbb{P}_C^{Y|JD}$ IO contractible. We need to show $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e J|(H, D, C)$.

Note that for each $d \in D$, the map $\omega \mapsto \mathbb{P}_C^{Y|JD}(\cdot | J(\omega), d^{|\mathbb{N}|})$ is measurable with respect to the exchangeable σ -algebra $\mathcal{E} \subset \mathcal{F}$, and is hence H -measurable. Furthermore, for any $\mathbf{d} \in D^{\mathbb{N}}$, the $\omega \mapsto \mathbb{P}_C^{Y|JD}(\cdot | J(\omega), \mathbf{d})$ is a function of the vector valued map $(\omega \mapsto \mathbb{P}_C^{Y|JD}(J(\omega), d^{|\mathbb{N}|}))_{d \in D}$ and is therefore also H -measurable. Thus $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e J|(H, D, C)$ as desired. \square

Theorem 5.3.8 can be extended to the case where decisions D are a one-to-one deterministic function of the choice, or a random mixtures of one-to-one deterministic functions of the choice. This extension is applicable a randomised controlled trial, where the treatments are deterministically controlled and randomly assigned.

Theorem 5.3.9. *Consider a sequential input-output model $(\mathbb{P}_{C'}, D, Y)$ where $\mathbb{P}_{C'}^{Y|WD}$ is IO contractible over W , and construct a second model (\mathbb{P}_C, D, Y) where \mathbb{P}_C is the union of $\mathbb{P}_{C'}$ and its convex hull. Then $\mathbb{P}_C^{Y|WD}$ is also IO contractible.*

Proof. For all $\alpha \in C$, there is some probability measure $\mu : C' \rightarrow [0, 1]$ such that

$$\begin{aligned} \mathbb{P}_\alpha^{Y|WD} &= \sum_{\beta \in C'} \mu(\beta) \mathbb{P}_\beta^{Y|WD} \\ &= \mathbb{P}_{C'}^{Y|WD} \end{aligned}$$

thus

$$\mathbb{P}_C^{Y|WD} = \mathbb{P}_{C'}^{Y|WD}$$

and in particular, $\mathbb{P}_C^{Y|WD}$ is IO contractible. \square

Theorem 5.3.9 can be used to argue that, given a sequence of experiments IO contractible under deterministic choices, adding random mixtures of these choices also yields a IO contractible sequence. Kasy (2016) argues that as long as the experimenter controls the treatment assignment, causal effects are identified (i.e. the randomisation step is not strictly necessary). Example 5.3.10 shows that this argument might be supported, but Example ?? shows that there are subtle ways that might lead to this argument failing.

We assume an infinite sequence, which is clearly unreasonable. Extending the representation theorems to the case of finite sequences, using for example the result of Diaconis and Freedman (1980) with establishes that finite exchangeable distributions are approximately mixtures of independent and identically distributed sequences, would allow some implausible assumptions in the following example to be removed.

Theorem 5.3.8 is used in the following example to argue that, under certain conditions, a controlled experiment supports a IO contractible model.

Example 5.3.10. A sequential experiment is modeled by a probability set \mathbb{P}_C with binary treatments $D := (D_i)_{i \in \mathbb{N}}$ and binary outcomes $Y := (Y_i)_{i \in \mathbb{N}}$. The set of choices C is the set of all probability distributions $\Delta(D^N)$ for some $N \subset \mathbb{N}$ (this is to ensure C is countable).

Each treatment D_i is given to a patient, and each patient provides a unique identifier l_i which for simplicity we assume is a number in \mathbb{N} (instead of, say, a driver's license number and state of issue), and that (implausibly) there is a positive probability for l_i to take any value in \mathbb{N} for any choice α .

The treatments are decided as follows: the analyst consults the model \mathbb{P}_C , and, according to \mathbb{P}_C and some previously agreed upon decision rule, comes up with a (possibly stochastic) sequence of treatment distributions $\alpha := (\mu_i)_{i \in \mathbb{N}}$ with each μ_i in $\Delta(\{0, 1\})$. If μ_i is deterministic – that is, it puts probability 1 on some treatment d_i , the experiment administrator will assign patient i the treatment d_i . Otherwise, if μ_i is nondeterministic, the administrator will consult a random number generator that yields treatment assignments according to μ_i , and treatment will then be assigned deterministically according to the result. Letting $C' \subset C$ be the deterministic elements of C , this scheme is assumed by the analyst to support the assumptions $Y|J \perp\!\!\!\perp_{\mathbb{P}_{C'}}^e D|C$ and $Y|J \perp\!\!\!\perp_{\mathbb{P}_{C'}}^e C|D$ for any J , and the randomisation procedure is deemed sufficient to ensure that for any mixed $\alpha \in C$ where $\alpha = \sum_{\beta \in C'} \mu(\beta)\beta$, $\mathbb{P}_\alpha = \sum_{\beta \in C'} \mathbb{P}_\beta$.

Furthermore, assume $\mathbb{P}_C^{Y|Dl}$ is IO contractible. Then by Theorem 4.6.7, there is some J such that, conditional on J , $\mathbb{P}_C^{Y_i|J D_i l_i}$ are conditionally independent and identical response functions. The analyst constructing the model has no particular knowledge about any identifier, and so for any choice the associated model is assumed invariant to permutations of identifiers - that is $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e l|C$ (see Lemma ??). The assumption that this holds given any choice can be tricky – not only must the identifiers appear symmetric to the analyst constructing the model, but nothing breaking this symmetry may be learned from the choice α (see the Example 5.3.4). One reason supporting this assumption is that the decision maker selects α according to a rule known in advance, so they do not

“learn” anything upon picking a particular α .

Then, for the deterministic subset $C' \subset C$, application of Theorem 5.3.8 yields $\mathbb{P}_{C'}^{Y|HD}$ is IO contractible over H , and by application of Corollary 5.3.9, so is $\mathbb{P}_C^{Y|D}$.

Permutability of identifiers can fail when the rule for selecting α is not known in advance. The following example is extreme in order to illustrate the issue clearly. The distinction between the analyst and the administrator is also intended to make the example easier to parse. The key point is that, when the rule for selecting α is not known in advance, symmetries that are apparent at the time of model construction do not necessarily hold for every choice α , and this remains true if e.g. the selection of choices leads to less extreme confounding or the analyst and the administrator are actually the same person.

The following example involves the choice α depending on some covariate U . It is not straightforward to express the idea that “ α depends on U ” in a probability set model \mathbb{P}_C , and they are intended to apply to situations where the choice doesn’t depend on anything not already expressed in the model (as in Example 5.3.10). However, the fact that probability sets don’t work well in situations where the choice depends on something not expressed in the model doesn’t mean that you can’t use a probability set to model such a situation, it just means that you shouldn’t do it. This is what the following example shows.

The condition $YIJ \perp\!\!\!\perp_{\mathbb{P}_C}^e C|D$ without also having $YIJ \perp\!\!\!\perp_{\mathbb{P}_C}^e D|C$ does *not* imply the conclusion of Theorem 5.3.8. Informally, if D gives some “extra information” over and above C , then any symmetry that holds before we observe D might not hold after D has been observed. We have argued in Section 5.3.4 somewhat informally that the choice C should be completely under the decision maker’s control – for Theorem 5.3.8, this perfect control has to extend to the sequence of inputs D . Constructing the following example requires the hypotheses that any given identifier $i \in \mathbb{N}$ could be associated with one of two input-output maps $D \rightarrow Y$. Thus the space of hypotheses is a sequence of binary values $H = \{0, 1\}^{\mathbb{N}}$. Equipped with the product topology, H is a countable product of separable, completely metrizable spaces and is therefore also separable and completely metrizable (Willard, 1970, Thm. 16.4, Thm. 24.11). Thus $(H, \mathcal{B}(H))$ is a standard measurable space and, because it is uncountable, it is isomorphic to $([0, 1], \mathcal{B}([0, 1]))$.

Example 5.3.11. Take $Y = C = D = \{0, 1\}$ and take (H, \mathcal{H}) to be $\{0, 1\}^{\mathbb{N}}$ equipped with the product topology. For any $i \neq 1$, $Y_i I_i D_i \perp\!\!\!\perp_{\mathbb{P}_C}^e C$, while $\mathbb{P}_\alpha^{D_1} = \delta_\alpha$ and $I_i \perp\!\!\!\perp_{\mathbb{P}_C}^e C$.

$YI \perp\!\!\!\perp_{\mathbb{P}_C}^e C|D$ follows from the fact that C can be (almost surely) written as a function of D .

For all $i \in \mathbb{N}$, $y, d \in \{0, 1\}$, $h \in H$ set

$$\mathbb{P}_C^{Y_i | H I_i D_i}(y|h, j, d) = \delta_1(p(j, h))\delta_d(y) + \delta_0(p(j, h))\delta_{1-d}(y)$$

where $p(j, h)$ projects the j -th component of h . That is, if h maps j to 1, Y goes

with D while if h maps j to 0, Y goes opposite D . Suppose also

$$Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (X_{<i}, Y_{<i}, I_{<i}, C) | (X_i, Y_i, H)$$

Then $\mathbb{P}_C^{Y|D_i}$ is IO contractible. Set \mathbb{P}_C^H to be the uniform measure on (H, \mathcal{H}) and for $i > 1$

$$\mathbb{P}_C^{D_i|I_i H}(d|j, h) = \delta_{p(j, h)}(d)$$

that is, if h maps j to 1, D is 1 while if h maps j to 0, D is 0. This also implies

$$\mathbb{P}_C^{I_i|D_i H}(p(\cdot, h)^{-1}(d)|d, h) = 1 \quad (5.7)$$

Then, for $i > 1$

$$\begin{aligned} \mathbb{P}_\alpha^{Y_i|HD_i}(y|h, d) &= \sum_{j \in \mathbb{N}} \delta_1(p(j, h)) \delta_d(y) \mathbb{P}_C^{I_i|D_i H}(j|d, h) + \delta_0(p(j, h)) \delta_{1-d}(y) \mathbb{P}_C^{I_i|D_i H}(j|d, h) \\ &= \sum_{j \in \mathbb{N}} \delta_1(d) \delta_d(y) \mathbb{P}_C^{I_i|D_i H}(j|d, h) + \delta_0(d) \delta_{1-d}(y) \mathbb{P}_C^{I_i|D_i H}(j|d, h) \quad \text{by Eq (5.7)} \\ &= \delta_1(y) \\ \implies \mathbb{P}_\alpha^{Y_i|D_i}(y|d) &= \delta_1(y) \end{aligned}$$

For $q \in I$, set

$$\mathbb{P}_C^{I|H}(q|h) = \begin{cases} 0.5 & q = (1, 2, 3, 4, \dots) \text{ or } (1, 3, 2, 4, \dots) \\ 0 & \text{otherwise} \end{cases}$$

and set

$$\mathbb{P}_C^{H|D}(h) = \begin{cases} 0.5 & h = (0, 1, 0, 1, 1, \dots) \text{ or } h = (0, 0, 1, 1, 1, \dots) \\ 0 & \text{otherwise} \end{cases}$$

Let \overline{H} be the support of $\mathbb{P}_C^{H|D}(h)$.

Then for $i = 1$

$$\begin{aligned} \mathbb{P}_\alpha^{Y_1|D_1}(y|h, d) &= \sum_{h \in H} \sum_{j \in \mathbb{N}} \mathbb{P}_\alpha^{I_1|D_1 H}(j|d, h) \mathbb{P}_C^{H|D_1}(h|d) (\delta_1(p(j, h)) \delta_d(y) + \delta_0(p(j, h)) \delta_{1-d}(y)) \\ &= \sum_{h \in \overline{H}} 0.5 (\delta_1(p(1, h)) \delta_d(y) + \delta_0(p(1, h)) \delta_{1-d}(y)) \\ &= \delta_{1-d}(y) \\ &\neq \mathbb{P}_\alpha^{Y_i|D_i}(y|h, d) \quad i \neq 1 \end{aligned}$$

Thus $\mathbb{P}_C^{Y|D}$ is not IO contractible by Theorem 4.3.9.

However, given any finite permutation $\rho : \mathbb{N} \rightarrow \mathbb{N}$

$$\begin{aligned} \mathbb{P}_\alpha^{Y|I}(y|q) &= \sum_{h \in \overline{H}} \sum_{d \in \{0,1\}^{\mathbb{N}}} \prod_{i \in \mathbb{N}} \mathbb{P}_C^{Y_i|I_i D_i H}(y_i|q_i, d_i, h) \mathbb{P}_\alpha^{D_i|I_i H}(d_i|q_i, h) \mathbb{P}_C^H(h) \\ &= \delta_{1-\alpha}(y_1) \delta_{(1)_{i \in \mathbb{N}}}(y_{>1}) \\ &= \mathbb{P}_\alpha^{Y|I}(y|\rho^{-1}(q)) \\ &= \mathbb{F}_\rho \mathbb{P}_\alpha^{Y|I}(y|q) \end{aligned}$$

5.3.4 Other examples

Example 5: Backdoor adjustment The “backdoor adjustment” formula is a fundamental tool for many kinds of causal inference. This is a short example to show the conditions under which it’s applicable, stated in terms of IO contractibility. Suppose a sequential input-output model $(\mathbb{P}_C, (D, X), Y)$ where $(\mathbb{P}_C^{Y|WDX})$ is IO contractible, and:

- $i > n \implies X_i \perp\!\!\!\perp_{\mathbb{P}_C}^e D_i | (H, C)$
- $\mathbb{P}_\alpha^{X_i|H} \cong \mathbb{P}_\alpha^{X_1|H}$ for all α

Then the model exhibits a kind of “backdoor adjustment”. Specifically, for $i > n$

$$\begin{aligned} \mathbb{P}_\alpha^{Y_i|D_i H}(A|d, h) &= \int_X \mathbb{P}_\alpha^{Y_i|X_i D_i H}(A|d, x, h) \mathbb{P}_\alpha^{X_i|D_i H}(dx|d, h) \\ &= \int_X \mathbb{P}_\alpha^{Y_1|X_1 D_1 H}(A|d, x, h) \mathbb{P}_\alpha^{X_i|H}(dx|h) \\ &= \int_X \mathbb{P}_\alpha^{Y_1|X_1 D_1 H}(A|d, x, h) \mathbb{P}_\alpha^{X_1|H}(dx|h) \end{aligned} \quad (5.8)$$

Equation (5.8) is identical to the backdoor adjustment formula (Pearl, 2009, Chap. 1) for an intervention on D_1 targeting Y_1 where X_1 is a common cause of both.

Example 6: the provenance of the choice variable

Use individual-level ccont

The point of this example is to clarify the idea of a “choice” variable. If we say that some value is the outcome of a choice, a straightforward interpretation of this term suggests that this value was chosen by someone, somewhere. However, for the purposes of decision making models, there are important differences between:

- Values chosen by someone, somewhere
- Values chosen by the decision make, using the decision making model

We call this example the “I choose vs you choose” problem. Suppose we have a decision maker (“DM”) and an administrator (“admin”) cooperating to collect data to support DM to make a choice.

First, consider the “I choose” condition. Here, DM’s choice $\alpha \in \{0, 1\}^2$ deterministically sets the value of binary inputs D_1, D_2 , and the decision maker is interested in evaluating the corresponding binary outputs Y_1, Y_2 . The decision maker assesses that their knowledge of the real-world mechanisms that gives rise to each output Y_i in the context on an input D_i render these mechanisms indistinguishable. From their point of view, the input-output relations for each step are indistinguishable. In particular, they assess that the marginal probabilities of the outputs are the same given a corresponding input, and that the evidence that the first experiment brings to bear on the second is equivalent to the evidence that the second brings to bear on the first. Thus, they assess that exchange commutativity is appropriate; for all α :

$$\mathbb{P}_\alpha^{Y_1 Y_2 | D_1 D_2} = \mathbb{P}_\alpha^{Y_2 Y_1 | D_2 D_1}$$

but this example suggests another reason one might want to avoid deterministic treatment assignments. If the choices α are a deterministic sequence of assignments for each index i , this means that there is an enormous set of possible choices, and many degrees of freedom if the choices “aren’t actually chosen” in the sense of the example above. In contrast, if the set of choices is a single parameter in $[0, 1]$ which is then used to assign all treatments according to a random procedure depending only on this parameter, there are many fewer degrees of freedom to exploit if the choice “isn’t actually chosen”.

A particular concern arises when the choice variable C is not associated with the output of a decision procedure involving the model \mathbb{P}_C . In this situation, the value of C can affect the model in potentially unexpected ways. “Potentially unexpected” is a vague notion, and we can’t say whether C being completely under the decision maker’s control avoids “unexpected” dependence on C , but it seems to be less problematic.

We set this up in terms of an “analyst” and an “administrator” who have responsibility for different parts of the procedure. They don’t strictly need to be different people, but it helps make the issue clearer. The analyst’s job is to construct a model \mathbb{P}_C , evaluate different options $\alpha \in C$ and offer advice regarding the choice. The administrator’s job is to choose some $\alpha \in C$ satisfying the analyst’s requirements and to carry out any procedure arising from this.

This separation of concerns gives the administrator a degree of freedom in their choice, and they can potentially use this to choose α with access to information that the analyst lacks.

In particular, suppose an experiment is modeled by a sequential input-output model $(\mathbb{P}_C, (D, U), Y)$ and the set of choices $C = [0, 1]^\mathbb{N}$ is a length \mathbb{N} sequence of probability distributions in $\Delta(\{0, 1\})$. The analyst, based on their knowledge of the experiment, constructs \mathbb{P}_C such that $\mathbb{P}_C^{Y_i | U_i D_i}(1 | \cdot, \cdot)$ is given by:

	$D_i = 0$	$D_i = 1$
$U_i = 0$	0	0
$U_i = 1$	1	1

and the triples (D_i, U_i, Y_i) are mutually independent given C . This makes $\mathbb{P}_C^{Y|UD}$ IO contractible over $*$. Suppose also

$$\mathbb{P}_\alpha^{D_i}(1) = \alpha_i$$

where $\alpha = (\alpha_i)_{i \in \mathbb{N}}$. From the analyst's point of view, both before and after making their recommendations the U_i are also IID. This will be expressed with a probability distribution \mathbb{Q} representing the analyst's prior knowledge:

$$\mathbb{Q}^{U_i}(1) = 0.5$$

one might be tempted to reason that, if \mathbb{Q} is the analyst's state of knowledge after making any recommendation, then we should take $\mathbb{P}_C^U = \mathbb{Q}^U$. Call the resulting model \mathbb{P}'_C . Together with the other assumptions above, this would imply

$$\mathbb{P}_C^{Y_i|D_i}(1|d) = 0.5 \quad \forall d \in \{0, 1\}$$

Thus $\mathbb{P}_C^{Y|D}$ is also IO contractible.

However, the analyst's recommendation *does not* fix the value of C . Suppose analyst actually recommends any α such that $\lim_{n \rightarrow \infty} \sum_i^n \frac{\alpha_i}{n} = 0.5$ (acknowledging that, in this contrived example, there's no obvious reason to do so). Suppose that the administrator operates by the following rule: *first* they observe the value of U_i , then they choose α_i equal to whatever they saw with an ϵ sized step towards 0.5. That is, if they see $U_i \bowtie 1$, they choose $\alpha_i = 1 - \epsilon$, where $\epsilon < 0.5$.

Then the analyst should instead adopt the model

$$\mathbb{P}_\alpha^{U_i}(1) = \mathbf{1}_{\alpha_i > 0.5}$$

Take α such that $\alpha_i = 1 - \epsilon$ and $\alpha_j = \epsilon$. Then

$$\begin{aligned} \mathbb{P}_\alpha^{Y_i|D_i}(1|1) &= 1 \\ &\neq \mathbb{P}_\alpha^{Y_j|D_j}(1|1) \\ &= 0 \end{aligned}$$

everything has been assumed IID, so $\mathbb{P}_C^{Y|HD}$ is not IO contractible.

The original justification for having a set of choices C is that C is the set of things that, after deliberation aided by the model \mathbb{P}_C , the decision maker might select. The present example does not conform to this understanding of the meaning of the set C , and it suggests that one should be cautious when modelling “decision problems” with “choices” that are not actually the things that are being chosen.

This point is related to the question of why experimenters randomise. Kasy (2016) argues that “randomised controlled trials are not needed for causal identifiability, only controlled trials”, and suggests that experiments should sometimes be designed with deterministic assignments of patients to treatment and control groups, optimised according to the experiment designer’s criteria. Following this, Banerjee et al. (2020) suggested that deterministic rules might falter when one can’t pick a function to balance covariates in a way that satisfies everyone in a panel of reviewers.

Without solving the problem, we observe that the terms “control” and “choice” here subsume both different kinds of choice indicated above, each of which has different implications for the construction of decision making models. We offer a speculative alternative explanation for randomisation: perhaps that the same model may be appropriate for both notions of “choice” under randomised choices, but not under nonrandomised choices.

Sävje (2021) argues that random assignment (under his definition) does not imply unconfoundedness

5.4 Conclusion

We review the decision making models implied by causal Bayesian networks and potential outcomes models. We find that these kinds of models have complementary “missing pieces” needed to induce the relevant decision making model – while causal Bayesian networks already have interventions that provide a kind of “choice set”, they need to be unrolled to a sequential model. On the other hand potential outcomes models can already be specified in an unrolled form, but need some notion of “choice set” to induce a decision making model. Common formulations of both models feature conditionally independent and identical response functions. We explore individual-level response functions as a means of establishing the widely accepted result that randomised trials. We note that the assumption of individual-level response functions seems to be a missing step in the often cited idea that exchangeability of individuals implies exchangeability of potential outcomes, and we show that with individual-level response functions, exchangeable individuals and completely controllable inputs, causal relationships are identified. The need for completely controllable inputs is also widely accepted, but to our knowledge it only appears as a formal assumption in Theorem 4.3.23.

Chapter 6

Discussion

Decision making models differ from regular probabilistic models in that they have a domain of choices or options that must be compared to one another.

In parallel, we’ve made use of a string diagram notation, both for writing out some proofs and for a visual aid to understanding different kinds of models. The string diagram aspect of this thesis can in principle be cleanly separated from the rest – it is simply a notation for reasoning using probability theory. Compared to the more common diagrammatic language of directed acyclic graphs (DAGs), the chief advantage of the string diagram notation is that it explicitly represents Markov kernels in the diagrams, and so it is possible (for example) to write that one diagram is equal to another different diagram. This facilitates mathematical reasoning using diagrams alone. A key missing ingredient from the use of string diagrams in this thesis is an analogue of the *d-separation* condition for traditional DAGs. Conditional independence is a very important feature in causal reasoning, and it would be useful to be able to map a collection of extended conditional independence statements to a diagrammatic representation. An analogue of d-separation could also be relevant to the abstract diagrammatic notion of conditional independence postulated by Fritz (2020), which would allow for generalisation of conditional independence beyond the concrete category of Markov kernels.

We focus on decision problems here, but the fact that we call the domain set C a set of “choices” is only a convention. There is no obvious reason it could not, for example, be a set of counterfactual propositions. The string diagram notation, extended conditional independence and the theorems proved here would be just as true for models under this alternative interpretation, but they may not always be as relevant. For example, we might not expect counterfactual responses to be independent of responses in the real world.

Conditionally independent and identical response functions (CIIRs) are a fundamental notion to our work. The intuitive idea is: if I am trying to discover an unknown function, I need to feed it different inputs and examine its outputs. The assumption of CIIRs allows me to reason as if a series of trials is, precisely, a series of trials of different inputs given to the *same* stochastic function.

Recall the symmetry identified in Chapter 4: the assumption of conditionally independent and identical response functions implies that infinite sequences of input-output pairs with sufficient support are interchangeable. This applies, for example, to sequences comprised of both experimental and passive observational data. To assume that the response of some variable to a treatment and covariates is identical for the experimental and the observational data is to hold that infinite sequences of each type of data are interchangeable. It is already widely accepted that this assumption is usually inappropriate for observational data, but it is often made nonetheless. We offer an alternative interpretation of this assumption when applied to observational data: it is often equivalent to assuming that the observational data is, in the limit, just as good as experimental data for the purposes of predicting consequences of actions.

There is clearly a need for weaker assumptions than conditionally independent and identical response functions. In Chapter 5, we introduce the idea of *preemption*. As an informal principle, the idea that whatever we can do has been done before is easy to understand, and seems like it might be applicable in many cases. We offer a tentative formalisation of this idea, but the question of how to precisely specify that something “has been done before” remains open. One interesting implication of our version of preemption is that, in conjunction with an assumption of a generic relationship between a “choice-making” function and the “consequences of that choice”, we find that it’s possible to reason from conditional independences in the data to the conclusion of identical response functions (Theorem 5.1.18). The notion of generic relationships between the functions that make up a causal graph has been used to motivate the assumption of *causal faithfulness*, as well as to provide a basis for the inference of the direction of causal relationships. Th

One option identified in Duvenaud et al. (2008) is to assume that inputs have no effect at all. In fact, in the guise of the *null hypothesis* this is the “default assumption” employed in almost every experimental trial in existence (it was first given this name by Fisher (1971)). By convention at least, this is a major alternative to the assumption of identical responses, particularly in cases where the data are not considered to be decisive one way or the other.

References

- Holy Bible : Contemporary English Version*. New York : American Bible Society, [1995] 1995, 1995. URL <https://search.library.wisc.edu/catalog/999953290302121>.
- David J. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, December 1981. ISSN 0047-259X. doi: 10.1016/0047-259X(81)90099-3. URL <https://www.sciencedirect.com/science/article/pii/0047259X81900993>.
- Joshua D. Angrist and Jörn-Steffen Pischke. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press, Princeton ; Oxford, with french flaps edition edition, December 2014. ISBN 978-0-691-15284-4.

- A. V. Banerjee, S. Chassang, and E. Snowberg. Chapter 4 - Decision Theoretic Approaches to Experiment Design and External Validity. We thank Esther Duflo for her leadership on the handbook and for extensive comments on earlier drafts. Chassang and Snowberg gratefully acknowledge the support of NSF grant SES-1156154. In Abhijit Vinayak Banerjee and Esther Duflo, editors, *Handbook of Economic Field Experiments*, volume 1 of *Handbook of Field Experiments*, pages 141–174. North-Holland, January 2017. doi: 10.1016/bs.hefe.2016.08.005. URL <https://www.sciencedirect.com/science/article/pii/S2214658X16300071>.
- Abhijit V. Banerjee, Sylvain Chassang, Sergio Montero, and Erik Snowberg. A Theory of Experimenters: Robustness, Randomization, and Balance. *American Economic Review*, 110(4):1206–1230, April 2020. ISSN 0002-8282. doi: 10.1257/aer.20171634. URL <https://www.aeaweb.org/articles?id=10.1257/aer.20171634>.
- Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On Pearl’s hierarchy and the foundations of causal inference. Technical report, 2020. URL <https://causalai.net/r60.pdf>.
- Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction, 1998. URL <http://jmvidal.cse.sc.edu/lib/sutton98a.html>.
- Richard Berk. What You Can and Cant Properly Do with Regression. *Journal of Quantitative Criminology*, 26(4):481–487, December 2010. ISSN 1573-7799. doi: 10.1007/s10940-010-9116-4. URL <https://doi.org/10.1007/s10940-010-9116-4>.
- David A. Blackwell. *Theory of Games and Statistical Decisions*. Dover Publications, New York, September 1979. ISBN 978-0-486-63831-7.
- Vladimir Bogachev and Ilya Malofeev. Kantorovich problems and conditional measures depending on a parameter. *Journal of Mathematical Analysis and Applications*, 486:123883, June 2020. doi: 10.1016/j.jmaa.2020.123883.
- Ethan D. Bolker. Functions Resembling Quotients of Measures. *Transactions of the American Mathematical Society*, 124(2):292–312, 1966. ISSN 0002-9947. doi: 10.2307/1994401. URL <https://www.jstor.org/stable/1994401>. Publisher: American Mathematical Society.
- Ethan D. Bolker. A Simultaneous Axiomatization of Utility and Subjective Probability. *Philosophy of Science*, 34(4):333–340, 1967. ISSN 0031-8248. URL <https://www.jstor.org/stable/186122>. Publisher: [The University of Chicago Press, Philosophy of Science Association].
- 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.

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

David Maxwell Chickering. Learning Equivalence Classes of Bayesian-Network Structures. *Journal of Machine Learning Research*, 2(Feb):445–498, 2002. ISSN ISSN 1533-7928. URL <http://www.jmlr.org/papers/v2/chickering02a.html>.

David Maxwell Chickering. Optimal Structure Identification with Greedy Search. *J. Mach. Learn. Res.*, 3:507–554, March 2003. ISSN 1532-4435. doi: 10.1162/153244303321897717. URL <https://doi.org/10.1162/153244303321897717>.

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

Kenta Cho and Bart Jacobs. Disintegration and Bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, 29(7):938–971, August 2019. ISSN 0960-1295, 1469-8072. doi: 10.1017/S0960129518000488.

Panayiota Constantinou and A. Philip Dawid. Extended conditional independence and applications in causal inference. *The Annals of Statistics*, 45(6): 2618–2653, 2017. ISSN 0090-5364. URL <http://www.jstor.org/stable/26362953>. Publisher: Institute of Mathematical Statistics.

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>.

A. P. Dawid. Influence Diagrams for Causal Modelling and Inference. *International Statistical Review*, 70(2):161–189, 2002. ISSN 1751-5823. doi: 10.1111/j.1751-5823.2002.tb00354.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2002.tb00354.x>.

A. Philip Dawid. Decision-theoretic foundations for statistical causality. *arXiv:2004.12493 [math, stat]*, April 2020. URL <http://arxiv.org/abs/2004.12493>. arXiv: 2004.12493.

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>.

Bruno de Finetti. Foresight: Its Logical Laws, Its Subjective Sources. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics: Foundations and Basic Theory*, Springer Series in Statistics, pages

- 134–174. Springer, New York, NY, [1937] 1992. ISBN 978-1-4612-0919-5. doi: 10.1007/978-1-4612-0919-5_10. URL https://doi.org/10.1007/978-1-4612-0919-5_10.
- P. Diaconis. Recent progress on de Finetti's notions of exchangeability. *Bayesian Statistics*, 3:111–125, 1988.
- P. Diaconis and D. Freedman. Finite Exchangeable Sequences. *The Annals of Probability*, 8(4):745–764, August 1980. ISSN 0091-1798, 2168-894X. doi: 10.1214/aop/1176994663. URL <https://projecteuclid.org/journals/annals-of-probability/volume-8/issue-4/Finite-Exchangeable-Sequences/10.1214/aop/1176994663.full>. Publisher: Institute of Mathematical Statistics.
- David Duvenaud, Daniel Eaton, Kevin Murphy, and Mark Schmidt. Causal learning without DAGs. In *Proceedings of the 2008th International Conference on Causality: Objectives and Assessment - Volume 6*, COA'08, pages 177–190, Whistler, Canada, December 2008. JMLR.org.
- Frederick Eberhardt. Almost optimal intervention sets for causal discovery. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, pages 161–168, Arlington, Virginia, USA, July 2008. AUAI Press. ISBN 978-0-9749039-4-1.
- Dean Eckles and Eytan Bakshy. Bias and High-Dimensional Adjustment in Observational Studies of Peer Effects. *Journal of the American Statistical Association*, 116(534):507–517, April 2021. ISSN 0162-1459. doi: 10.1080/01621459.2020.1796393. URL <https://doi.org/10.1080/01621459.2020.1796393>. Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/01621459.2020.1796393>.
- Thomas S. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, July 1967. ISBN 978-1-4832-2123-6.
- R.P. Feynman. *The Feynman lectures on physics*. Le cours de physique de Feynman. Interditions, France, 1979. ISBN 978-2-7296-0030-3. INIS Reference Number: 13648743.
- Ronald A. Fisher. Cancer and Smoking. *Nature*, 182(4635):596–596, August 1958. ISSN 1476-4687. doi: 10.1038/182596a0. URL <https://www.nature.com/articles/182596a0>. Number: 4635 Publisher: Nature Publishing Group.
- Ronald A. Fisher. *The Design of Experiments*. Macmillan Pub Co, New York, 9th edition edition, June 1971. ISBN 978-0-02-844690-5.
- 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.

- Tobias Fritz. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, August 2020. ISSN 0001-8708. doi: 10.1016/j.aim.2020.107239. URL <https://www.sciencedirect.com/science/article/pii/S0001870820302656>.
- Brett R. Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook. SSRN Scholarly Paper ID 3033144, Social Science Research Network, Rochester, NY, September 2018. URL <https://papers.ssrn.com/abstract=3033144>.
- Brett R. Gordon, Robert Moakler, and Florian Zettelmeyer. Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement. *arXiv:2201.07055 [econ]*, January 2022. URL <http://arxiv.org/abs/2201.07055>. arXiv: 2201.07055.
- Sander Greenland and James M Robins. Identifiability, Exchangeability, and Epidemiological Confounding. *International Journal of Epidemiology*, 15(3): 413–419, September 1986. ISSN 0300-5771. doi: 10.1093/ije/15.3.413. URL <https://doi.org/10.1093/ije/15.3.413>.
- Jinyong Hahn, Petra Todd, and Wilbert Van der Klaauw. Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica*, 69(1):201–209, 2001. ISSN 0012-9682. URL <https://www.jstor.org/stable/2692190>. Publisher: [Wiley, Econometric Society].
- J. Y. Halpern. A Counter Example to Theorems of Cox and Fine. *Journal of Artificial Intelligence Research*, 10:67–85, February 1999. ISSN 1076-9757. doi: 10.1613/jair.536. URL <https://www.jair.org/index.php/jair/article/view/10223>.
- D. Heckerman and R. Shachter. Decision-Theoretic Foundations for Causal Reasoning. *Journal of Artificial Intelligence Research*, 3:405–430, December 1995. ISSN 1076-9757. doi: 10.1613/jair.202. URL <https://www.jair.org/index.php/jair/article/view/10151>.
- Miguel A Hernán. Beyond exchangeability: The other conditions for causal inference in medical research. *Statistical Methods in Medical Research*, 21(1): 3–5, February 2012. ISSN 0962-2802. doi: 10.1177/0962280211398037. URL <https://doi.org/10.1177/0962280211398037>. Publisher: SAGE Publications Ltd STM.
- Miguel A Hernán and James M Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health*, 60(7): 578–586, July 2006. ISSN 0143-005X. doi: 10.1136/jech.2004.029496. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2652882/>.

- Miguel A. Hernán and James M Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020. URL <https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/>.
- Eric Horvitz, David Heckerman, and Curtis Langlotz. A Framework for Comparing Alternative Formalisms for Plausible Reasoning. January 1986. URL <https://openreview.net/forum?id=rJNeX0gdbR>.
- Guido W. Imbens and Joshua D. Angrist. Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467–475, 1994. ISSN 0012-9682. doi: 10.2307/2951620. URL <https://www.jstor.org/stable/2951620>.
- Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge, 2015. ISBN 978-0-521-88588-1. doi: 10.1017/CBO9781139025751. URL <https://www.cambridge.org/core/books/causal-inference-for-statistics-social-and-biomedical-sciences/71126BE90C58F1A431FE9B2DD07938AB>.
- 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.
- Richard C. Jeffrey. *The Logic of Decision*. University of Chicago Press, July 1965. ISBN 978-0-226-39582-1.
- Olav Kallenberg. The Basic Symmetries. In *Probabilistic Symmetries and Invariance Principles*, Probability and Its Applications, pages 24–68. Springer, New York, NY, 2005a. ISBN 978-0-387-28861-1. doi: 10.1007/0-387-28861-9_2. URL https://doi.org/10.1007/0-387-28861-9_2.
- Olav Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Probability and Its Applications. Springer-Verlag, New York, 2005b. ISBN 978-0-387-25115-8. doi: 10.1007/0-387-28861-9. URL <http://link.springer.com/10.1007/0-387-28861-9>.
- Maximilian Kasy. Why Experimenters Might Not Always Want to Randomize, and What They Could Do Instead. *Political Analysis*, 24(3):324–338, 2016. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mpw012. URL <https://www.cambridge.org/core/journals/political-analysis/article/abs/why-experimenters-might-not-always-want-to-randomize-and-what-they-could-do-instead/79A12E7AB0921C50ECC1F1EA61EC1F8D>. Publisher: Cambridge University Press.
- G. Jay. Kerns and Gábor J. Székely. Definettis Theorem for Abstract Finite Exchangeable Sequences. *Journal of Theoretical Probability*, 19(3):589–608,

- December 2006. ISSN 1572-9230. doi: 10.1007/s10959-006-0028-z. URL <https://doi.org/10.1007/s10959-006-0028-z>.
- Ron Kohavi and Stefan Thomke. The Surprising Power of Online Experiments. *Harvard Business Review*, September 2017. ISSN 0017-8012. URL <https://hbr.org/2017/09/the-surprising-power-of-online-experiments>. Section: Experimentation.
- Chayakrit Krittanawong, Bharat Narasimhan, Zhen Wang, Joshua Hahn, Hafeez Ul Hassan Virk, Ann M. Farrell, HongJu Zhang, and WH Wilson Tang. Association between chocolate consumption and risk of coronary artery disease: a systematic review and meta-analysis. *European Journal of Preventive Cardiology*, July 2020. doi: 10.1177/2047487320936787. URL <http://journals.sagepub.com/doi/10.1177/2047487320936787>. Publisher: SAGE PublicationsSage UK: London, England.
- Finnian Lattimore and David Rohde. Causal inference with Bayes rule. *arXiv:1910.01510 [cs, stat]*, October 2019a. URL <http://arxiv.org/abs/1910.01510>. arXiv: 1910.01510.
- Finnian Lattimore and David Rohde. Replacing the do-calculus with Bayes rule. *arXiv:1906.07125 [cs, stat]*, December 2019b. URL <http://arxiv.org/abs/1906.07125>. arXiv: 1906.07125.
- Finnian Rachel Lattimore. Learning how to act: making good decisions with machine learning. 2017. doi: 10.25911/5d67b766194ec. URL <https://openresearch-repository.anu.edu.au/handle/1885/144602>. Accepted: 2018-06-27T06:17:13Z Last Modified: 2020-05-19 Publisher: The Australian National University.
- 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>.
- Alessandro Liberati, Douglas G. Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C. Gøtzsche, John P. A. Ioannidis, Mike Clarke, P. J. Devereaux, Jos Kleijnen, and David Moher. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLoS Medicine*, 6(7): e1000100, July 2009. ISSN 1549-1676. doi: 10.1371/journal.pmed.1000100. URL <https://dx.plos.org/10.1371/journal.pmed.1000100>.
- D. V. Lindley and Melvin R. Novick. The Role of Exchangeability in Inference. *The Annals of Statistics*, 9(1):45–58, 1981. ISSN 0090-5364. URL <https://www.jstor.org/stable/2240868>. Publisher: Institute of Mathematical Statistics.
- Christopher Meek. Strong Completeness and Faithfulness in Bayesian Networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI’95, pages 411–418, San Francisco, CA, USA, 1995.

- Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-385-1. URL <http://dl.acm.org/citation.cfm?id=2074158.2074205>. event-place: Montréal, Qué, Canada.
- Karl Menger. Random Variables from the Point of View of a General Theory of Variables. In Karl Menger, Bert Schweizer, Abe Sklar, Karl Sigmund, Peter Gruber, Edmund Hlawka, Ludwig Reich, and Leopold Schmetterer, editors, *Selecta Mathematica: Volume 2*, pages 367–381. Springer, Vienna, 2003. ISBN 978-3-7091-6045-9. doi: 10.1007/978-3-7091-6045-9_31. URL https://doi.org/10.1007/978-3-7091-6045-9_31.
- Seán M. Muller. Causal Interaction and External Validity: Obstacles to the Policy Relevance of Randomized Evaluations. *The World Bank Economic Review*, 29(suppl_1):S217–S225, January 2015. ISSN 0258-6770. doi: 10.1093/wber/lhv027. URL <https://doi.org/10.1093/wber/lhv027>.
- Ignavier Ng, Shengyu Zhu, Zhitang Chen, and Zhuangyan Fang. A Graph Autoencoder Approach to Causal Structure Learning, November 2019. URL <http://arxiv.org/abs/1911.07420>. Number: arXiv:1911.07420 arXiv:1911.07420 [cs, stat].
- Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, March 2018. doi: 10.1073/pnas.1708274114. URL <https://www.pnas.org/doi/abs/10.1073/pnas.1708274114>. Publisher: Proceedings of the National Academy of Sciences.
- Katsuhiko Ogata. *Discrete-Time Control Systems*. Pearson, Englewood Cliffs, N.J, 2 edition edition, January 1995. ISBN 978-0-13-034281-2.
- Masashi Okamoto. Distinctness of the Eigenvalues of a Quadratic form in a Multivariate Sample. *The Annals of Statistics*, 1(4):763–765, 1973. ISSN 0090-5364. URL <https://www.jstor.org/stable/2958321>. Publisher: Institute of Mathematical Statistics.
- Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, August 2015. doi: 10.1126/science.aac4716. URL <https://www.science.org/doi/abs/10.1126/science.aac4716>. Publisher: American Association for the Advancement of Science.
- Naomi Oreskes and Erik M. Conway. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Climate Change: How a Handful of Scientists ... Issues from Tobacco Smoke to Global Warming*. Bloomsbury Press, New York, NY, June 2011. ISBN 978-1-60819-394-3.
- 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>.
- Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, 1 edition edition, May 2018. ISBN 978-0-465-09760-9.
- Robert N. Proctor. The history of the discovery of the cigarettelung cancer link: evidentiary traditions, corporate denial, global toll. *Tobacco Control*, 21(2):87–91, March 2012. ISSN 0964-4563, 1468-3318. doi: 10.1136/tobaccocontrol-2011-050338. URL <https://tobaccocontrol.bmj.com/content/21/2/87>. Publisher: BMJ Publishing Group Ltd Section: The shameful past.
- 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.
- Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov Properties for Acyclic Directed Mixed Graphs. *arXiv:1701.06686 [stat]*, January 2017. URL <http://arxiv.org/abs/1701.06686>. arXiv: 1701.06686.
- 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>.
- Olli Saarela, David A. Stephens, and Erica E. M. Moodie. The role of exchangeability in causal inference. June 2020. doi: 10.48550/arXiv.2006.01799. URL <https://arxiv.org/abs/2006.01799v3>.
- L. J. Savage. The theory of statistical decision. *Journal of the American Statistical Association*, 46:55–67, 1951. ISSN 1537-274X(Electronic),0162-1459(Print). doi: 10.2307/2280094.
- Leonard J. Savage. *The Foundations of Statistics*. Wiley Publications in Statistics, 1954.
- Bernhard Schölkopf. Causality for Machine Learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 765–804. February 2022. ISBN 978-1-4503-9586-1. doi: 10.1145/3501714.3501755.
- P. Selinger. A Survey of Graphical Languages for Monoidal Categories. In Bob Coecke, editor, *New Structures for Physics*, Lecture Notes in Physics,

- pages 289–355. Springer, Berlin, Heidelberg, 2011. ISBN 978-3-642-12821-9. doi: 10.1007/978-3-642-12821-9_4. URL https://doi.org/10.1007/978-3-642-12821-9_4.
- Eyal Shohar. 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.
- Ilya Shpitser and Judea Pearl. Complete Identification Methods for the Causal Hierarchy. *Journal of Machine Learning Research*, 9(Sep):1941–1979, 2008. ISSN 1533-7928. URL <https://www.jmlr.org/papers/v9/shpitser08a.html>.
- Brian Skyrms. Causal Decision Theory. *The Journal of Philosophy*, 79(11):695–711, November 1982. doi: 10.2307/2026547. URL https://www.pdcnet.org/pdc/bvdb.nsf/purchase?openform&fp=jphil&id=jphil_1982_0079_0011_0695_0711.
- Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.
- Statista. Cigarettes - worldwide | Statista Market Forecast, 2020. URL <https://www.statista.com/outlook/50010000/100/cigarettes/worldwide>.
- Katie Steele and H. Orri Stefánsson. Decision Theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2020 edition, 2020. URL <https://plato.stanford.edu/archives/win2020/entries/decision-theory/>.
- Wolfgang Stroebe. What Can We Learn from Many Labs Replications? *Basic and Applied Social Psychology*, 41(2):91–103, March 2019. ISSN 0197-3533. doi: 10.1080/01973533.2019.1577736. URL <https://doi.org/10.1080/01973533.2019.1577736>. Publisher: Routledge _eprint: <https://doi.org/10.1080/01973533.2019.1577736>.
- Fredrik Sävje. Randomization does not imply unconfoundedness, July 2021. URL <http://arxiv.org/abs/2107.14197>. arXiv:2107.14197 [stat].
- J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Theory of games and economic behavior. Princeton University Press, Princeton, NJ, US, 1944.
- N. N. Vorobev. Consistent Families of Measures and Their Extensions. *Theory of Probability & Its Applications*, 7(2), 1962. doi: 10.1137/1107014. URL http://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=tv&paperid=4710&option_lang=eng.
- Abraham Wald. *Statistical decision functions*. Statistical decision functions. Wiley, Oxford, England, 1950.

Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, 1991.

Robert Wiblin. Why smoking in the developing world is an enormous problem and how you can help save lives, 2016. URL <https://80000hours.org/problem-profiles/tobacco/>.

Stephen Willard. *General topology*. Reading, Mass., Addison-Wesley Pub. Co, 1970. ISBN 978-0-201-08707-9. URL http://archive.org/details/generaltopology00will_0.

World Health Organisation. Tobacco Fact sheet no 339, 2018. URL <https://www.webcitation.org/6gUXrCDKA>.

Sewall Wright. The Method of Path Coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215, 1934. ISSN 0003-4851. URL <https://www.jstor.org/stable/2957502>. Publisher: Institute of Mathematical Statistics.

Lily Zhang. The Abdul Latif Jameel poverty action lab: bringing evidence-based policy into international development. *Harvard International Review*, 35(4):4–6, March 2014. ISSN 07391854. URL <https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=07391854&v=2.1&it=r&id=GALE%7CA370890303&sid=googleScholar&linkaccess=abs>. Publisher: Harvard International Relations Council, Inc.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL <https://proceedings.neurips.cc/paper/2018/hash/e347c51419ffb23ca3fd5050202f9c3d-Abstract.html>.

Appendix: