=9 =1

Decision theoretic foundations for statistical causal modelling

David Johnston

A thesis submitted for the degree of Doctor of Philosophy

College of Engineering and Computer Science

Australian National University

September, 2022

Declaration

This thesis is an account of research undertaken between January 2018 and September 2022 at the College of Engineering and Computer Sciences, The Australian National University, Canberra, Australia.

The work presented in this thesis is that of the candidate alone, except where indicated by due literature reference and acknowledgements in the text. It has not been submitted in whole or in part for any other degree at this or any other university.

The development of ideas and research was undertaken with guidance from my primary supervisors Robert Williamson and Cheng Soon Ong, and the thesis was written by myself. The overall direction of the research was developed in collaboration with my supervisors, who also provided a lot of detailed feedback on my work along the way. The original results presented here were primarily my work.

David Johnston
5 July 2023

Acknowledgements

The research in this thesis would not have been possible without my primary supervisors Robert Williamson, Cheng Soon Ong and Amanda Barnard. They have all been extremely patient, they have provided enormous amounts of good advice and discussion about technical details of my research, writing and about where I might be trying to go in the end and how I might get there. This research was motivated by a seemingly compelling idea about the relationship between causal models and the purpose of causal modelling, and a great deal of time was taken up with struggling to fashion this idea into a coherent and comprehensible theory. I sometimes doubted whether there was a worthwhile payoff in the end, or whether I would be able to find it. The support from my supervisors in all its forms helped me to find a path forward and to believe it was still a path worth following.

I would also like to thank Parastoo Sadeghi, Tom Everitt, Sarita Rosenstock and Zhen-Yue Chin for listening to my ideas, offering feedback, encouragement and organisational advice, and Amardeep Wander for offering flexible employment throughout much of the time I have been working on this research. Both the flexibility and the employment are deeply appreciated. I would also like to thank my parents Julie Permezel and Dennis Johnston for their help with childcare at the end of writing.

I would finally like to express my deepest gratitude to my partner, Mevlana Adil, who has gone far out of her way to support me both to commence and to finish writing this thesis through a few wild years. Without her love and support finishing would have been much less likely. I also want to acknowledge my daugher Anaïs for her understanding and cooperation beyond my expectations for a person her age.

Abstract

Mathematical formalisms of causal inference usually depend on theories of causation, and are often used to analyse problems of data-driven decision making. We show that it is possible to formalise data-driven decision problems and analyse key assumptions using a more minimal theory that aims only to satisfy the requirements of decision makers, and not to additionally offer an account of causation.

Motivated by the literature on decision theory, we consider maps from a decision maker's set of options to probability distributions on a common sample space to be the object of our study, which we call a *decision model*. We extend standard probability theory to a theory of *probability sets* to support reasoning with models of this type. We also make use of a string diagram notation for stochastic functions.

Drawing nontrivial conclusions from decision making models requires nontrivial assumptions. Such assumptions are usually formulated using a theory of causation. We propose that symmetries of decision models may cut out this "causal detour". In particular, we investigate the assumption that a sequence of pairs is related by *conditionally independent and identical responses* (henceforth: CIIR sequences). We show that this assumption is equivalent to the assumption that different infinite sequences of pairs are, in a certain sense, interchangeable – an assumption that we argue is usually unreasonable if the pairs in question are observable.

We show how causal models formulated using both the causal Bayesian network and potential outcomes approach can be represented as decision models with CIIR sequences involving latent variables. The two approaches each require a different extra assumption in order to be made compatible with ours. Causal Bayesian networks require a specification of how to "unroll" a structural model into a sequential model. A potential outcomes model requires a specification of how the decision maker's options relate to the rest of the model. Both approaches avoid the criticism of CIIR sequences we raise as the pairs in question are not fully observable.

The assumption of *precedent* is the assumption that "whatever we can do has been done before", and is weaker than the assumption of CIIR sequences. We show that the assumption of precedent in conjunction with a technical condition of *regular relationships between conditionals* can yield a conclusion of CIIR sequences when the data displays the right kind of conditional independence. The aforementioned technical condition is similar to a number of assumptions found in the literature that license the conclusion of a directed causal relationship from certain features of the given data. We speculate the assumption of precedent may offer an alternative way to understand directed causal relationships.

Contents

List of Symbols

Name	Notation	Meaning	Reference
Miscellaneous symbols			
Numbers from m to n	$[m, n]$	The set of natural numbers $\{m, m+1, ..., n\}$	
Numbers up to n	$[n]$	The set of natural numbers $\{1, 2, ..., n\}$	
Complement of $[n]$	$[n]^{\complement}$ The set $\mathbb{N} \setminus [n]$		
Iverson bracket	$\llbracket \cdot \rrbracket$	Function equal to 1 if $\cdot$ is true, false otherwise	
Directed graphs	G, V, E	Directed graph G, set of nodes V, set of edges E	Definition 5.1.1
Probability theory			
Variable	X	Measurable function $(\Omega, \mathcal{F}) \to (X, \mathcal{X})$	Definition 2.1.13
Trivial variable	$*$	Any single-valued random variable	Definition 2.1.18
Variable sequence	(X, Y)	The variable given by $\omega \mapsto (\mathsf{X}(\omega), \mathsf{Y}(\omega))$	Definition 2.1.14
Probability measure	$\mathbb{P} \in \Delta(\Omega)$	Countably additive measure on $(\Omega, \mathcal{F})$ with $\mathbb{P}(\Omega) = 1$	Definition 2.1.4
Set of probability measures	$\Delta(\Omega)$	Set of probability measures on $(\Omega, \mathcal{F})$	Notation 2.1.5
Markov kernel	$\mathbb{K} : X \rightarrow Y$	Measurable map from $(X, \mathcal{X})$ to probability measures on $(Y, \mathcal{Y})$	Definition 2.1.6
Dirac measure	δ_x	Probability measure where $\delta_x(A) = 1$ if $x \in A$, 0 otherwise	Definition 2.1.9

Name	Notation	Meaning	Reference
Markov kernel associated with a function	$\mathbb{F}_f$	Markov kernel associated with $f : X \to Y$ that maps $x \mapsto \delta_{f(x)}$	Definition 2.1.10
Marginal distribution	$\mathbb{P}^{\mathsf{X}}$	$\mathbb{P}\mathbb{F}_{\mathsf{X}}$	Definition 2.1.16
Conditional distribution	$\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$	Arbitrary Markov kernel $X \dashrightarrow Y$ such that $\mathbb{P}^{\mathsf{XY}}(A \times B) = \int_A \mathbb{P}^{\mathsf{Y}|\mathsf{X}}(B|x)\mathbb{P}^{\mathsf{X}}(\mathrm{d}x)$	Definition 2.1.17
Conditional independence	$\mathsf{X} \perp\!\!\!\perp_{\mathbb{P}} \mathsf{Y}|\mathsf{Z}$	$\mathbb{P}^{\mathsf{X}|\mathsf{YZ}}(A|y,z)$ does not depend on z	Definition 2.3.16
Uniform conditional probability	$\mathbb{P}_A^{\mathsf{Y}|\mathsf{X}}$	Arbitrary Markov kernel $X \dashrightarrow Y$ that is a conditional distribution for every $\alpha \in A$	Definition 2.3.4
Kernel product	$\mathbb{K}\mathbb{L}$	The Markov kernel given by $(A|x) \mapsto \int_Y \mathbb{L}(A|y)\mathbb{K}(\mathrm{d}y|x)$	Definition 2.1.20
Semidirect product	$\mathbb{K} \odot \mathbb{L}$	The Markov kernel given by $(A \times B|x) \mapsto \int_A \mathbb{L}(B)|y)\mathbb{K}(\mathrm{d}y|x)$	Definition 2.1.24
Permuted sequence	Y_ρ	Given $\mathsf{Y} := (\mathsf{Y}_i)_{i \in \mathbb{N}}$, $\mathsf{Y}_\rho := (\mathsf{Y}_{\rho(i)})_{i \in \mathbb{N}}$	

String diagrams			
Identity map	Id_X	Markov kernel associated with the identity function $X \to X$	Definition 2.2.1
Erase map	Del_X, $\overset{*}{|}$	Markov kernel associated with the trivial variable $*_X : X \to \{*\}$	Definition 2.2.2
Swap map	Swap_{XY}, $\times$	Markov kernel associated with the function that swaps its inputs $(x,y) \mapsto (y,x)$	Definition 2.2.3
Swap according to permuation	Swap_ρ	Markov kernel that swaps inputs in a manner specified by permuation ρ	

Name	Notation	Meaning	Reference
Copy map	Copy_X, $\curlyvee$	Markov kernel associated with the function that makes two copies of its inputs	Definition 2.2.4
Probability sets and decision models			
Decision model	$(\mathbb{P}_., (\Omega, \mathcal{F}), C)$	An option set C, a sample space $(\Omega, \mathcal{F})$ and a stochastic map from options to the sample space	Definition 2.3.1
Probability set	$\mathbb{P}_A$	A collection of probability measures $\{\mathbb{P}_\alpha | \alpha \in A\}$ on a common sample space	Definition 2.3.2
Option set	C	Interpreted as the set of options available to a decision maker	Definition 2.3.1
Nonstochastic variable	ϕ	Function defined on the option set $C \to A$	Definition 2.3.18
Complementary variables	(ϕ, ξ)	Sequence of nonstochastic variables that induces an invertible function	Definition 2.3.19
Extended conditional independence	$\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}, \phi) | (\mathsf{Z}, \xi)$	Generalisation of conditional independence to decision models	Definition 2.3.21
Choice variable	id_C	Identity function on option set C; corresponds to the choice made by decision maker	
Tabular conditional	Y^X	Variable with the property that $\mathsf{Y} = \sum_{x \in X} [\![\mathsf{X} = x]\!] \mathsf{Y}^x$; not necessarily interpretable as potential outcomes	Definition 4.3.10
Input-output model	$(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$	Shorthand for $((\mathbb{P}_., (\Omega, \mathcal{F}), (C, \mathcal{C})), \mathsf{D}, \mathsf{Y})$ with sequence of inputs D and corresponding outputs Y	Definition 4.3.1

Chapter 1

Introduction

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 (a slowing or stopping of the beating of cilia in the upper respiratory tract), 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 Ronald 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 (Wiblin, 2016; World Health Organisation, 2018).

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 and a great many facts relevant to the question

are agreed upon. Understanding how the world might be influenced is often both valuable and difficult.

There are a number of different ways one could go about learning how to influence the world from data. One option is to try to obtain data that we're confident tells us directly about the consequences of the different options under consideration. For some purposes, data produced from well-conducted experiments is widely agreed to provide reliable information about the effectiveness and safety of treatments tested in the experiment.

Alternatively, we could use the data to solve an intermediate problem, and make use of pre-existing knowledge about how to influence the world given a solution to this problem. For example, if I am on a long car trip, my tank is three quarters empty and I'm at the last fuel station for 200km, then given an answer to the question of how far my car will travel on one quarter a tank of fuel it is easy to decide whether or not I should fill up right now, and if I've logged my mileage previously I might use the data I collected to answer this question. In this example, the data don't tell me directly whether or not I should fill up.

However, we might be in a position where, we aren't so confident that the data we can acquire can provide us with reliable guidance directly about our choice, and we don't know of any surrogate problems that make the causal question straightforward. In this case, we may want to make use of a formal theory of causal inference. When we can't see the solution immediately, a theory of causal inference can help us see more clearly the consequences of things we already know. It can also provide a language that we can use to discuss assumptions and conclusions with other people.

A lot can be said for the first two options. "Collecting the right data" has driven some of the most significant recent developments in algorithmic decision making. Operational advances that enable controlled experiments to be conducted at large scales have driven substantial changes in 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 many different development interventions in many different contexts (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 in particular, replication attempts have shown that causal conclusions from experimental psychological data are less robust than many had hoped or (perhaps) believed (Open Science Collaboration, 2015; Stroebe, 2019). At the same time, standards for what constitutes a "well-conducted" experiment have risen across many fields (Liberati et al., 2009; Nosek et al., 2018).

"Solving intermediate problems" has also been behind tremendous technological advances. A machine that recognises your face has no useful impact in the world by itself, but there are a lot of people who know how to use such a machine for commercial or other purposes.

While a lot of progress can be made by getting around the problem that it's hard to make data-driven decisions that directly aim to influence the world, we think it is also interesting to attack the problem directly. One question we might want to ask is: *why* is this problem hard? While we don't claim to know for sure, a major source of difficulty seems to come from the fact that decision making requires us to consider hypotheticals.

1.1.1 Decision making requires thinking about hypotheticals

Data driven prediction problems and data driven decision making problems can have a lot in common. The outcomes some people are interested in predicting 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 vise-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, but there's a lot of overlap nonetheless.

A distinguishing feature of decision making problems is that they demand the decision maker consider a collection of hypotheticals, most of which are never realised. The strategist must consider a number of different strategies to pursue, and ultimately only learns of the outcome under the strategy their party *did* pursue. The election forecaster, on the other hand, can consider a number of different forecasts, but after the fact they learn exactly how accurate *every* forecast was and, in particular, whether the forecast they made was better than others they could have made.

Statistical probability is well-established and widely used as a formal theory of data-driven prediction. We can speculate, on the basis of the observations above, that a formal theory of data-driven decision making may be obtained by augmenting statistical probability with the right kind of hypotheticals. In fact, even though causal inference is not quite synonymous with data-driven decision making, the most widely used theories of causal inference are theories of statistical probability augmented with a particular notion of hypothetical.

To understand the role of hypotheticals in theories of causal inference and decision making, it helps to think about the role of random variables in statistical probability. Random variables have two "faces"; on one hand, they are defined as measurable functions whose domain is the sample space, and this allows us to reason about them using the tools of mathematics. However, they *also* refer to the results of measurements conducted in the real world. This feature of random variables is not a consequence of any mathematical definition, but it is this connection to the real world that allows us to use insights derived from mathematical reasoning to inform predictions we can make about real-world events.

Hypotheticals are similarly two-faced. On the one hand, particular kinds of hypotheticals have formal definitions in theories of causal inference and decision making, and on the other hand the fact that they point to something outside the mathematical model is what allows us to use the model to help us make decisions (or to do whatever else we might want to do with a causal model). Just what it is that hypotheticals in causal models refer to is a trickier question than what random variables refer to.

1.1.2 Structural interventions

One widely-used theory of causal inference uses the term *interventions* for the relevant class of hypotheticals. Interventions are defined with respect to a particular set of variables that we will call *causal variables*. A graphical causal model (for the purposes of this introduction) assigns to each causal variable a possibly empty set of *parents* (or causes) selected from the rest of the causal variables. Usually this assignment of parents has no cycles, but there are versions of interventional models that allow cycles (Bongers et al., 2016; Forré and Mooij, 2017).

In this manner, the assignment of parents can be represented with a directed acyclic graph – each causal variable is associated with a node of the graph, and an arrow $X \rightarrow Y$ appears in the graph just when X is a parent of Y. For example, the following graph
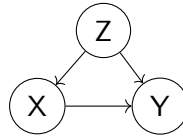
identifies three causal variables X, Y and Z, and identifies X and Z as parents of Y, Z as a parent of X and Y as a parent of X.

Given a graphical causal model and a joint probability distribution over the causal variables, an intervention on a causal variable X is formally an operation that alters the distribution of X conditional on its parents in a known way, while not affecting the distribution of any other causal variable conditional on *its* parents.

Beyond the formalism of structural causal models, *interventions* on X are usually taken to refer to actions that can be taken that alter the real thing represented by X in a predictable way while also avoiding directly influencing anything else (or at least, any other causal variable that appears in the model). While this sounds complicated, it can sometimes be reasonably clear what it means for an action to avoid influencing other causal variables; for example, take Z to be the last month's average rainfall, X to be the average number of flowers I see when I walk for the past month and Y to be how often I walk in the past month. In a common sense way, deciding to go for a walk today changes Y but has no effect on rainfall or flower growth. Also in a common sense way, someone else planting flowers might induce me to walk more often, but does not affect the rainfall or my inclination to walk holding weather and flowers constant. Finally, seeding clouds might cause more rainfall, which could cause more flowers and might affect my walking in an unpredictable way, but it probably doesn't alter the dependence of my walking on the weather and the scenery together.

However, it's not always clear how to interpret interventions. A particular example that has received extended discussion is the idea of an intervention on "obesity". Hernán and Taubman (2008); Hernán (2016) have argued that "intervention" on obesity – as measured by a patient's body mass index, or their weight divided by the square of their height – are ill-defined. They note that several different actions might alter a person's body mass index, such as diet, exercise or gastric bypass surgery, and it isn't clear which of these if any count as an intervention. Pearl (2018) responded that an intervention could be defined as an ideal action setting body mass index "performed by nature", although we find this idea hard to understand.

1.1.3 Potential outcomes

The other widely used theory of causal inference uses the term *potential outcomes* for its class of hypotheticals. Formally, potential outcomes are statistical random variables associated with a pair of "ordinary" statistical random variables. For example, if we have X again representing the mean number of flowers I see every walk and Y again representing my frequency of walks, we can define a vector of potential outcomes as a copy of Y for every possible value of X: $(Y^x)_{x \in [0,100]}$ (X is real-valued because it is a mean).

Beyond the formalism of the model, a potential outcome Y^{50} represents "how often I would've taken walks last month if I had seen 50 flowers each time on average". In cases like this, it's not clear that there is a common-sense interpretation of potential outcomes. Counterfactual

statements themselves are often difficult enough to grasp that some additional theory seems needed to make sense of them.

There are some cases where potential outcomes have a somewhat easier interpretation. A version of a classic example is when X represents whether or not a patient took antibiotics and Y represents the presence of an ear infection at a follow-up appointment. In this case Y^0 represents the presence of an ear infection "had the patient not taken antibiotics" and Y^1 represents the presence of an ear infection "had the patient taken antibiotics". If we adopt the patient's point of view, we can view these as the consequences that the patient should consider when deciding whether or not to take the medicine.

In fact, some authors have argued that potential outcomes are underpinned by choices – that is, we have potential outcomes Y^X precisely when X is a set of options that somebody could, in principle, have chosen (Imbens and Rubin, 2015, pg. 4).

In the philosophical investigation of the interpretation of counterfactuals, accounts based on structural interventions are one of the most prominent theories (Starr, 2021, Section 3.3), though there are several versions of this account and like all of the other theories of counterfactuals they are controversial.

1.1.4 Successes of theories of causal inference

Formal theories of causal inference exist to help us to draw reliable conclusions from data when informal reasoning isn't good enough to do so. A full account of the successes of intervention and potential outcomes based theories is beyond the scope of this introduction, but a brief overview will help to situate the work in this thesis in the broader context of causal inference theories.

Successes of potential outcomes

Potential outcomes models are characterised by the inclusion of potential outcomes variables, typically notated with superscripts Y^0, Y^1. These variables represent counterfactual notions – Y^0 can be read "the value that Y would have taken, had X been 0". The potential outcomes framework has been particularly influential in econometrics, with use of potential outcomes in that field predating the actual term "potential outcomes". In econometrics, and in other areas that potential outcomes have been widely used, models often involve people acting deliberately (or, sometimes, rationally) and associate potential outcomes with prospective consequences of people's choices.

Models involving rational agents: As we have noted, in some situations the potential outcomes Y^0, Y^1 and so forth look like a set of prospective consequences that a decision maker is choosing between. In some settings, we can let them be precisely that - a decision maker's expectation of the consequences of different actions they can choose. An early application of potential outcomes was to the problem of determining supply and demand curves given data on the quantity of goods exchanged and the price at which they were exchanged (Haavelmo, 1943; Tinbergen, 1930). In this early work, a "supply curve" is defined as a function that maps a hypothetical price to the quantity of goods that would be supplied, were that the actual price of goods, which is for all intents and purposes a vector of potential outcomes $(\mathsf{Y}^x)_{x \in A}$ for some set A of prices under consideration.

Note that, in this case, "what potential outcomes mean" might be able to be grounded in a theory of behaviour of buyers and sellers. We can suppose that sellers are actually asking questions like "how much would I sell if I asked for a price x?" and getting hints about the answer from buyers and other sellers. Under some idealisations – for example, maybe we

require the buyers and sellers to all agree on questions of this nature – we might be able to consider the potential outcomes to represent the the traders' answers to these questions.

Analysis of randomised experiments: The potential outcomes framework offers an account of what it is that a randomised experiment achieves so that it enables causal conclusions to be drawn. Under this framework, the critical condition is the statistical independence of the input X from the vector of potential outcomes $(\mathsf{Y}^x)_{x \in X}$ (in the potential outcomes framework, X is often called an *assignment*). If the value of X is completely determined by some physical randomisation procedure then (so the argument goes) it must be independent of the potential outcomes.

A more general kind of randomised experiment completely determines the values of inputs based on a collection of other variables W called *covariates* and some physical randomisation procedure. In this kind of experiment, the inputs are independent of the potential outcomes conditional on the covariates. When this holds, the inputs X depend probabilistically on the covariates W, and this dependence is usually called the *assignment mechanism*. When this dependence is known or able to be estimated, it can facilitate the calculation of many causal effects of interest (see Section 5.2 for more details). Using the potential outcomes framework, many techniques have been developed to estimate the assignment mechanism and to estimate causal quantities of interest given the assignment mechanism – for example, many methods and worked examples can be found in Imbens and Rubin (2015).

Because the potential outcomes framework doesn't offer a theory of counterfactuals, it doesn't come with an explanation of why the inputs are independent of the potential outcomes in a randomised experiment – this is a matter that stands outside the formal theory. One could consider an explanation that goes something like: we imagine the potential outcomes to be fixed at the time that the inputs are decided, so the inputs have no influence over them. Furthermore, if the inputs are physically randomised, the potential outcomes can have no influence over the inputs, and so the two are independent. We might also consider the converse of this claim: perhaps any reasonable theory of counterfactuals must yield the conclusion that potential outcomes are (statistically) independent of any physically randomised inputs.

Unlike the case of supply and demand, the deliberate action in a randomised experiment is the assignment carried out by the experimenter rather than the choices made by subjects of the model.

Subpopulations with different behaviour: Rather than having an input X that is completely determined by a physical randomization mechanism, sometimes experiments of interest have some physically randomized Z that influences but does not fully determine X. For example, if X records whether or not someone took a medicine, Z might record whether or not the medicine was prescribed. Experimenters might be able to have prescriptions randomized, but not the actual act of taking the medicine. The potential outcomes framework first offers us a way to understand that this is not analogous to an experiment where X is randomised: by application of probability theory, the fact that the potential outcomes are independent of Z does not mean that they are independent of X.

A notable result proven using the potential outcomes framework is that, under an assumption that prescribing a medicing never induces anyone to avoid the medicine who would otherwise have taken it ("no defiers"), it is possible to determine the effect of taking the medicine on the subpopulation of "compliers" – people who were induced by the prescription to switch from not taking the medicine to taking it. See (Imbens and Angrist, 1994) for more details.

In this setting, there are often deliberate choices carried out by the experimenter – namely, the assignment of Z – and also deliberate choices carried out by experimental subjects – the choice of X.

Successes of structural interventional models

Structural interventional models feature a collection of causal variables, and each variable is assigned a set of parents from the remaining causal variables. Each causal variable may be intervened on, which in general alters the distribution of the variable conditional on its parents while not changing the distribution of any other causal variable conditional on that variable's parents.

A formal theory of directed causal relationships: A point that is repeatedly made by Pearl (2009) and Pearl and Mackenzie (2018) is that informal notions of causal relationships play a key role in the formulation of many statistical models. For example, in the account of randomised experiments above, we said that the treatment assignment was "determined by" a physically random procedure. This statement is not backed by a formal theory, but the relation invoked by the phrase "determined by" is something like a causal relationship. It implies that the treatment assignment and the output of the randomisation are deterministically related, but "being determined by something" is not a symmetric relationship.

Structural interventional models offer a theory of causal relationships that aims to clarify these intuitions. They have been used to analyse questions like "what is the likelihood that the medicine caused the reaction?" (Pearl (2009, ch. 9), Pearl (2015)), which differ from traditional causal inference questions that are more focused on the consequences of actions than on attributing responsibility.

The question of whether the structural interventional account explains causal intuitions has been taken on by philosophers (see, for example, Woodward (2016)), but whether it is successful in doing so is contested (Cartwright, 2001).

Interventions might not be the *only* possible way to ground intuitions about directed causal relationships. An alternative proposition is that the distribution of a cause and the distribution of an effect conditional on the distribution of the cause should be *algorithmically independent*, or that the relation between them should be *generic* (Lemeire and Janzing, 2013). Because this principle can induce directed relationships between pairs of variables, it can potentially offer an account of directed causal relationships without appealing to interventions, though it has been substantially less studied than the structural intervention account of directed relationships. However, unlike the interventional theory, this does not seem to tell us how causal relationships should inform our ideas of the consequences of taking an action[1].

One of the contributions of the present work is Theorem 5.1.21 in Section 5.1.5, which shows that under an assumption of generic relationships between the conditional distributions of causes and effects together with the assumption that a proposed plan of action has a precedent then conditional independences in observed data can imply certain relationships do not change under any action the decision maker might take. Invariant relationships under action, which are taken to be axiomatic in the theory of structural interventions, can be shown to follow from the assumption of generic relationships between conditional distributions and the previously mentioned assumption that actions are precedented. We think this suggests that it may be possible to forge a unified view of directed causal relationships that subsumes both the notion

---

[1]Throughout this thesis, we use the term "intervention" to refer to an operation defined in the structural interventional account of causal inference, or the interpretation of this operation. We use the word "action" to refer to something that may or may not be interpretable as an intervention.

that causal relationships should be invariant under action and the notion that conditional distributions should be generically related in the causal direction, although precisely how to do this is an open question.

Causal inference under generic assumptions: Traditionally, analysis of causal inference problems involves certain non-generic assumptions like the assumption of independence of inputs and potential outcomes. These assumptions are non-generic because they do not apply to arbitrary causal inference problems, and so the analysis made under these assumptions can only be applied to data generated in particular contexts (for example, in controlled experiments).

The structural models tradition, however, has fostered the analysis of *causal discovery*, which is the problem of learning causal relationships from data which may not be known to satisfy certain strong assumptions. There are two main approaches to causal discovery: conditional independence-based causal discovery infers a family of causal graphs from conditional independences inferred from a dataset, while the previously mentioned theory that conditionally probabilities should be algorithmically independent in the causal direction has led to a wide variety of different approaches for discovering the direction of causation. Early examples of conditional independence based inference are the PC algorithm and the Causal Inference Algorithm (Spirtes et al., 2000, Ch. 5 & 6) and Greedy Equivalent Search (Chickering, 2002, 2003), while more recently it has been discovered that the problem of searching for a graph satisfying inferred conditional independences can be posed as a continuous optimization problem (Ng et al., 2019; Zheng et al., 2018). Examples of the applications of the idea of algorithmic independence include methods based on the assumption of additive noise (Hoyer et al., 2009; Shimizu et al., 2006) and so-called *information geometric causal inference* (IGCI) methods (Daniusis et al., 2012). Reviews of these methods can be found in (Peters, Janzing et al., 2017, ch. 4, 5, 6 & 7) and (Mooij, J.M. et al., 2016).

While the aim of this analysis is to discover causal relationships from generic data, in practice the key assumptions are not completely generic. Uhler et al. (2013) examined how frequently the key assumption of λ-*faithfulness* underpinning the conditional independence based approach is violated, and finds that (under their assumptions) models with more than 10 variables and relatively dense causal connections almost always violate the condition. Owing to the fact that algorithmic independence is incomputable and suitable approximations have to be found for practical algorithms, the algorithmic independence based approach has typically involved special conditions like linear causal relationships with non-Gaussian additive noise.

Causal identification in complex models: Potential outcomes approaches have proposed a wide variety of sufficient assumptions for estimating causal effects, but there are models in which causal effects can be estimated that are typically ignored by the potential outcomes approach. The graphical models community usually separates the problem of *identification* and *estimation*; a causal effect is *identified* if it can be computed from the joint probability of the observed variables. If this is possible, then the causal effect could in principle be calculated from an estimate of the joint probability (though estimating the entire joint probability is usually more than is needed).

One of the simplest graphical models in which causal effects are identified that haven't received much attention in the potential outcomes literature is the "front-door" condition. Given a graphical model for which it is impossible to identify the effect of X on Y, but the effect of X on W and the effect of W on Y can be separately identified, then it is possible to compose the two to identify the effect of X on Y (Pearl, 2009, Section 3.3.2). In fact, a complete characterisation of the identifiability of graphical models has been given by Shpitser

and Pearl (2008), and a more recent alternative characterisation is presented in Richardson, Evans et al. (2017).

1.1.5 Challenges to popular theories of causal inference

Despite the fact that both the potential outcomes and structural interventional approaches have substantially advanced everyone's understanding of causal inference, we believe that both approaches face difficulties that make them hard to apply outside of special settings. The difficulty with potential outcomes is easy to state: the potential outcomes framework offers no clear theory of counterfactuals. This makes it difficult to even ask if we are making valid counterfactual inferences in general, and pragmatically we rely on the appropriate interpretation of counterfactuals being obvious to everyone who needs to interpret the model.

As we noted earlier, some authors have suggested that potential outcomes should represent the potential consequences of choices or actions, and that this might perhaps form the basis of a theory of counterfacutals. The work in this thesis builds a theory of causal inference starting from the view that the fundamental problem to be addressed is a problem of making informed choices. We don't have a particularly strong view on whether our theory is fundamentally different to potential outcomes, or whether the two approaches are fundamentally similar but look different because we focus on modelling an abstract problem of making informed choices, while potential outcomes is often focused much more on various concrete problems.

Alternatively, some authors have argued that the theory of structural interventions is the appropriate theory of counterfactuals for potential outcomes (Pearl, 2009, chap. 7). In this case, the following remarks on structural interventional theories apply.

Difficulties for structural models

If a decision maker has a sound informal understanding of causal relationships relevant to their problem, structural models are often an excellent tool to formalise this understanding and derive conclusions from it. However, if a decision maker is dealing with a problem where they do *not* have a sound informal understanding of causal relationships, what does the structural approach offer them? The structural model community might suggest that they perform *causal discovery*; this is some procedure that takes their data and offers them a best guess of the structural model associated with this data.

What role should this learned structural model play in the decision maker's subsequent deliberation? We propose three answers to this question:

1. The structural model tells the decision maker what their options are and what consequences are likely to follow

2. The structural model can be combined with the decision maker's prior knowledge of what their options are to offer an assessment of their consequences

3. The structural model and the decision maker's options coevolve; perhaps the decision maker has an initial idea of what their options are which motivates a particular avenue of causal discovery which, in turn, migh prompt the decision maker to reevaluate their options and so forth

We consider the second and third answers reasonable, though the third answer is beyond the scope of this thesis. However, these two answers seem to be in tension with typical practice in causal discovery. Causal discovery algorithms typically take only the given data as input, and depend in no way on any specification of the decision maker's options. Despite this, whether

or not a structural model can offer an assessment of the consequences of a set of options is sensitive to the options under consideration.

Example 1.1.1 (Different sets of options require different models)**.** Suppose on day i, at some point during the day, a volunteer Ella looks at the current outside temperature and logs whether it is "cold" or "hot" as T_i, whether or not she's wearing a jumper as X_i and whether she feels cold, comfortable or hot as Y_i. Under normal circumstances, in cold weather she usually wears a jumper and feels comfortable, and in hot weather she wears no jumper and feels comfortable. If she is uncharacteristically not wearing a jumper on cold days, she feels cold, and if she is wearing one on hot days she feels hot.

Suppose she's asked to wear her jumper no matter what. In this case, she will feel comfortable on cold days as before, and will feel hot on hot days also as before. Under this "intervention", the relationship between her perceived body temperature and the joint specification of her clothing and the weather is unchanged. If we assume an acyclic structural model, that the instruction to wear a jumper should indeed be modeled by an intervention in this model, and that the temperature may be a cause but not an effect of body temperature and jumper wearing, then we can conclude from the results of the intervention that jumper wearing causes body temperature perception and that this relationship is unconfounded given the daily temperature:

$$(1.1)$$

Suppose we *also* want to consider the effect of actions affecting Ella's perceived body temperature. We could ask her to exercise intensely before filling out the survey on some days. What we find is, because exercise raises her body temperature, after exercising she feels comfortable with no jumper in cold weather and feels hot in hot weather with no jumper. However, we *also* find that she now prefers not to wear a jumper in cold weather. This finding does *not* correspond to the structural model (1.1).

The following modification to this structure also does not yield the desired result:
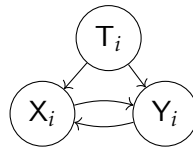
Under normal circumstances, the condition $\mathsf{Y}_i =$ "feel hot" always happened when Ella was wearing a jumper, but under the exercise condition it corresponds to no jumper wearing. Thus the conditional distribution of X_i given T_i and Y_i is not the same in the exercise condition to the normal condition.

If we introduce a new unobserved variable Y^X that represents Ella's beliefs about how she would feel if she were or were not to wear a jumper, then the following structural model can

yield the desired result for both interventions (note: there are also other possibilities):
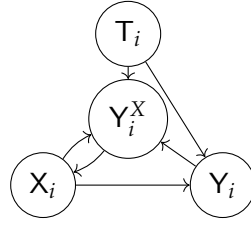
Here, we say that Ella's beliefs about her perceived body temperature are influenced by the outside temperature, whether or not she's wearing a jumper and her actual perceived body temperature. Here, once again, her perceived temperature is caused only by T_i and X_i, which reproduces the fact that her perceived temperature depends on these variables in exactly the same way after intervention on her jumper wearing. However, the only cause of her jumper wearing is her beliefs about how comfortable she will feel with a jumper on, which can be influenced by interventions on her body temperature. Thus this model can also reproduce the assumed fact that when her body temperature is intervened on, she acts to maintain a comfortable equilibrium.

Concretely, taking -1 to be "cold", 0 to be "comfortable", 1 to be "hot", 0 also for "no jumper" and 1 for "jumper", we specify the model as

$$\mathsf{T}_i \sim U(\{-1,0\})$$

$$\mathsf{Y}_i^X \leftarrow \begin{cases} x \mapsto x-1 & \text{if } \mathsf{Y}_i = \mathsf{X}_i - 1 \\ x \mapsto x & \text{if } \mathsf{Y}_i = \mathsf{X}_i \\ x \mapsto 1 & \text{if } \mathsf{Y}_i = \mathsf{X}_i + 1 \end{cases}$$

$$\mathsf{X}_i \leftarrow \begin{cases} 1 & \text{if } \mathsf{Y}_i^X = x \mapsto x-1 \\ 0 & \text{otherwise} \end{cases} \tag{1.2}$$

$$\mathsf{Y}_i \leftarrow \mathsf{T}_i + \mathsf{X}_i \tag{1.3}$$

Here, a left arrow indicates a causal assignment. This has a unique solution for each value of T_i. Instructing Ella to wear a jumper is modelled by replacing the right hand side of Eq. (1.2) with 1 and instructing Ella to exerciese is modelled by replacing the right hand side of Eq. (1.3) with $\max(1, \mathsf{T}_i + \mathsf{X}_i + 1)$. See Bongers et al. (2016); Forré and Mooij (2020) for a much more in-depth treatment of structural models with cycles, and see Eberhardt, Hoyer et al. (2010); Ghassami (2020) for algorithms for discovering linear cyclic causal structures

Suppose, finally, we are interested in the results of a) tampering with the device Ella uses to determine the temperature each day and b) putting up a large structure to shade Ella's house. Both of these actions are likely to affect the reading T_i, but otherwise have different consequences. It does not seem possible, therefore, that any single structural model limited to the variables X_i, Y_i and T_i and functions thereof can represent the consequences of both of these actions.

In this example, different kinds of models are suitable for assessing the consequences of different sets of options, and a model accommodating all of the options considered is substantially more complex than a model accommodating only the request to wear a jumper. Maybe it is in principle possible to come up with a structural model that covers every set of options anyone might want to consider – though it's far from obvious that this really is possible – but it would be very surprising to us if there was any practical way to do so.

Practically, there seems to be some tension between the views that structural models should prescribe exactly what can be done and the view that they should be flexible enough to accommodate a decision maker's needs. For example, we can find in the literature a wide variety of types of intervention that can be considered alongside structural models: beyond the standard "perfect interventions" (Hauser and Bühlmann, 2012; Pearl, 2009, ch. 1) we have soft interventions (Correa and Bareinboim, 2020; Eberhardt and Scheines, 2007), general or fat-hand interventions (Eberhardt and Scheines, 2007; Glymour and Spiegelman, 2017; Yang et al., 2018) and general interventions with unknown targets (Brouillard et al., 2020). Offering such a variety of different kinds of interventions seems to acknowledge that decision makers need some flexibility to specify structural models that will suit their needs.

On the other hand, evaluation of causal discovery research almost invariably employs a measure that does not depend on any set of options under consideration at all – as in the structural hamming distance, which counts the number of edges that differ between an inferred structure and a putative "true" structure – or that assumes that options are given by perfect interventions with respect to the true structure, as in the structural intervention distance (Peters and Bühlmann, 2015). To offer just a few examples, Brouillard et al. (2020); Chickering (2003); Forré and Mooij (2018); Ng et al. (2019); Scherrer et al. (2022); Spirtes et al. (2000); Toth et al. (2022); Zheng et al. (2018) all evaluate their methods according to one or both of these kinds of measures.

Why do we have on the one hand an acknowledgement of the need to allow "interventions" in structural models to be flexible enough to accommodate a decision maker's needs, while on the other hand causal discovery methods are evaluated only according to a rigid interpretation of perfect interventions? One of the reasons for this, we presume, is that if we consider a very broad class of interventions – say, general interventions with unknown targets – then a structural model places no constraints on what consequences can be achieved by an arbitrary intervention. However, we want causal discovery to tell us something useful, and restricting our attention to what it tells us about perfect interventions ensures that at least it tells us something nontrivial. Exactly when this is also something useful, we don't know.

In short, there is some conceptual difficulty for structural models in determining how they should interface with a given decision maker's options. In this thesis, we ask (loosely speaking) the reverse of this question; rather than starting with a structural model and asking how we can accommodate a decision maker, we start with a decision maker and ask what sorts of models they might want to use.

1.2 Outlining our approach

As we have noted, decision making requires the consideration of a collection of hypotheticals – specifically, a decision maker must consider the options she has available, and specifically wants to consider the consequences of choosing each of these options. Our approach is to suppose that our job is to help a decision maker evaluate their options. This is an idealisation; a lot of causal analysis doesn't end up directly making a decision, but it might help a third party's decisions in ways the analyst may or may not anticipate. However, it's a different idealisation to the frameworks discussed above. Potential outcomes depend on certain counterfactual statements, structural models depend on ideal interventions and our approach depends on a decision maker's set of options. However, decision making is often at least indirectly the aim of causal inference, and decision problems require the decision maker to consider a set of options, while they do not require a decision to consider counterfactuals or interventions.

This approach, which can be called a *decision theoretic approach to casual inference*, has previously been explored by Heckerman and Shachter (1995) as well as Dawid (2000, 2002); Dawid (2020); Dawid (2012). Our approach builds on this earlier work, with a particular focus on the way assumptions of symmetry or regularity in sequential decision problems can lead to models that support non-trivial inferences from data.

To a decision maker, our approach offers the possibility of analysing their problem with fewer assumptions. The two approaches surveyed above require a decision maker who wishes to formally pose their problem to accept a set of options to consider, probability theory, an account of causal effects (whether structural or some other kind of counterfactual) and some "bridging" assumption that links their options to the causal effects. In contrast, we only require them to accept a set of options and probability theory. In either case, the decision maker will also have to make additional assumptions that reflect their best guesses about how to use their available data to make a good choice. A key question is whether this approach facilitates solutions to practical decision making problems with assumptions that differ from those that must be made under the existing widely used frameworks.

A basic condition that corresponds approximately to unconfoundedness in standard causal analysis is the assumption of *conditionally independent and identical responses*. In the spirit of De Finetti's analysis of conditionally independent and identically distributed sequences, we examine in Chapter 4 the relationship between conditionally independent and identical responses and symmetries of a decision making model. This offers an alternative means of analysing assumptions of unconfoundedness in terms of the interchangeability of different datasets. The basic idea here is, instead of making assumptions about which structural causal relationships hold or how counterfactual quantities are distributed, we phrase assumptions in terms of which experiments are essentially identical. For example, the assumption that the *response functions* in a model of a combined sequence of observational and experimental data are identical, we show, is equivalent to the assumption that collecting only the observational data is essentially identical to collecting only the experimental data. We regard this as a mostly negative result; for most decision problems, a precise assumption of identical response functions is usually untenable, though it may be tenable in some approximate form.

The established approaches to causal inference have received a great deal of cumulative development effort, and many specific applications have already been studied. In Chapter 5, we show how to represent Causal Bayesian Networks (a kind of structural causal model) and Potential Outcomes models as decision making models. In particular, we show how each framework by default leaves complementary pieces of a model underspecified. Causal Bayesian Networks are, by default, non-sequential and so one must specify how they can be "unrolled" into a sequence model in order to use them to model a decision problem. Potential outcomes models are sequential, but do not provide an unambiguous way to include the options available to the decision maker into the model. Because we can perform this translation, this means that it's possible in principle to translate established application specific reasoning to our framework.

The results from Chapter 4 are mostly negative in their practical relevance – certain assumptions that would make inference from data possible are usually untenable. In Chapter 5 we explore the weaker assumption of *precedent* ("whatever I can do, it's been done by somebody before"). We prove Theorem 5.1.21, which establishes that, subject to the additional assumption of *generic probabilistic relations between conditionals*, one can conclude from a conditional independence in observed data that there is a corresponding invariant relationship among the consequences of every option available to the decision maker.

In the structural models literature generic probabilistic relations between conditionals have been used as an assumption to justify the inference of directed causal relationships. A classic result from Meek (1995) establishes that an assumption of generic probabilistic relations between conditionals "in the causal direction" can support the assumption of *faithfulness*, which facilitates the inference of structural models from conditional independences in observed data. An observation subsequently made by a number of authors, for example Lemeire and Janzing (2013), is that these kinds of generic relationships often intuitively hold in the causal direction, and can be violated in the countercausal direction. We use the same idea of generic relationships between conditionals, but by making the assumption of precedent we can skip many steps of reasoning. To use the assumption of generic relationships between conditionals to infer the consequences of choosing different options, a structural modeller needs to infer the causal structure, assume sufficiency (or some other assumption strong enough to support structural identifiability) and then make additional assumptions to connect the intervention operations in the structural model back to the actual options under consideration. With our result, a decision maker can make a similar assumption of generic relationships between conditionals, assume that the consequences of their choices have precedent, and immediately conclude from an observed conditional independence that an observed relationship is invariant over their actual options at hand. Furthermore, we argue in Section 5.1.5 that an assumption similar to the assumption of precedent is very often built into structural models.

In Theorem 5.1.21, we assume generic relationships between the key conditional distributions. There is a large literature on methods to identify "generic relationships" between conditional probabilities in order to infer causal directions. An open question is whether these methods provide evidence the right kind of "genericity" for the application of Theorem 5.1.21.

In Chapter 5 we also examine additional justifications of the assumption of independent and identical response functions. We show how, in addition to the criterion of equivalent prediction problems presented in Chapter 4, this assumption can also in some circumstances be justified by an assumption of equivalent options, or by considering certain variables that determine the manner in which outputs respond to inputs to be "pseudo-observable" and impose constraints that we might impose on observed variables that play a similar role.

Chapters 2 and 3 provide context and prerequisites for the work that follows. In particular, 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 and Sutton, 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).

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 3 we survey the literature on modelling decision problems, and show that many different decision theories posit that decision models, including *evidential decision theory* and *causal decision theory* share the same basic type. We also show how a particular class of

decision making models – a class that contains the models investigated in all subsequent chapters – induce classical statistical decision problems.

The mathematical basis for essentially all of the work in this thesis is probability theory, though we make use of nonstandard constructions within the theory to facilitate the representation of sets of options in the models we consider. We introduce the relevant theory in Chapter 2. As is common in causal inference, we often use a graphical language to represent probabilistic decision models. The language we use is somewhat different to the directed graphs that are standard in the area. We use a string diagram notation that can be related to ordinary directed graphs as in (Fong, 2013), but which also supports equality statements and a collection of transformations that can be applied to a diagram to yield an equivalent diagram.

We conclude in Chapter 6, revisiting key results from this thesis and surveying important questions that they raise.

1.2.1 Summary of motivation and key differences of new approach

This is a brief summary of the motivation for exploring a new approach to causal modelling and key differences to existing systems. As this is a summary, it necessarily neglects many subtleties of the discussion.

A key challenge in causal inference is that it requires assumptions that are not generally valid and that are hard to empirically evaluate in the context of interest. There are two existing approaches to causal modelling: potential outcomes models and structural interventional models. Both are used to formally model data-driven decision problems, and both feature a representation of causation that does not depend on the options a decision maker has available. While it is sometimes the case that the correspondence between options and counterfactuals or interventions is intuitively obvious, it is difficult to give a general account of the correspondence, and in fact explaining counterfactuals in general is an open philosophical problem. The approach to causal modelling presented here instead takes a decision maker's options as the causal primitive, and does not aim to offer a problem independent account of causation. A motivating hypothesis – and one that is far from proven – is that by avoiding the need to offer a general account of causation, we may be better placed to understand the assumptions requierd for valid causal inference.

	Decision Models	Potential Outcomes Mo…
Causal primitive	options	potential outcomes
What they don't say	what causes what/what would have happened counterfactually	counterfactual to option… respondence
problem specific set of options being considered	alternative versions of key variables under counterfactual premises	consequences of actin… change key variables in p… ular ways
Identification assumptions	pairs of variables with invariant conditional probabilities	conditional indepen… of potential outcomes … premise
Graphical notation	string diagrams	various
Key questions	observation to consequence bridge principles	conditional independen… counterfactual variables

Chapter 2

Technical Prerequisites

Our approach to causal inference is based on probability theory. Many results and conventions will be familiar to readers, and these are collected in Section 2.1.1.

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. This is introduced in Section 2.2.

We also extend the theory of probability to a theory of probability sets, which we introduce in Section 2.3. This section goes over some ground already trodden by Section 2.1.1; this structure was chosen so that people familiar with the Section 2.1.1 can skip to Section 2.3 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.

Sections 2.4 and 2.5 are not critical for understanding the work in the rest of this thesis.

We introduce the assumption of *validity* in Section 2.4, a condition that ensures probability sets constructed by "assembling" collections of uniform conditionals are non-empty. This is relevant to sanity checking models with interventions (see Definition 5.1.12 and the remarks following it).

We use many standard definitions in our setup – for example, *variables* (Definition 2.1.13) are a standard definition in probability theory. On the other hand, we also make use of non-standard notions, like probability sets and decision models (Definition 2.3.1). In Section 2.5 we attempt an account of how the formal elements of our theory could be interpreted. The two objects we focus our attention on are *variables* and *options*.

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 introduce one additional probabilistic primitive: *combs*. They have been moved to this chapter as we feel that additional context is helpful for understanding them.

2.1 Probability Theory

2.1.1 Standard Probability Theory

σ-algebras

Definition 2.1.1 (Sigma algebra)**.** Given a set A, a σ-algebra $\mathcal{A}$ is a collection of subsets of A where

- $A \in \mathcal{A}$ and $\emptyset \in \mathcal{A}$

- $B \in \mathcal{A} \implies B^{\complement} \in \mathcal{A}$

- $\mathcal{A}$ is closed under countable unions: For any countable collection $\{B_i | i \in Z \subset \mathbb{N}\}$ of elements of $\mathcal{A}$, $\cup_{i \in Z} B_i \in \mathcal{A}$

Definition 2.1.2 (Measurable space)**.** A measurable space $(A, \mathcal{A})$ is a set A along with a σ-algebra $\mathcal{A}$.

Definition 2.1.3 (Sigma algebra generated by a set)**.** Given a set A and an arbitrary collection of subsets $U \supset \mathcal{P}(A)$, 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 $(\mathbb{R}, B(\mathbb{R}))$ or $(C, \mathcal{P}(C))$ for denumerable C (Çinlar, 2011, Chap. 1).

Probability measures and Markov kernels

Definition 2.1.4 (Probability measure)**.** Given a measurable space $(E, \mathcal{E})$, a map $\mu : \mathcal{E} \to [0, 1]$ 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 2.1.5 (Set of all probability measures)**.** The set of all probability measures on $(E, \mathcal{E})$ is written $\Delta(E)$.

Definition 2.1.6 (Markov kernel)**.** Given measurable spaces $(E, \mathcal{E})$ and $(F, \mathcal{F})$, a *Markov kernel* or *stochastic function* is a map $\mathbb{M} : E \times \mathcal{F} \to [0, 1]$ 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 2.1.7 (Signature of a Markov kernel)**.** Given measurable spaces $(E, \mathcal{E})$ and $(F, \mathcal{F})$ and $\mathbb{M} : E \times \mathcal{F} \to [0, 1]$, we write the signature of $\mathbb{M} : E \rightarrow F$, read "$\mathbb{M}$ maps from E to probability measures on F".

Definition 2.1.8 (Deterministic Markov kernel). A *deterministic* Markov kernel $\mathbb{A} : E \to \Delta(\mathcal{F})$ 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 2.1.9 (Dirac measure). The *Dirac measure* $\delta_x \in \Delta(X)$ is a probability measure such that $\delta_x(A) = [\![x \in A]\!]$

Definition 2.1.10 (Markov kernel associated with a function). Given measurable $f : (X, \mathcal{X}) \to (Y, \mathcal{Y})$, $\mathbb{F}_f : X \rightarrowtail Y$ is the Markov kernel given by $x \mapsto \delta_{f(x)}$

Definition 2.1.11 (Markov kernel associated with a probability measure). Given $(X, \mathcal{X})$, a one-element measurable space $(\{*\}, \{\{*\}, \varnothing\})$ and a probability measure $\mu \in \Delta(X)$, the associated Markov kernel $\mathbb{Q}_\mu : \{*\} \rightarrowtail X$ is the unique Markov kernel $* \mapsto \mu$

Lemma 2.1.12 (Products of functional kernels yield function composition). *Given measurable $f : X \to Y$ and $g : Y \to Z$, $\mathbb{F}_f \mathbb{F}_g = \mathbb{F}_{g \circ f}$.*

Proof.

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

$\square$

Variables, conditionals and marginals

We offer the standard definition of a random variable.

Definition 2.1.13 (Random variable). Given a measurable space $(\Omega, \mathcal{F})$ and a measurable space of values $(X, \mathcal{X})$, an *X-valued random variable* is a measurable function $\mathsf{X} : (\Omega, \mathcal{F}) \to (X, \mathcal{X})$.

A sequence of random variables is also a random variable.

Definition 2.1.14 (Sequence of variables). Given a measurable space $(\Omega, \mathcal{F})$ and two random variables $\mathsf{X} : (\Omega, \mathcal{F}) \to (X, \mathcal{X})$, $\mathsf{Y} : (\Omega, \mathcal{F}) \to (Y, \mathcal{Y})$, $(\mathsf{X}, \mathsf{Y}) : \Omega \to X \times Y$ is the random variable $\omega \mapsto (\mathsf{X}(\omega), \mathsf{Y}(\omega))$.
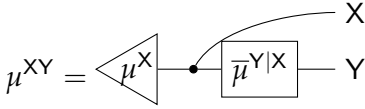
We define a partial order on random variables such that Y is higher than X if X is given by application of a function to Y. For example, $\mathsf{Y} \preccurlyeq (\mathsf{W}, \mathsf{Y})$ as Y can be obtained by composing a projection with (W, Y).

Definition 2.1.15 (Random variables determined by another random variable). Given a sample space $(\Omega, \mathcal{F})$ and variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, $\mathsf{X} \preccurlyeq \mathsf{Y}$ if there is some $f : Y \to X$ such that $\mathsf{X} = f \circ \mathsf{Y}$.

Definition 2.1.16 (Marginal distribution). Given a probability space $(\mu, \Omega, \mathcal{F})$ and a variable $\mathsf{X} : \Omega \to (X, \mathcal{X})$, the *marginal distribution* of X with respect to μ, $\mu^{\mathsf{X}} : \mathcal{X} \to [0, 1]$ by $\mu^{\mathsf{X}}(A) := \mu(\mathsf{X}^{-1}(A))$ for any $A \in \mathcal{X}$.

Definition 2.1.17 (Conditional distribution)**.** Given a probability space $(\mu, \Omega, \mathcal{F})$ and variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, the *conditional distribution* of Y given X is any Markov kernel $\mu^{\mathsf{Y}|\mathsf{X}} : X \rightarrow Y$ such that

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

$$\Longleftrightarrow$$

Definition 2.1.18 (Trivial variable)**.** We let $*$ stand for any single-valued variable $* : \Omega \to \{*\}$.

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 2.1.19 (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 2.1.20 (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

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

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

Definition 2.1.21 (Kernel-function product)**.** Given $\mathbb{M} : X \rightarrow Y$ and $f : Y \to Z$, the *kernel-function product* $\mathbb{M}f : X \to Z$ is given by

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

for all $x \in X$.

Definition 2.1.22 (Tensor product)**.** Given $\mathbb{M} : X \rightarrow Y$ and $\mathbb{L} : W \rightarrow Z$, the tensor product $\mathbb{M} \otimes \mathbb{N} : X \times W \rightarrow Y \times Z$ 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 2.1.23.

Lemma 2.1.23 (Marginal distribution as a kernel product)**.** *Given a probability space* $(\mu, \Omega, \mathcal{F})$ *and a variable* $\mathsf{X} : \Omega \to (X, \mathcal{X})$, *define* $\mathbb{F}_{\mathsf{X}} : \Omega \dashrightarrow X$ *by* $\mathbb{F}_{\mathsf{X}}(A|\omega) = \delta_{\mathsf{X}(\omega)}(A)$, *then*

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

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

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

$\square$

Semidirect product

Given a marginal μ^{X} and a conditional $\mu^{\mathsf{Y}|\mathsf{X}}$, the product of the two yields the marginal distribution of Y: $\mu^{\mathsf{Y}} = \mu^{\mathsf{X}} \mu^{\mathsf{Y}|\mathsf{X}}$. We define another product – the *semidirect* product $\odot$ – as the product that yields the joint distribution of (X, Y): $\mu^{\mathsf{XY}} = \mu^{\mathsf{X}} \odot \mu^{\mathsf{Y}|\mathsf{X}}$. The semidirect product is associative (Lemma 2.1.25)
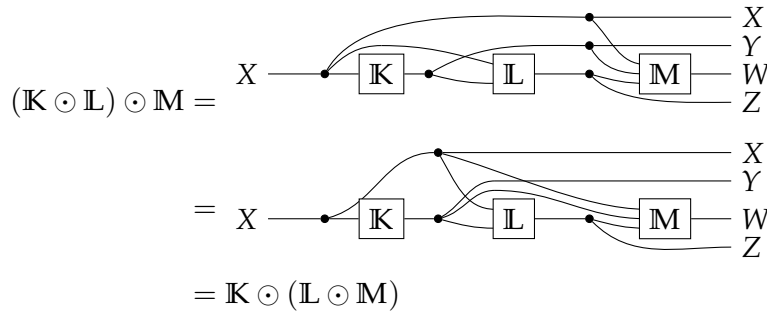
Definition 2.1.24 (Semidirect product)**.** Given $\mathbb{K} : X \dashrightarrow Y$ and $\mathbb{L} : Y \times X \dashrightarrow Z$, the semidirect product $\mathbb{K} \odot \mathbb{L} : X \to Y \times Z$ is given by

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

Lemma 2.1.25 (Semidirect product is associative)**.** *Given* $\mathbb{K} : X \dashrightarrow Y$, $\mathbb{L} : Y \times X \dashrightarrow Z$ *and* $\mathbb{M} : Z \times Y \times X \dashrightarrow W$

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

Proof.

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

$\square$

The semidirect product can be used to define a notion of almost sure equality: two kernels $\mathbb{K} : X \dashrightarrow Y$ and $\mathbb{L} : X \dashrightarrow Y$ 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, $\mathbb{K} \overset{\mu}{\cong} \mathbb{L}$ if and only if $\mathbb{K} = \mathbb{L}$ μ-almost everywhere.

Definition 2.1.26 (Almost sure equality)**.** Two Markov kernels $\mathbb{K} : X \nrightarrow Y$ and $\mathbb{L} : X \nrightarrow Y$ are almost surely equal $\overset{\mathbb{P}_C}{\cong}$ with respect to a probability space $(\mu, X, \mathcal{X})$, written $\mathbb{K} \overset{\mu}{\cong} \mathbb{L}$ if

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

Theorem 2.1.27. *Given* $(\mu, X, \mathcal{X})$, $\mathbb{K} : X \nrightarrow Y$ *and* $\mathbb{L} : X \nrightarrow Y$, $\mathbb{K} \overset{\mu}{\cong} \mathbb{L}$ *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. $\qquad\square$

We often want to talk about almost sure equality of two different versions $\mathbb{K}$ and $\mathbb{L}$ of a conditional distribution $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ with respect to some ambient probability space $(\mathbb{P}, \Omega, \mathcal{F})$. This simply means $\mathbb{K}$ and $\mathbb{L}$ satisfy Definition 2.1.17 with respect to $\mathbb{P}$, X and Y, and they are almost surely equal with respect to the marginal $\mathbb{P}^{\mathsf{X}}$. The relevant variables are usually obvious from the context and we leave them implicit and we will write $\mathbb{K} \overset{\mathbb{P}}{\cong} \mathbb{L}$. If the relevant marginal is ambiguous, we will instead write $\mathbb{K} \overset{\mathbb{P}^{\mathsf{X}}}{\cong} \mathbb{L}$

Definition 2.1.28 (Almost sure equality with respect to a pair of variables)**.** Given $(\mathbb{P}, \Omega, \mathcal{F})$ and $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, two Markov kernels $\mathbb{K} : X \nrightarrow Y$ and $\mathbb{L} : X \nrightarrow Y$ are X-almost surely equal with respect to $\mathbb{P}$, written $\mathbb{K} \overset{\mathbb{P}}{\cong} \mathbb{L}$, if they are almost surely equal with respect to the marginal $\mathbb{P}^{\mathsf{X}}$.

2.2 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 Cho and Jacobs (2019); Fritz (2020).

2.2.1 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:

$$\mathbb{K} := \boxed{\mathbb{K}}$$

$$\mu := \triangleleft\!\boxed{\mathbb{P}}$$

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:

$$\mathbb{L}\mathbb{M} := X\,\boxed{\mathbb{K}}\!\!-\!\!\boxed{\mathbb{M}}\,Z$$

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:

$$\mathbb{K} \otimes \mathbb{L} := \begin{matrix} W\,\boxed{\mathbb{K}}\,Y \\ X\,\boxed{\mathbb{L}}\,Z \end{matrix}$$

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

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

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

$$\mathrm{Id}^X := X \underline{\qquad\qquad} X$$

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

$$X \times Y \cong \mathrm{Id}^X \otimes \mathrm{Id}^Y := \begin{matrix} X \!-\! X \\ Y \!-\! Y \end{matrix}$$

$$= X \times Y \underline{\qquad} X \times Y$$

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:

$$X - \boxed{\mathbb{L}} = \begin{matrix} Y \\ Z \end{matrix}$$

$$\equiv$$

$$X - \boxed{\mathbb{L}} - Y \times Z$$

We read diagrams from left to right (this is somewhat different to Cho and Jacobs (2019); Fong (2013); Fritz (2020) 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.

2.2.2 Special maps

Definition 2.2.1 (Identity map). The identity map $\mathrm{Id}_X : X \to X$ defined by $(\mathrm{id}_X)(A|x) = \delta_x(A)$ for all $x \in X$, $A \in \mathcal{X}$, is represented by a bare line.

$$\mathrm{id}_X := \quad X - X$$

Definition 2.2.2 (Erase map). Given some 1-element set $\{*\}$, the erase map $\mathrm{Del}_X : X \to \{*\}$ is defined by $(\mathrm{Del}_X)(*|x) = 1$ for all $x \in X$. It "discards the input". It looks like a lit fuse:

$$\mathrm{Del}_X := \quad \rightarrow\!\!* \; X$$

Definition 2.2.3 (Swap map). The swap map $\mathrm{Swap}_{X,Y} : X \times Y \to Y \times X$ is defined by $(\mathrm{Swap}_{X,Y})(A \times B|x,y) = \delta_x(B)\delta_y(A)$ 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:

$$\mathrm{Swap}_{X,Y} := \quad \times$$

Definition 2.2.4 (Copy map). The copy map $\mathrm{Copy}_X : X \to X \times X$ is defined by $(\mathrm{Copy}_X)(A \times B|x) = \delta_x(A)\delta_x(B)$ for all $x \in X$, $A, B \in \mathcal{X}$. It makes two identical copies of the input, and is drawn as a fork:

$$\mathrm{Copy}_X := \quad X \prec \begin{matrix} X \\ X \end{matrix}$$

Definition 2.2.5 (n-fold copy map). The n-fold copy map $\mathrm{Copy}_X^n : X \to X^n$ is given by the recursive definition

$$\mathrm{Copy}_X^1 = \mathrm{Copy}_X$$

$$\mathrm{Copy}_X^n = \quad \boxed{\mathrm{Copy}_X^{n-1}} \qquad\qquad n > 1$$

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 $i \in A$,

$$\bigotimes_{i \in A} \mathbb{K}_i := \boxed{\begin{array}{c} \mathbb{K}_i \\ i \in A \end{array}} \qquad \mathrm{Copy}_X^{|A|}(\bigotimes_{i \in A} \mathbb{K}_i) \qquad := \qquad \boxed{\begin{array}{c} \mathbb{K}_i \\ i \in A \end{array}}$$

2.2.3 Commutative comonoid axioms

Diagrams in Markov categories satisfy the commutative comonoid axioms.

$$\tag{2.1}$$

$$\tag{2.2}$$

$$\tag{2.3}$$

as well as compatibility with the monoidal structure

and the naturality of *Del*, which means that

2.2.4 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 2.2.6.

Theorem 2.2.6 (Copy map commutes for deterministic kernels ([Fong, 2013](#))). *For $\mathbb{K} : X \twoheadrightarrow Y$*

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 $\mathbb{K} : X \times Y \twoheadrightarrow Z$, $\mathbb{L} : W \twoheadrightarrow Y$

$$= (\mathrm{id}_X \otimes \mathbb{L})\mathbb{K}$$

$$[(\mathrm{id}_X \otimes \mathbb{L})\mathbb{K}](A|x,w) = \int_Y \int_X \mathbb{K}(A|x',y')\mathbb{L}(\mathrm{d}y'|w)\delta_x(\mathrm{d}x')$$

$$= \int_Y \mathbb{K}(A|x,y')\mathbb{L}(dy'|w)$$

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

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

$$= (\mathrm{id}_X \otimes \mathrm{Copy}_Y)\mathbb{K}$$

$$[(\mathrm{id}_X \otimes \mathrm{Copy}_Y)\mathbb{K}](A|x,y) = \int_Y \int_Y \mathbb{K}(A|x,y',y'')\delta_y(\mathrm{d}y')\delta_y(\mathrm{d}y'')$$

$$= \mathbb{K}(A|x,y,y)$$

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

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

$$= \mathrm{Swap}_{YX}\mathbb{K}$$

$$(\mathrm{Swap}_{YX}\mathbb{K})(A|y,x) = \int_{X \times Y} \mathbb{K}(A|x',y')\delta_y(\mathrm{d}y')\delta_x(\mathrm{d}x')$$

$$= \mathbb{K}(A|x,y)$$

The swap map before a kernel switches the input arguments.

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

$$\begin{aligned}
&\raisebox{0pt}{\fbox{$\mathbb{K}$}}\!\!\prec\qquad = \mathbb{K}\mathrm{Swap}_{YZ}\\
(\mathbb{K}\mathrm{Swap}_{YZ})(A \times B|x) &= \int_{Y \times Z} \delta_y(B)\delta_z(A)\mathbb{K}(\mathrm{d}y \times \mathrm{d}z|x)\\
&= \int_{B \times A} \mathbb{K}(\mathrm{d}y \times \mathrm{d}z|x)\\
&= \mathbb{K}(B \times A|x)
\end{aligned}$$

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

$$(\mathbb{K} \odot \mathbb{L})(\mathrm{id}_Y \otimes \mathrm{Del}_Z) =$$

$$=\qquad\qquad\qquad\qquad \text{by Eq. (2.3)}$$

$$=\qquad X \!-\!\fbox{$\mathbb{K}$}\!-\! Y \qquad\qquad \text{by Eq. (2.2)}$$

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

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

2.3 Probability Sets and Decision Models

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.

First, we define a *decision model* as a triple of a probability set along with an indexing *option set* and the common sample space. We could also consider a decision model to be a function $C \rightarrow \Omega$, where we do not distinguish options options $\alpha, \alpha' \in C$ that induce the same probability distributions (i.e. $\mathbb{P}_\alpha = \mathbb{P}_{\alpha'}$).

Definition 2.3.1 (Decision model)**.** A decision model is a triple $(\mathbb{P}_., (\Omega, \mathcal{F}), (C, \mathcal{C}))$ where $\mathbb{P}_. : C \rightarrow \Omega$ is a Markov kernel, $(\Omega, \mathcal{F})$ is the shared sample space and C is the option set.

Decision models induce sets of probability distributions: for some $A \subset C$, $\mathbb{P}_A := \{\mathbb{P}_\alpha | \alpha \in A\}$ is the set of probability distributions induced by A.

Definition 2.3.2 (Probability set). A probability set $\mathbb{P}_C$ on $(\Omega, \mathcal{F})$ is a collection of probability measures on $(\Omega, \mathcal{F})$. 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 2.1.16 and 2.1.17 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 2.3.8) with respect to probability sets.

Definition 2.3.3 (Marginal probability with respect to a probability set). Given a sample space $(\Omega, \mathcal{F})$, a variable $\mathsf{X} : \Omega \to X$ and a probability set $\mathbb{P}_C$, the marginal distribution $\mathbb{P}_C^{\mathsf{X}} = \mathbb{P}_\alpha^{\mathsf{X}}$ for any $\mathbb{P}_\alpha \in \mathbb{P}_C$ if a distribution satisfying this condition exists. Otherwise, it is undefined.

Definition 2.3.4 (Uniform conditional distribution). Given a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$ and a probability set $\mathbb{P}_C$, a uniform conditional distribution $\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}$ is any Markov kernel $X \rightarrow Y$ such that $\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}$ is an $\mathsf{Y}|\mathsf{X}$ conditional probability of $\mathbb{P}_\alpha$ for all $\mathbb{P}_\alpha \in \mathbb{P}_C$. If no such Markov kernel exists, $\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}$ is undefined.

Given a conditional distribution $\mu^{\mathsf{ZY}|\mathsf{X}}$ we can define a higher order conditional $\mu^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$, which is a version of $\mu^{\mathsf{Z}|\mathsf{XY}}$. 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 $\mathbb{P}_C^{\mathsf{ZY}|\mathsf{X}}$ then it also has a uniform conditional $\mathbb{P}_C^{\mathsf{Z}|\mathsf{XY}}$ (Theorems 2.3.15 and 2.4.10). Given $\mu^{\mathsf{XY}|\mathsf{Z}}$ and $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ standard measurable, it has recently been proven that a higher order conditional $\mu^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ exists Bogachev and Malofeev (2020), Theorem 3.5.

Definition 2.3.5 (Higher order conditionals). Given a probability space $(\mu, \Omega, \mathcal{F})$ and variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and $\mathsf{Z} : \Omega \to Z$, a higher order conditional $\mu^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})} : X \times Y \to Z$ is any Markov kernel such that, for some $\mu^{\mathsf{Y}|\mathsf{X}}$,

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

$$\Longleftrightarrow$$

$$\mu^{\mathsf{ZY}|\mathsf{X}} =$$

Definition 2.3.6 (Uniform higher order conditional). Given a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and $\mathsf{Z} : \Omega \to Z$ and a probability set $\mathbb{P}_C$, if $\mathbb{P}_C^{\mathsf{ZY}|\mathsf{X}}$ exists then a uniform higher order conditional $\mathbb{P}_C^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ is any Markov kernel $X \times Y \rightarrow Z$ that is a higher order conditional of some version of $\mathbb{P}_C^{\mathsf{ZY}|\mathsf{X}}$. If no $\mathbb{P}_C^{\mathsf{ZY}|\mathsf{X}}$ exists, $\mathbb{P}_C^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ is undefined.

Definition 2.3.7 (Almost sure equality). Two Markov kernels $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : X \rightarrow Y$ are $\mathbb{P}_C, \mathsf{X}, \mathsf{Y}$-almost surely equal if for all $A \in \mathcal{X}$, $B \in \mathcal{Y}$, $\alpha \in C$

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

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 $\mathbb{P}_\alpha^\mathsf{X}$ for all $\alpha \in C$.

2.3.1 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 $\mathbb{P}_\alpha^\mathsf{X}$ with each Markov kernel is identical.

Definition 2.3.8 (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 $\mathsf{X} : \Omega \to X$ if for all $\mathbb{P}_\alpha \in \mathbb{P}_C$,

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

Lemma 2.3.9 (Uniform conditional distributions are almost surely equal). *If $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : X \rightarrow Y$ are both versions of $\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}$ then $\mathbb{K} \stackrel{\mathbb{P}_C}{\cong} \mathbb{L}$*

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

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

$\square$

Lemma 2.3.10 (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 $\mathbb{P}_\alpha \in \mathbb{P}_C$*

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

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

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

$\square$

Theorem 2.3.11 (Semidirect product of uniform conditional distributions is a joint uniform conditional distribution). *Given a probability set $\mathbb{P}_C$ on $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and uniform conditional distributions $\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}$ and $\mathbb{P}_C^{\mathsf{Z}|\mathsf{XY}}$, then $\mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}$ exists and is equal to*

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

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

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

$\square$

2.3.2 Existence of conditional distributions

It is known that conditional distributions do not exist in general for probability measures defined on arbitrary measurable sets. However, the requirement that the sets are *standard measurable* is sufficient for the existence of conditional distributions. We also further show that, given a borel measurable map to standard measurable probability distributions, what we call "higher order conditionals" also exist.

Lemma 2.3.12 (Conditional pushforward)**.** *Suppose we have a sample space $(\Omega, \mathcal{F})$, variables* $X : \Omega \to X$ *and* $Y : \Omega \to Y$, $Z : \Omega \to Z$ *and a probability set* $\mathbb{P}_C$ *with conditional* $\mathbb{P}_C^{X|Y}$ *such that* $Z = f \circ Y$ *for some* $f : Y \to Z$. *Then there exists a conditional probability* $\mathbb{P}_C^{Z|X} = \mathbb{P}_C^{Y|X} \mathbb{F}_f$.

Proof. Note that $(X, Z) = (\mathrm{id}_X \otimes f) \circ (X, Y)$. Thus, by Lemma 2.1.23, for any $\mathbb{P}_\alpha \in \mathbb{P}_C$

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

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

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

Thus

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

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 2.3.13 always

applies. We will simply assume that conditional probabilities exist, and avoid referencing this theorem every time.

Theorem 2.3.13 (Existence of regular conditionals)**.** *Suppose we have a sample space* $(\Omega, \mathcal{F})$, *variables* $\mathsf{X} : \Omega \to X$ *and* $\mathsf{Y} : \Omega \to Y$ *with* Y *standard measurable and a probability model* $\mathbb{P}_\alpha$ *on* $(\Omega, \mathcal{F})$. *Then there exists a conditional* $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}}$.

Proof. Çinlar (2011, Theorem 2.18) □

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

Theorem 2.3.14. *Given a Borel measurable map* $m : X \rightarrow Y \times Z$ *let* $\Pi_Y : Y \times Z \to Y$ *be the projection onto* Y. *Then there exists a Borel measurable map* $n : X \times Y \rightarrow Y \times Z$ *such that*

$$n(\Pi_y^{-1}(y)|x, y) = 1$$

$$m(\mathsf{Y}^{-1}(A) \cap B|x) = \int_A n(B|x, y) m\mathbb{F}_{\Pi_Y}(dy|x) \qquad \forall A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z} \qquad (2.4)$$

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

Proof. Take $\mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}$ to be the Borel measurable map m from Theorem 2.3.14, and note that $\Pi_Y \circ (\mathsf{Y}, \mathsf{Z}) = \mathsf{Y}$. Then equation (2.4) implies for all $A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z}$ there is some n such that

$$\mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}(\mathsf{Y}^{-1}(A) \cap B|x) = \int_A n(B|x, y)\mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}\mathbb{F}_{\Pi_Y}(dy|x)$$
$$= \int_A n(B|x, y)\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}(dy|x) \qquad (2.5)$$

where Equation (2.5) follows from Lemma 2.3.12.

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

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

which implies n is a version of $\mathbb{P}_C^{\mathsf{YZ}|(\mathsf{Y}|\mathsf{X})}$. By Lemma 2.3.12, $n\mathbb{F}_{\Pi_Y}$ is a version of $\mathbb{P}_C^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$. □

2.3.3 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 2.3.17).

Definition 2.3.16 (Conditional independence). For a *probability model* $\mathbb{P}_\alpha$ and variables $\mathsf{A}, \mathsf{B}, \mathsf{Z}$, we say B is conditionally independent of A given id_C, written $\mathsf{B} \perp\!\!\!\perp_{\mathbb{P}_\alpha} \mathsf{A} | \mathrm{id}_C$, if

$$\mathbb{P}^{\mathsf{Y}|\mathsf{W}\mathsf{X}} \overset{\mathbb{P}}{\cong} \quad \begin{array}{c} \mathsf{W} \rule[0.5ex]{1em}{0.4pt} \boxed{\mathbb{K}} \rule[0.5ex]{1em}{0.4pt} \mathsf{Y} \\ \mathsf{X} \rule[0.5ex]{2em}{0.4pt} * \end{array}$$

$$\iff \mathbb{P}^{\mathsf{Y}|\mathsf{W}\mathsf{X}}(A|w,x) \overset{\mathbb{P}}{\cong} \mathbb{K}(A|w) \qquad\qquad \forall A \in \mathcal{Y}$$

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

Theorem 2.3.17. *Given standard measurable* $(\Omega, \mathcal{F})$, *a probability model* $\mathbb{P}$ *and variables* $\mathsf{W} : \Omega \to W$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}} \mathsf{X} | \mathsf{W}$ *if and only if there exists some version of* $\mathbb{P}^{\mathsf{Y}|\mathsf{W}\mathsf{X}}$ *and* $\mathbb{K} : W \twoheadrightarrow Y$ *such that*

$$\mathbb{P}^{\mathsf{Y}|\mathsf{W}\mathsf{X}} \overset{\mathbb{P}}{\cong} \quad \begin{array}{c} \mathsf{W} \rule[0.5ex]{1em}{0.4pt} \boxed{\mathbb{K}} \rule[0.5ex]{1em}{0.4pt} \mathsf{Y} \\ \mathsf{X} \rule[0.5ex]{2em}{0.4pt} * \end{array}$$

$$\iff \mathbb{P}^{\mathsf{Y}|\mathsf{W}\mathsf{X}}(A|w,x) \overset{\mathbb{P}}{\cong} \mathbb{K}(A|w) \qquad\qquad \forall A \in \mathcal{Y}$$

Proof. See Cho and Jacobs (2019). $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \square$

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

Definition 2.3.18 (Nonstochastic variable). Given a sample space $(\Omega, \mathcal{F})$, 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 $\phi : C \to X$.

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 $\phi := \mathrm{idf}_C$, then

Definition 2.3.19 (Complementary nonstochastic variables). A pair of nonstochastic variables ϕ and ξ are complementary if (ϕ, ξ) is invertible.

Nxample 2.3.20. 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 2.3.21 (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 $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_C}^{e} \mathsf{X}\phi | \mathsf{Z}\xi$ holds if for each $a \in \xi(C)$, $\mathbb{P}^{\mathsf{Y}|\mathsf{X}\mathsf{Z}}_{\xi^{-1}(a)}$ and $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_{\xi^{-1}(a)}$ exist

and

$$\mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}_{\xi^{-1}(a)} \overset{\mathbb{P}_{\xi^{-1}(a)}}{\cong} \quad \begin{array}{c} \mathsf{Z} \longrightarrow \boxed{\mathbb{P}^{\mathsf{Y}|\mathsf{Z}}_{C}} \longrightarrow \mathsf{Y} \\ \mathsf{X} \longrightarrow \ast \end{array}$$

$$\Longleftrightarrow$$

$$\mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}_{\xi^{-1}(a)}(A|x,z) \overset{\mathbb{P}_{\xi^{-1}(a)}}{\cong} \mathbb{P}^{\mathsf{Y}|\mathsf{Z}}_{\xi^{-1}(a)}(A|z) \qquad\qquad \forall A \in \mathcal{Y}, (x,z) \in X \times Z$$

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 2.3.22 (Uniform conditional independence). Given a probability set $\mathbb{P}_C$ and variables X, Y and Z, the uniform conditional independence $\mathsf{Y} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} (\mathsf{X}, \mathrm{id}_C)|\mathsf{Z}$ holds if $\mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}_{C}$ and $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_{C}$ exist and

$$\mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}_{C} \overset{\mathbb{P}_C}{\cong} \quad \begin{array}{c} \mathsf{Z} \longrightarrow \boxed{\mathbb{P}^{\mathsf{Y}|\mathsf{Z}}_{C}} \longrightarrow \mathsf{Y} \\ \mathsf{X} \longrightarrow \ast \end{array}$$

$$\Longleftrightarrow$$

$$\mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}_{C}(A|x,z) \overset{\mathbb{P}_C}{\cong} \mathbb{P}^{\mathsf{Y}|\mathsf{Z}}_{C}(A|z) \qquad\qquad \forall A \in \mathcal{Y}, (x,z) \in X \times Z$$

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: $\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{Y}\phi|\mathsf{Z}\xi$ iff $\mathsf{Y} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{X}\phi|\mathsf{Z}\xi$

2. $\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{Y}C|\mathsf{Y}C$

3. Decomposition: $\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{Y}\phi|\mathsf{W}\xi$ and $\mathsf{Z} \preccurlyeq \mathsf{Y}$ implies $\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{Z}\phi|\mathsf{W}\xi$

4. Weak union:

   (a) $\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{Y}\phi|\mathsf{W}\xi$ and $\mathsf{Z} \preccurlyeq \mathsf{Y}$ implies $\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{Y}\phi|(\mathsf{Z},\mathsf{W})\xi$

   (b) $\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{Y}\phi|\mathsf{W}\xi$ and $\lambda \preccurlyeq \phi$ implies $\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{Y}\phi|(\mathsf{Z},\mathsf{W})(\xi,\lambda)$

5. Contraction: $\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{Z}\phi|\mathsf{W}\xi$ and $\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{Y}\phi|(\mathsf{Z},\mathsf{W})\xi$ implies $\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} (\mathsf{Y},\mathsf{Z})\phi|\mathsf{W}\xi$

The following forms of these properties are often used here:

1. Symmetry: $\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} (\mathsf{Y}\mathrm{id}_C)|\mathsf{Z}$ iff $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}} (\mathsf{X}\mathrm{id}_C)|\mathsf{Z}$

2. Decomposition: $\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} (\mathsf{Y},\mathsf{Z},\mathrm{id}_C)|\mathsf{W}$ implies $\mathsf{X} \perp\!\!\!\perp_{\mathbb{P}} \mathsf{Y}C|\mathsf{W}$ and $\mathsf{X} \perp\!\!\!\perp_{\mathbb{P}} (\mathsf{Z},\mathrm{id}_C)|\mathsf{W}$

3. Weak union: $\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} (\mathsf{Y},\mathsf{Z},\mathrm{id}_C)|\mathsf{W}$ implies $\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} (\mathsf{Y},\mathrm{id}_C)|(\mathsf{Z},\mathsf{W})$

4. Contraction: $\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} (\mathsf{Z},\mathrm{id}_C)|\mathsf{W}$ and $\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} (\mathsf{Y},\mathrm{id}_C)|(\mathsf{Z},\mathsf{W})$ implies $\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} (\mathsf{Y},\mathsf{Z},\mathrm{id}_C)|\mathsf{W}$

Conditional independence is sometimes given in terms of a factorisation of the joint conditional distribution (or joint conditional expectations). Theorem 2.3.23 shows that the definition given here is equivalent to the definition given in terms of factorisation.

Theorem 2.3.23 (Uniform conditional independence representation)**.** *Given a probability set* $\mathbb{P}_C$ *with a uniform conditional probability* $\mathbb{P}_C^{XY|Z}$,

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

$$
\Longleftrightarrow
$$

$$
\mathbb{P}_C^{XY|Z}(A \times B|z) \overset{\mathbb{P}_C}{\cong} \mathbb{P}_C^{X|Z}(A|z)\mathbb{P}_C^{Y|Z}(B|z) \qquad\qquad \forall A \in \mathcal{X}, B \in \mathcal{Y}, z \in Z
$$

if and only if $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e (X, \mathrm{id}_C)|Z$

Proof. If: By Theorem 2.4.10

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

$$
\overset{\mathbb{P}_C}{\cong}
\begin{array}{c}
Z \longrightarrow
\begin{cases}
\boxed{\mathbb{P}_C^{Y|Z}} - Y \\
\boxed{\mathbb{P}_C^{X|Z}} - X
\end{cases}
\end{array}
$$

$$
=
\begin{array}{c}
Z \longrightarrow
\begin{cases}
\boxed{\mathbb{P}_C^{Y|Z}} - Y \\
\boxed{\mathbb{P}_C^{X|Z}} - X
\end{cases}
\end{array}
$$

Only if: Suppose

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

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

$$
\mathbb{P}_C^{Y|XZ}(B|x,z) > \mathbb{P}_C^{Y|Z}(B|z) \qquad\qquad \forall (x,z) \in A \times C \qquad (2.6)
$$

then

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

a contradiction. An analogous argument follows if we replace ">" with "<" in Eq. (2.6). $\quad\square$

2.4 Maximal probability sets and valid conditionals

So far, we have been implicitly supposing that we first set up a non-empty probability set and from that set we may sometimes derive conditional probabilities, extended conditional independences and so forth. However, sometimes we want to work backwards: start with a collection of conditional probabilities, and work with the probability set implicitly defined by this collection. This is similar to the case in which we specify how some system works by specifying, based on a priori knowledge, a set of equations that govern it. A sanity check on proposing such a set of equations is to check whether they admit any solutions, and (possibly) to check how large the set of solutions they admit is.

Similarly, proposing a collection of condiitonal probabilities may similarly be compatible with a large set of base probability distributions, a unique probability distribution or no probability distributions at all. *Validity* is a sufficient (but not necessary) condition to ensure that certain collections of conditional probabilities yield a non-empty set of compatible distributions.

A particular example of specifying conditional probabilities a priori is given by Causal Bayesian Networks (see 5.1.12 for more details on this kind of model). The collection of operations of the form "$\mathbf{do}(\mathsf{X} = x)$" force the conditional distribution of X given its parents to a particular value. Specifically:

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

Given a causal Bayesian network with the graph $\mathsf{Y} \to \mathsf{X}$ and some observational probability distribution $\mathbb{P}^{\mathsf{XY}}_{\mathrm{obs}} \in \Delta(X \times Y)$, we conclude that $\mathbb{P}^{\mathsf{X}|\mathsf{Y}}_{\mathrm{do}_{\mathsf{X}},x} = \mathrm{del}_Y \delta_x$. Suppose, however, we had $\mathsf{X} = \mathsf{Y}$ – i.e. X and Y are actually two different labels for the same variable. Then there would be no distribution $\mathbb{Q}$ in $\Delta(X \times Y)$ with $\mathbb{Q}^{\mathsf{X}|\mathsf{Y}} = \mathrm{del}_Y \delta_x$.

The key result of this section 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 Vorob'ev (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.

In functional causal models, we have the notions of *global compatibility* from Forré and Mooij (2020) and *unique solvability* in Bongers et al. (2016). These are more general than validity; they are not just sufficient but also necessary for the existence of a solution. In addition, we note that the intervention operation discussed in Forré and Mooij (2020) preserves global compatibility, unlike the acyclic notion of intervention discussed above.

Definition 2.4.1 (Valid distribution). Given $(\Omega, \mathcal{F})$ and a variable $\mathsf{X} : \Omega \to X$, an X-valid probability distribution is any probability measure $\mathbb{K} \in \Delta(X)$ such that $\mathsf{X}^{-1}(A) = \varnothing \implies \mathbb{K}(A) = 0$ for all $A \in \mathcal{X}$.

Definition 2.4.2 (Valid conditional). Given $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ a $\mathsf{Y}|\mathsf{X}$-*valid conditional probability* is a Markov kernel $\mathbb{L} : X \nrightarrow Y$ that assigns probability 0 to impossible events, unless the argument itself corresponds to an impossible event:

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

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

Theorem 2.4.3 (Equivalence of validity definitions)**.** *Given* $\mathsf{X} : \Omega \to X$, *with* Ω *and* X *standard measurable, a probability measure* $\mathbb{P}^{\mathsf{X}} \in \Delta(X)$ *is valid if and only if the conditional* $\mathbb{P}^{\mathsf{X}|*} := * \mapsto \mathbb{P}^{\mathsf{X}}$ *is valid.*

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

$$\forall A \in \mathcal{X} : \mathsf{X} \bowtie A = \varnothing \implies \mathbb{P}^{\mathsf{X}|*}(A|*) = 0$$

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

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

$\square$

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

Definition 2.4.4 (Maximal probability set)**.** Given $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and a $\mathsf{Y}|\mathsf{X}$-valid conditional probability $\mathbb{L} : X \dashrightarrow Y$ 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 $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}}$.

Theorem 2.4.5 shows that the semidirect product of any pair of valid conditional probabilities is itself a valid conditional. Suppose we have some collection of $\mathsf{X}_i|\mathsf{X}_{[i-1]}$-valid conditionals $\{\mathbb{P}_i^{\mathsf{X}_i|\mathsf{X}_{[i-1]}} | i \in [n]\}$; then recursively taking the semidirect product $\mathbb{M} := \mathbb{P}_1^{\mathsf{X}_1} \odot (\mathbb{P}_2^{\mathsf{X}_2|\mathsf{X}_1} \odot ...)$ yields a $\mathsf{X}_{[n]}$ 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 Pearl (2009); Richardson and Robins (2013).

Note that validity is not a necessary condition for a conditional to define a non-empty probability set. Given some $\mathbb{K} : X \dashrightarrow Y$, $\mathbb{K}$ 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 2.4.7).

Theorem 2.4.5 (Semidirect product of valid conditional distributions is valid)**.** *Given* $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ *(all spaces standard measurable) and any valid candidate conditional* $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ *and* $\mathbb{Q}^{\mathsf{Z}|\mathsf{Y}\mathsf{X}}$, $\mathbb{P}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{Q}^{\mathsf{Z}|\mathsf{Y}\mathsf{X}}$ *is also a valid candidate conditional.*

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

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

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

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

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

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

$\square$

Corollary 2.4.6 (Valid conditionals are validly extendable to valid distributions)**.** *Given* Ω, $\mathsf{U} : \Omega \to U$, $\mathsf{W} : \Omega \to W$ *and a valid conditional* $\mathbb{T}^{\mathsf{W}|\mathsf{U}}$, *then for any valid conditional* $\mathbb{V}^{\mathsf{U}}$, $\mathbb{V}^{\mathsf{U}} \odot \mathbb{T}^{\mathsf{W}|\mathsf{U}}$ *is a valid probability.*

Proof. Applying Lemma 2.4.5 choosing $\mathsf{X} = *$, $\mathsf{Y} = \mathsf{U}$, $\mathsf{Z} = \mathsf{W}$ and $\mathbb{P}^{\mathsf{Y}|\mathsf{X}} = \mathbb{V}^{\mathsf{U}|*}$ and $\mathbb{Q}^{\mathsf{Z}|\mathsf{YX}} = \mathbb{T}^{\mathsf{W}|\mathsf{U}*}$ we have $\mathbb{R}^{WU|*} := \mathbb{V}^{\mathsf{U}|*} \odot \mathbb{T}^{\mathsf{W}|\mathsf{U}*}$ is a valid conditional probability. Then $\mathbb{R}^{\mathsf{WU}} \cong \mathbb{R}^{WU|*}$ is valid by Theorem 2.4.3. $\square$

Theorem 2.4.7 (Validity of conditional probabilities)**.** *Suppose we have* Ω, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, *with* Ω, X, Y *discrete. A conditional* $\mathbb{T}^{\mathsf{Y}|\mathsf{X}}$ *is valid if and only if for all valid distributions* $\mathbb{V}^{\mathsf{X}}$, $\mathbb{V}^{\mathsf{X}} \odot \mathbb{T}^{\mathsf{Y}|\mathsf{X}}$ *is also a valid distribution.*

Proof. If: this follows directly from Corollary 2.4.6.

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

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

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

Example 2.4.8. Body mass index is defined as a person's weight divided by the square of their height. Thus, given the random variables W, H modelling $\mathcal{W}, \mathcal{H}$, $\mathcal{B}$ is the random variable given by $\mathsf{B} = \frac{\mathsf{W}}{\mathsf{H}^2}$.

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

$$\mathbb{P}_\square^{\mathsf{B}|\mathsf{WHD}} = \begin{array}{c} \mathsf{H} \longrightarrow * \\ \mathsf{D} \underline{\hspace{2cm}} \mathsf{B} \\ \mathsf{W} \longrightarrow * \end{array} \tag{2.7}$$

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

$$(\mathsf{B}, \mathsf{W}, \mathsf{H}) \bowtie \{(x, w, h)\} = \{\omega | (\mathsf{W}, \mathsf{H})(\omega) = (w, h), \mathsf{B}(\omega) = \frac{w}{h^2}\}$$
$$= \varnothing$$

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

Validity rules out conditional probabilities like (2.7). We conjecture that in many cases this condition is implicitly taken into account. We note, however, that presuming the authors intended their model to be interpreted according to the usual semantics of causal Bayesian networks with hard interventions, the invalid conditional probability (2.7) appears in the structural model of the causal effect of body mass index found in Shahar (2009).

One question that arises is whether we can generally choose valid versions of the higher order conditionals whose existence is established in Theorem 2.3.15. This would potentially have applications to causal Bayesian network style reasoning – where we might want to start with some conditional probability, disintegrate it into higher order conditionals, keep one and replace the other under some intervention operation. In general, this is still an open question – Theorem 2.3.15 does not imply the resulting higher order conditional is valid – but we can show that if we limit our attention to discrete sets then higher order conditionals can be chosen to be valid.

Lemma 2.4.9. *Given a probability space $(\mu, \Omega, \mathcal{F})$ and variables $\mathsf{X}: \Omega \to X$, $\mathsf{Y}: \Omega \to Y$, if there is a regular proper conditional probability $\mu^{|\mathsf{X}}: X \rightarrowtail \Omega$ then there is a valid conditional distribution $\mu^{\mathsf{Y}|\mathsf{X}}$.*

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

Defining $\mathsf{O} := \mathrm{id}_{\Omega}$ (the identity function $\Omega \to \Omega$), $\mu^{|\mathsf{X}}$ is a version of $\mu^{\mathsf{O}|\mathsf{X}}$. Note also that $\mathsf{Y} = \mathsf{Y} \circ \mathsf{O}$. Thus by Lemma 2.3.12, $\mathbb{K}$ is a version of $\mu^{\mathsf{Y}|\mathsf{X}}$.
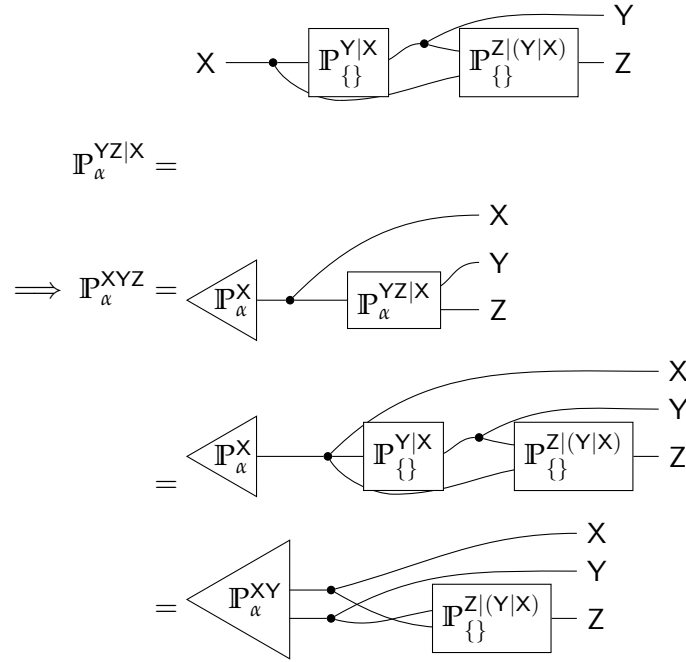
It remains to be shown that $\mathbb{K}$ is valid. Consider some $x \in X$, $A \in \mathcal{Y}$ such that $\mathsf{X}^{-1}(\{x\}) \cap \mathsf{Y}^{-1}(A) = \varnothing$. Then by the assumption $\mu^{|\mathsf{X}}$ is proper

$$\mathbb{K}(\mathsf{Y} \bowtie A | x) = \delta_x(\mathsf{Y}^{-1}(A))$$
$$= 0$$

Thus $\mathbb{K}$ is valid. $\qquad \square$

Theorem 2.4.10 (Higher order conditionals). *Suppose we have a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X}: \Omega \to X$ and $\mathsf{Y}: \Omega \to Y$, $\mathsf{Z}: \Omega \to Z$ and a probability set $\mathbb{P}_C$ with conditional $\mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}$. Then $\mathbb{P}_C^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ is a version of $\mathbb{P}_C^{\mathsf{Z}|\mathsf{YX}}$*

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

$$\mathbb{P}_\alpha^{\mathsf{YZ|X}} =$$

$$\implies \mathbb{P}_\alpha^{\mathsf{XYZ}} =$$

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

Theorem 2.4.11 (Valid higher order conditionals). *Suppose we have a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ and a probability set $\mathbb{P}_C$ with valid regular conditional $\mathbb{P}_C^{\mathsf{YZ|X}}$, Y discrete and Z standard measurable. Then there exists a valid regular $\mathbb{P}_C^{\mathsf{Z|XY}}$.*

Proof. By Theorem 2.3.15, we have a higher order conditional $\mathbb{P}_C^{\mathsf{Z|(Y|X)}}$ which, by Theorem 2.4.10 is also a version of $\mathbb{P}_C^{\mathsf{Z|XY}}$.

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

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

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

For all x, y such that $\mathbb{P}_C^{\mathsf{Y|X}}(\{y\}|x)$ is positive, we have

$$\mathbb{P}_C^{\mathsf{YZ|X}}(\{y\} \times A|x) = 0 \implies \mathbb{P}_C^{\mathsf{Z|XY}}(A|x,y) = 0 =: \mathbb{Q}(A|x,y)$$

Furthermore, where $\mathbb{P}_C^{\mathsf{Y|X}}(\{y\}|x) = 0$, we either have $(\mathsf{X}, \mathsf{Y}, \mathsf{Z}) \bowtie \{(x,y)\} \times A = \emptyset$ or can choose some $\omega \in (\mathsf{X}, \mathsf{Y}, \mathsf{Z}) \bowtie \{(x,y)\} \times A$ and let $\mathbb{Q}(\mathsf{Z}(\omega)|x,y) = 1$. 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 $\mathbb{P}_C^{Y|X}(\{y\}|x) = 0$ is a measure zero set, and so $\mathbb{Q}$ differs from $\mathbb{P}_C^{Y|X}$ on a measure zero set. $\qquad\square$

2.5 Interpretation of probabilistic decision models

In this thesis, we will use the notions of probability distributions, random variables and sample spaces that are standard to probability theory. Beyond these, decision models also feature *option sets*, which are not standard elements of the theory of probability and necessitate the theory of probability sets worked out above. With regard to how decision models are interpreted, we will consider them to be an aid to a decision maker who wants to predict the consequences of different choices they might make and have little else to say about them in this thesis. However, in this section we will explore in more detail the question of interpretating variables and options. On the question of interpreting options, our approach differs in some respects from the interventional and potential outcomes approaches.

In statistics, variables aren't *just* measurable functions defined on the sample space (Definition 2.1.13). Typically, they're also 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 "conditional measurement procedure", where the decision maker, on choosing an option $\alpha \in C$, executes a measurement $\mathcal{S}_\alpha$. $\mathcal{S}_\alpha$ is considered to be a measurement procedure that measures all quantities of interest, and particular quantities of interest can be obtained by composing $\mathcal{S}_\alpha$ with an appropriate function. The function X that, when applied to the result of $\mathcal{S}$ is the variable associated with this particular quantity of interest.

2.5.1 Random 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, 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.

Thus, in order to fully make sense of the equation $\mathsf{F} = \mathsf{MA}$, it must be understood relative to some measurement procedure $\mathcal{S}$ that simultaneously measures the force on an object, its mass and its acceleration. If the procedure yields a triple ("force", "mass", "acceleration"), then the required quantities can be recovered by composing functions with the procedure's result. For example, "force" can be recovered by applying the function $\mathsf{F} : (a, b, c) \mapsto a$, and one can similarly define M and A. The equation then says that, whatever result s this procedure yields, $\mathsf{F}(s) = \mathsf{M}(s)\mathsf{A}(s)$ will hold.

One could also consider imposing some coherence requirements on $\mathcal{S}$. For example, perhaps we require that $\mathsf{F} \circ \mathcal{S}$ gives the same result as some procedure $\mathcal{F}$ that measures only the net force on the given object, or that different people can execute $\mathcal{S}$ and obtain the same result. However, details like this are beyond the scope of this work. The one requirement we do place on $\mathcal{S}$ is that it is sure to return values in some set $F \times M \times A$ that is known in advance. No actual procedure can be guaranteed to return elements of a mathematical set known in advance – anything can fail – but for our purposes we assume that we can study procedures reliable enough that we don't lose much by ignoring this possibility.

A measurement procedure $\mathcal{S}$ 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. $\mathcal{S}$ itself is not a well-defined mathematical thing, but the set of values it may yield *is* a well-defined mathematical set. This is what facilitates the idea of "composing a function with $\mathcal{S}$". Because $\mathcal{S}$ is not a purely mathematical thing, we avoid attempting to reason mathematically about $\mathcal{S}$ beyond the requirement that we can compose functions with it. As a result, $\mathcal{S}$ is required to interpret the mathematical models, but mainly stays in the background while the actual details of the reasoning concern the functions F, M and A (or whatever variables the problem actually gives us).

2.5.2 Defining measurement procedures

Motivated by the example above, we define a measurement procedure as something an individual can "do" which leaves them, in the end, with an element of a mathematical set.

Definition 2.5.1 (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 2.5.2 (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 $\vee_{x \in A} \mathcal{B} \bowtie x$.

Definition 2.5.3 (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 without measuring anything else. However, if we suppose that both yield the same result in the end we can treat them as equivalent.

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

Definition 2.5.4 (Composition of functions with procedures)**.** Given a procedure $\mathcal{B}$ that takes values in some set B, and a function $f : B \to C$, 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:

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

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, and in the spirit of not analysing procedures too deeply we avoid defining any such notion.

Our approach here is to suppose that there is only one measurement procedure $\mathcal{S}$ that yields all the information required to determine the values of all observed variables. Observed variables are, precisely, functions defined on the observable sample space $(\Psi, \mathcal{E})$. Thus we never need to combine real world actions – we assume that they are all taken care of by $\mathcal{S}$.

Given that measurement processes are in practice finite precision and with finite range, Ψ will generally be a finite set. We could in this case equip Ψ with the collection of measurable sets given by the power set $\mathcal{E} := \mathcal{P}(\Psi)$, and $(\Psi, \mathcal{E})$ is a standard measurable space. More generally, we assume that there is a set $\mathcal{E}$ of observable events such that $(\Psi, \mathcal{E})$ is standard measurable, though beyond the suggestion that Ψ is likely to be finite in practice, we don't know if there are reasonable measurement procedures that do not conform to this requirement.

2.5.3 Observable variables

The measurement procedure $\mathcal{S}$ represents a large collection of quantities of interest, each of which can be obtained by composition of some function with $\mathcal{S}$. Given some measurable function $\mathsf{X} : (\Psi, \mathcal{E}) \to (X, \mathcal{X})$, we call the pair $(\mathsf{X}, \mathcal{S})$ an *observable variable.* Typically, however, we omit the mention of $\mathcal{S}$ for brevity. Unlike the definition of a variable (Definition 2.1.13), an *observable variable* is the function along with the procedure required to determine which "bit of the world" the function is referring to.

Definition 2.5.5 (Observable variable)**.** Given a measurement procedure $\mathcal{S}$ taking values in $(\Psi, \mathcal{E})$, an observable variable is a pair $(\mathsf{X}, \mathcal{S})$ where $\mathsf{X} : (\Psi, \mathcal{E}) \to (X, \mathcal{X})$ is a measurable function.

For the model $\mathsf{F} = \mathsf{MA}$, for example, suppose we have a 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 (S, F) where $\mathcal{F} := \mathsf{F} \circ \mathsf{S}$ and $\mathsf{F} : X \times Y \times Z \to X$ 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 "$\mathsf{X} \circ \mathsf{S}$ yields x" an *observation*. The proposition "X yields x" is equivalent to the proposition "S yields a value in $\mathsf{X}^{-1}(x)$". Because of this, we define the *event* $\mathsf{X} \bowtie x$ to be the set $\mathsf{X}^{-1}(x)$.

Definition 2.5.6 (Event)**.** Given the complete procedure S taking values in Ψ and an observable variable $(\mathsf{X} \circ \mathsf{S}, \mathsf{X})$ for $\mathsf{X} : \Psi \to X$, the *event* $\mathsf{X} \bowtie x$ is the set $\mathsf{X}^{-1}(x)$ for any $x \in X$.

If we are given an observation "$\mathsf{X} \circ \mathsf{S}$ yields x", then the corresponding event $\mathsf{X} \bowtie x$ is the set of measurement results 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 $\mathsf{Y} = y$ already has a meaning, namely that Y is a constant function everywhere equal to y.

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

2.5.4 Unobservable variables

Statistical models often include unobservable variables. Formally, we can define a sample space $(\Omega, \mathcal{F})$ and a function $\mathsf{S} : \Omega \to \Psi$ which tells us, for any observed result of S, which values of the larger sample space Ω are consistent.

A general variable is then a measurable function on Ω, and an unobservable variable is a general variable X such that $\mathsf{X} \not\preceq \mathsf{S}$ (Definition 2.1.15). However, unobservable variables come with a warning – we assume there is some measurement procedure S that tells us how to interpret all of the observable variables, but we do not assume any such object that facilitates an interpretation of unobservable variables.

Observable variables are special in the sense that they are tied to a particular measurement procedure S. However, the measurement procedure 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.

Definition 2.5.7 (Unobservable variable)**.** Given a sample space $(\Omega, \mathcal{F})$, a measurement procedure S yielding values in Ψ and $\mathsf{S} : \Omega \to \Psi$ that determines the outcomes compatible with measurement results, a random variable $\mathsf{Y} : \Omega \to Y$ is *unobservable* if $\mathsf{Y} \not\preceq \mathsf{S}$.

2.5.5 Decision procedures

The kind of problem we want to solve requires us to compare the consequences of different choices from a set of options C. We take the *consequences of* $\alpha \in \mathsf{C}$ to refer to the values obtained by some measurement procedure S_α associated with the choice α, and assume that we have in hand a *conditional measurement procedure* S_C which is the procedure given by "do S_α if you choose α".

We could instead contemplate an unconditional measurement procedure $\mathcal{T}$ of the form:

1. Choose id_C

2. Proceed according to $\mathcal{S}_\alpha$

One of the reasons we might want to do this is that we already have something that looks like a procedure for part 1: first, construct a model of the consequences of each option, and then pick the best option according to the model. However, this only requires a model of the consequences of each action – i.e. a model of part 2 – not a model of part 1. Furthermore, modelling part 1 is difficult. Suppose we construct a model $\mathbf{Q}$ of parts 1 and 2, and as a result we decide on option α. Then, seemingly, we can construct a better model $\mathbf{Q}'$ which assigns certainty to the outcome $\mathsf{id}_C \bowtie \alpha$. But *then* the consequences of any $\alpha' \neq \alpha$ seem to be moot, as we are at this point certain that α will be chosen.

There may well be reasonable solutions to the problem of modelling the full procedure $\mathcal{T}$ that also yields a useful model of the consequences of every available option, but we only need to model part 2 in order to execute the procedure for picking an option. Exploring and resolving any difficulties related to modelling the part 1 are beyond the scope of this thesis.

We can therefore consider a decision procedure to be a collection of subprocedures $\mathcal{S}_\alpha$ for each option α, and these correspond to probability measures $\mathbb{P}_\alpha$ for each option α (see Definition 2.3.1). Our analysis isn't maximally general here – we could imagine some decision problems where different options in general lead to different sample spaces. Exploring this variation is also beyond the scope of this thesis.

Definition 2.5.8 (Decision procedure)**.** A decision procedure is a collection $\mathcal{S}_C := \{\mathcal{S}_\alpha\}_{\alpha \in C}$ of measurement procedures. As in Definition 2.3.1, we call C the *option set.*

2.5.6 Interpretation of potential outcomes and interventions

Given an option set C, we could consider a vector of potential outcomes Y^C where Y^α is interpreted as "the value that Y would take had I chosen α". Because Y^α is defined whether or not I choose α, it's clear that the vector Y^C must be an unobservable variable. Thus it comes with the warning we mentioned – we don't necessarily assume that there is any obvious way to interpret this variable.

On the other hand, the assumed decision procedure $\mathcal{S}_C$ does provide a means of interpreting the decision model $(\mathbb{P}_., \Omega, C)$. In practice, someone might analyse a decision problem with a decision procedure that is too vague, or that fails some consistency condition (questions which we don't analyse in any detail here). However, we do, at least, point to decision procedures as the kind of thing that facilitates the interpretation of a decision model.

Whether our interpretation applies to interventional models depends on what kinds of interventions the model contemplates and on just how vague decision procedures are allowed to be before we disqualify them. On the one hand, an interventional model that allows only interventions on a choice variable id_C will induce a collection of probability distributions for each value of the option set C, and so would be suitable for modelling a decision procedure with this option set. It is not clear that there is a meaningful distinction in this kind of model between id_C taking a particular value and intervening on id_C "to set it to a particular value". On the other hand, whether an interventional model that features hard interventions on arbitrary variables can be interpreted as a model of a decision procedure depends on whether we accept that a statement like "have Nature set an individual's body mass index to 11" is enough to define a measurement procedure $\mathcal{S}_\alpha$ (Pearl, 2018). Such a statement seems too vague to qualify to us, but we don't analyse this question in detail and the subsequent work doesn't depend on a strong commitment either way.

Chapter 3

Models with choices and consequences

Decision models are used to model *decision problems*. 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. The role of the decision model is to associate each option with a prediction of the consequences of choosing this option. We have already suggested that the type of a decision model is a map from the option set to probability distributions over the consequence set. In this chapter, we examine the literature on decision theory, and find that it offers essentially the same prescription for the type of a decision model.

In practice, a lot of empirical causal analysis is concerned with problems a step removed from choosing among options. Often, the purpose of causal analysis is to support other decision makers deliberating on a course of action, rather than to recommend an action directly. These are still problems involving a choice among options, but the procedure by which the choice is made is somewhat opaque. We consider problems where the analyst makes the choice because it is an important class of problems in its own right, and may serve as a useful idealisation of more opaque decision problems. In Section 3.1, we point out that many works in the causal inference literature regard problems of decision making or control as a particularly important class of problem. While some authors have held that counterfactual questions are a more general class of causal problem, the extra assumptions needed for counterfactual models are not required by decision makers, and we regard the extra complications of counterfactual models as mostly beyond the scope of this thesis.

In Section 3.2, we discuss in more detail what we take to be a prototypical decision problem and the need for some kind of "relation" (broadly understood) between options and consequences. This relation must be able to represent uncertainty somehow, and with this in mind we offer two elaborations on the definition of decision models given in the previous chapter. The first of these may be loosely considered a "Bayesian" decision model and the second a "non-Bayesian" decision model.

The structure of the decision models we propose relies on some apparently arbitrary choices – for example, the choice to use probability to represent uncertainty. The literature on decision theory has offered a number of axiomatisations of "rational choice" or "coherent preferences" that aim to put such choices on a clearer footing. Section 3.3 provides an overview of four major decision theories along along with (where applicable) their axiomatisations. These are *Savage decision theory*, *Jeffrey decision theory* (or evidential decision theory), Lewis' *causal decision theory* and *statistical decision theory*. While there are significant controversies surrounding the question of how decision problems should be modeled, there is substantial

(though not perfect) agreement between the different theories on the type of model a decision maker uses to evaluate their options, and this type is typically a Bayesian decision model.

Section 3.3 explores in particular detail the connections between *statistical decision theory* (Wald, 1950) and decision models. We demonstrate a close connection between a certain family of decision models and the classical statistical notion of the *risk* of a decision rule. This family of decision models is defined by a particular conditional independence structure which we find is shared by the kinds decision models that we construct to represent potential outcomes and causal Bayesian networks in Chapter 5. That is, we show that the conditional independence structure of the decision models underlying classical statistical decision theory is the same as the structure of the decision models underlying modern approaches to causal inference.

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 we only focus 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) introduce causal inference as the study of "outcomes of manipulations" and (Spirtes et al., 2000) highlight 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. Secondly, 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. Thirdly, 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.

3.2 Modelling decision problems

People who need to make decisions might (and often do) make them with no mathematical reasoning at all. However, this work is concerned with making decisions supported by mathematical reasoning, and this requires a mathematical representation of the decision making problem. We suppose that a decision maker finds themselves in the following kind of situation:

1. They are contemplating a collection of different options and must choose one of them (which may include "do nothing")

2. They know what could possibly happen in the future, and prefer some of these possibilities to others

Such a decision maker could choose to mathematically represent their set of options with an option set C and the set of possible things that could happen in the future with Ω. They could also choose to represent their preferences over future possibilities with a scoring or utility function u. This means that their preferences must form a total order on the set of future possibilities, and we accept this assumption.

The set Ω here does not necessarily correspond to the possible results of a measurement procedure Ψ, as we discussed in Section 2.5.2, which reflects the fact that decision makers might have preferences about the values of things that are never measured. However, in this thesis we only consider the simpler family of problems where the consequences of interest are all observed.

This decision maker faces the problem that they can only choose from among their options, but they only have the ability to evaluate different future possibilities. What they need is some means of relating the options C and the possibilities Ω, which will allow them to consider the different future possibilities brought about by choices among their options, and thereby to also determine which of their options are preferred. In general, they won't know

exactly which future will be brought about by any of their choices, and so whatever method they use to relate C to Ω must allow them to represent uncertainty about this relationship.

We can consider two kinds of model (though these are far from exhaustive!):

- The decision maker considers the consequences of each option to be uncertain, and uses probability alone to represent this uncertainty; their model is a Markov kernel $C \twoheadrightarrow \Omega$

- The decision maker considers the consequences of each option to be probabilistically uncertain, and furthermore considers the the appropriate Markov kernel to lie in some set H, the correct choice being non-probabilistically uncertain; their model is a map $C \times H \twoheadrightarrow \Omega$

The first kind of model enables the decision maker to use the principle of expected utility to induce a total order on their set of options, while the second kind of model in conjunction with expected utility induces only a preorder on the decision maker's option set, and some further decision rule is generally needed to select a "best" option given this structure.

A decision maker might also contemplate a decision model with no probabilistic uncertainty at all, such as a binary relation between C and Ω, although they might be substantially hamstrung by neglecting any means of representing graduations of uncertainty. Alternatively, they might consider a model that associates each option with an *imprecise probability* over the consequences (Walley, 1991). Whether or not this is equivalent to the second kind of model is an open question.

There have been a number of attempts to show that a rational reasoner or decision maker *must* use probability to represent uncertain knowledge. For example, Finetti ([1937] 1992); Horvitz et al. (1986) propose a number of principles they claim coherent reasoning under uncertainty must follow, and demonstrate that a reasoner who follows these principles must be able to represent their uncertainty with a probability distribution. These principles have, in turn, been criticised (Halpern, 1999). Our question – how to represent decision models – has been more directly addressed in the literature on decision theory which is surveyed in Section 3.3.

While we believe the project of axiomatising model choices is valuable because it helps decision makers to understand what commitments they are making when they adopt a certain type of model to aid their deliberation, we think it is likely to be very difficult to sustain an argument that a decision maker *must* commit to any given set of principles and do not make such an argument here.

3.2.1 Formal definitions

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 $u : \Omega \to \mathbb{R}$. We call these ingredients a "decision problem".

Definition 3.2.1 (Decision problem). A decision problem is a triple $((C,\mathcal{C}), (\Omega,\mathcal{F}), u)$ consisting of a measurable set $(C,\mathcal{C})$ of choices, $(\Omega,\mathcal{F})$ consequences and a measurable utility function $u : \Omega \to \mathbb{R}$.

Our task is to find a *model* that relates choice C to consequences Ω. We assume two forms of model – a "Bayesian" model, which associates each choice with a unique probability distribution, and a "non-Bayesian" model that consists of a set of Bayesian models.

Definition 3.2.2 (Bayesian decision model). Given a decision problem $((C,\mathcal{C}), (\Omega,\mathcal{F}), u)$, a *Bayesian decision model* is a triple $(\mathbb{P}_., (\Omega,\mathcal{F}), (C,\mathcal{C}))$ where $\mathbb{P}_. : C \twoheadrightarrow \Omega$ is a Markov kernel.

Definition 3.2.3 (non-Bayesian decision model)**.** Given a decision problem $((C, \mathcal{C}), (\Omega, \mathcal{F}), u)$, a *non-Bayesian decision model* is a triple $(\mathbb{P}_\cdot, (\Omega, \mathcal{F}), (C \times H, \mathcal{C} \otimes \mathcal{H}))$ where H is a set of *hypotheses* and $\mathbb{P}_\cdot : C \twoheadrightarrow \Omega$ is a Markov kernel.

By convention, we use $\mathbb{P}_\cdot$ with the subscript $\cdot$ to denote the map from options to consequences, subscripts $\mathbb{P}_\alpha$ refer to the model evaluated at $\alpha \in C$ (or in $C \times H$) and the subscript $\mathbb{P}_A$ refers to the probability set formed by the image of $A \subset C$ of A under the model.

3.3 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 3.2.3, see also Chapter 4)

- Most theories have some notion of a "prior" over hypotheses, which induces a choice only model (Definition 3.2.2)

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 3.3.23) 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 decision models we call conditionally independent see-do models (Definition 3.3.15) combined with the principle of expected utility maximisation. With this in hand, we show that the complete class theorem applies to this class of decision models, and extend it to a slightly broader class.

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 $a \sim b$.

3.3.1 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 $(\Omega, \mathcal{F})$) 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 $\mathbb{P}_\alpha$ and $\mathbb{P}_{\alpha'}$, there exists some $u : \Omega \to \mathbb{R}$ 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 is the set of all probability measures on $(\Omega, \mathcal{F})$. 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.

3.3.2 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 $u : \Omega \to \mathbb{R}$ in combination with the principle of expected utility. As Theorem 3.3.3 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. 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, which is usually much bigger than the set of all the things the decision maker might actually choose to do. This could be considered a requirement of extendability: given a non-Bayesian model $(\mathbb{P}_{C \times H}, \Omega, C \times H)$, we might consider it a Savage decision model if there is some $(\mathbb{Q}_{C' \times H}, \Omega, C' \times H)$ where C' is defined as the convex closure of the set of all deterministic functions $H \to \Omega$ and $\alpha \in C$ implies $\mathbb{Q}_\alpha = \mathbb{P}_\alpha$ and the Savage axioms hold (see Appendix A.1).

The reason why Savage's theory has such a rich set of options is because the derivation needs to deduce a preference relation over consequences from the preference relation over the options (in contrast, we simply assume preferences over consequences are available at the outset). All a decision maker actually needs is the ordering on the options that they're actually considering, and this might be compatible with many orderings of consequences. Sufficiently enriching the set of options can restrict this to a unique relation. We don't

know if there are cases of decision problems where this extendability requirement introduces difficulties.

Definition 3.3.1 (Elements of a Savage decision problem)**.** A *Savage decision problem* features a measurable set of states $(S, \mathcal{S})$, a set of consequences $(\Omega, \mathcal{F})$ and a set of acts C such that $|C| = \Omega^S$ and a measurable evaluation function $T : S \times C \to \Omega$ such that for any $f : S \to \Omega$ there exists $c \in C$ such that $T(\cdot, c) = f$.

Theorem 3.3.2 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 A.1.

Theorem 3.3.2. *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* $\mu \in \Delta(\mathcal{S})$ *and a utility* $u : \Omega \to \mathbb{R}$ *unique up to affine transformation such that*

$$\alpha \preceq \alpha' \iff \int_S u(T(s, \alpha))\mu(\mathrm{d}s) \leq \int_S u(T(s, \alpha'))\mu(\mathrm{d}s) \qquad \forall \alpha, \alpha' \in C$$

Proof. Savage (1954) □

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

Theorem 3.3.3. *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* $\mathbb{P}_. : C \to \Delta(\Omega)$ *and a utility* $u : \Omega \to \mathbb{R}$ *unique up to affine transformation such that*

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

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

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

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

Then

$$\int_\Omega u(f)\mathbb{P}_\alpha(\mathrm{d}f) = \int_S u \circ T_\alpha(s)\mu(\mathrm{d}s)$$
$$= \int_S u(T(s, \alpha))\mu(\mathrm{d}s)$$

□

3.3.3 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 $\underline{\mathcal{F}}$, which is a complete atomless boolean algebra. Elements of $\underline{\mathcal{F}}$ are said to be propositions. We note that $\underline{\mathcal{F}}$ cannot be understood as the set of events with respect to a finite measurement procedure (Section 2.5). 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 Savages'; where Savage identifies a set of acts C with all functions $S \to F$ and proposes axioms that constrain a preference relation on C, Jeffrey assumes that choices are elements of the algebra $\underline{\mathcal{F}}$, accompanied by other propositions that do not correspond precisely to choices. Jeffrey's axioms pertain to a preference relation on $\underline{\mathcal{F}}$, 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 $\varnothing$ and 1 with $\mathbb{R}$, but it has atoms: any singleton $\{x\}$ has only the subsets $\varnothing$ 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 3.3.4 (Complete atomless boolean algebra). A complete atomless boolean algebra $\underline{\mathcal{F}}$ is a tuple $(A, \wedge, \vee, \neg, 0, 1)$ 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 A.2).

Theorem 3.3.5. *Suppose there is a complete atomless Boolean algebra $\underline{\mathcal{F}}$ with a preference relation $\preceq$. If $\preceq$ satisfies the* Bolker axioms *then there exists a desirability function des :* $\underline{\mathcal{F}} \to \mathbb{R}$ *and a probability distribution $\mu \in \Delta(\underline{\mathcal{F}})$ such that for $A, B \in \underline{\mathcal{F}}$ and finite partition $D_1, ..., D_n \in \underline{\mathcal{F}}$:*

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

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

Proof. Bolker (1966)) $\qquad \square$

As mentioned, in Jeffrey's theory the *choices* under consideration C are assumed to be some subset of $\underline{\mathcal{F}}$. Thus we can deduce from a Jeffrey model a function $C \to \Delta(\underline{\mathcal{F}})$ that "represents the consequences of choices" in the sense of Theorem 3.3.6.

Theorem 3.3.6. *Suppose there is a complete atomless Boolean algebra $\underline{\mathcal{F}}$ 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 $\mathbb{P}. : C \to \Delta(\underline{\mathcal{F}})$ such that for any $\alpha, \alpha' \in C$ and finite partition $D_1, ..., D_n \in \underline{\mathcal{F}}$:*

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

Where μ and des *are as in Theorem 3.3.5*

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

3.3.4 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 3.3.7 (Causal decision rule). Given a set C of choices, sample space $(\Omega, \mathcal{F})$, variables $\mathsf{H} : \Omega \to H$ (the *dependency hypothesis*) and $\mathsf{S} : \Omega \to S$ (the *consequence*) and a utility $u : \Omega \to \mathbb{R}$, the *causal utility* of a choice $\alpha \in C$ is given by

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

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

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 $\mathbb{P}. : C \to \Delta(\Omega)$. 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 an important idea in causal reasoning. Lewis' decision rule connects the theory of probability sets with *statistical decision theory*, as Section 3.3.5 will show. Chapter 4 goes into considerable detail concerning the question of when probability sets support certain types of dependency hypotheses. 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.

3.3.5 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 3.3.8 (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, $l : D \times H \to \mathbb{R}$ is a loss function and $\mathbb{P}. : H \rightarrow X$ is a Markov kernel from hypotheses to 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 3.3.9 (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 reuquires a prior over hypotheses, while minimax optimality does not.

Definition 3.3.10 (SDP Risk)**.** Given a statistical decision problem $(X, H, D, l, \mathbb{P}.)$ and decision functions C, the *risk* functional $R : C \times H \to \mathbb{R}$ 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 in other manners. Thus Bayes optimality and minimax optimality are defined in terms of risk functions in general, not SDP risk functions.

Definition 3.3.11 (Bayes risk)**.** Given decision functions C, hypotheses $(H, \mathcal{H})$, risk $R : C \times H \to \mathbb{R}$ 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 3.3.12 (Bayes optimal)**.** Given decision functions C, hypotheses $(H, \mathcal{H})$, risk $R : C \times H \to \mathbb{R}$ 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 3.3.13 (Minimax optimal)**.** Given decision functions C, hypotheses $(H, \mathcal{H})$, risk $R : C \times H \to \mathbb{R}$, a *minimax decision function* is any decision function $\mathbb{D}_\alpha$ 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 3.3.14 (See-do model)**.** A probability set model of a statistical decision problem, or a *see-do model* for short, is a tuple $(\mathbb{P}_{C \times H}, \mathsf{X}, \mathsf{Y}, \mathsf{D})$ where $\mathbb{P}_{C \times H}$ is a probability set indexed by elements of $C \times H$ on $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$ are the observations, $\mathsf{Y} : \Omega \to Y$ are the consequences and $\mathsf{D} : \Omega \to D$ are the decisions. $\mathbb{P}_{C \times H}$ must observe the following conditional independences:

$$\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}_{C \times H}} \mathrm{id}_C | \mathsf{H}$$
$$\mathsf{D} \perp\!\!\!\perp^e_{\mathbb{P}_{C \times H}} \mathsf{H} | \mathrm{id}_C$$

where $\mathrm{id}_C : C \times H \to C$ and $\mathsf{H} : C \times H \to H$ are the respective projections (refer to Definition 2.3.21 for the definition of extended conditional independence $\perp\!\!\!\perp^e_{\mathbb{P}_{C \times H}}$).

Definition 3.3.15 (Conditionally independent see-do model)**.** A conditionally independent see-do model is a see do model $(\mathbb{P}_{C \times H}, \mathsf{X}, \mathsf{Y}, \mathsf{D})$ where the following additional conditional independence holds:

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

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

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

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

A see-do model induces a set of decision functions: for each $\alpha \in C$, there is an associated probability distribution $\mathbb{P}^{\mathsf{D}|\mathsf{X}}_\alpha$. 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 3.3.17 (Induced SDP risk)**.** *Given a conditionally independent see-do model* $(\mathbb{P}_{C \times H}, \mathsf{X}, \mathsf{Y}, \mathsf{D})$ *along with a utility* $u : Y \to \mathbb{R}$, *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* $\mathbb{P}^{\mathsf{D}|\mathsf{X}}_\alpha$ *and hypothesis* h.

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

Proof. The expected utility given α and h is

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

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

Theorem 3.3.17 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 3.3.18) provides a notion of risk for the more general case, while Theorem 3.3.17 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 3.3.18 (See-do risk)**.** Given a see-do model $(\mathbb{P}_{C \times H}, \mathsf{X}, \mathsf{Y}, \mathsf{D})$ along with a utility $u : X \times Y \to \mathbb{R}$, the *see-do risk* $R : C \times H \to \mathbb{R}$ is given by

$$R(\alpha, h) := -\mathbb{P}^{\mathsf{XY}}_{\alpha,h} u \qquad\qquad \forall \alpha \in C, h \in H$$

Section 3.2 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 $(\mathbb{P}_{C \times H}, X, Y, D)$ 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 3.3.19 (Bayesian see-do model). A Bayesian see-do model is a tuple $(\mathbb{P}_C, X, Y, D, H)$ where $\mathbb{P}_C$ is a probability set on $(\Omega, \mathcal{F})$, $X : \Omega \to X$ are the observations, $Y : \Omega \to Y$ are the consequences, $D : \Omega \to D$ are the decisions and $H : \Omega \to H$ is the hypothesis. $\mathbb{P}_C$ must observe the following conditional independences:

$$X \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathrm{id}_C | H$$
$$D \perp\!\!\!\perp^e_{\mathbb{P}_C} H | \mathrm{id}_C$$
$$H \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathrm{id}_C$$

Definition 3.3.20 (Induced Bayesian see-do model). Given a see-do model $(\mathbb{P}_{C \times H}, X, Y, D, H)$ on $(\Omega, \mathcal{F})$ 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

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

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

Theorem 3.3.21 (Induced SDP Bayes risk). *Given a conditionally independent see-do model $(\mathbb{P}_C, X, Y, D, H)$ along with a utility $u : Y \to \mathbb{R}$ 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 $h \mapsto \mathbb{P}^{Y|XD}_{C \times \{h\}}$ is a version of $\mathbb{P}^{Y|XD}_C$ and hence $Y \perp\!\!\!\perp^e_{\mathbb{P}_C} (X, \mathrm{id}_C) | (H, D)$, a property it inherits from the underlying see-do model.

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

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

$$\begin{aligned}
\mathbb{P}^Y_\alpha u &= \int_Y u(y) \mathbb{P}^Y_\alpha(\mathrm{d}y) \\
&= \int_Y \int_D \int_X \int_H u(y) \mathbb{P}^{Y|DXH}_\alpha(\mathrm{d}y|d,x,h) \mathbb{P}^{D|XH}_\alpha(\mathrm{d}d|x,h) \mathbb{P}^{X|H}_\alpha(\mathrm{d}x|h) \mathbb{P}^H_\alpha(\mathrm{d}h) \\
&= \int_X \int_D \int_Y \int_H u(y) \mathbb{P}^{Y|DH}_\alpha(\mathrm{d}y|d,h) \mathbb{P}^{D|X}_\alpha(\mathrm{d}d|x) \mathbb{P}^{X|H}_\alpha(\mathrm{d}x|h) \mathbb{P}^H_\alpha(\mathrm{d}h) \\
&= \int_X \int_D \int_Y \int_H u(y) \mathbb{P}^{Y|DH}_C(\mathrm{d}y|d,h) \mathbb{P}^{D|X}_\alpha(\mathrm{d}d|x) \mathbb{P}^{X|H}_C(\mathrm{d}x|h) \mu(\mathrm{d}h) \\
&= -\int_D \int_X \int_H l(d,h) \mathbb{P}^{D|X}_\alpha(\mathrm{d}d|x) \mathbb{P}^{X|H}_C(\mathrm{d}x|h) \mu(\mathrm{d}h) \\
&= -\int_H R(\mathbb{P}^{D|X}_\alpha, h) \mu(\mathrm{d}h) \\
&= -R_\mu(\mathbb{P}^{D|X}_\alpha)
\end{aligned}$$

$\square$

Complete class theorem

The *complete class theorem* is a key theorem of classical SDT that establishes, under certain conditions, any *admissible* decision rule (Definition 3.3.23) 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 3.3.17 that conditionally independent see-do models induce statistical decision problems. However, the complete class theorem iteself (Theorem 3.3.24) 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 3.3.31 and 3.3.32.

Definition 3.3.22 (Risk function). Given a set of choice C and a set of hypotheses H, a risk function is a map $R : H \times C \to \mathbb{R}$.

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 3.3.23 (Admissible choice). Given a risk function $R : C \times H \to \mathbb{R}$, a choice r $\alpha \in C$ 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 3.3.24 (Complete class). A *complete class* is any $B \subset C$ such that, for any α' inB 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 3.3.25. *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 3.3.26 (Risk set). Given a finite set of hypotheses H, a set of choices C and a risk function $R : C \times H \to \mathbb{R}$, the risk set is the subset of $\mathbb{R}^{|H|}$ given by

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

Theorem 3.3.27 (Complete class theorem). *Given a risk function $R : C \times H \to \mathbb{R}$, 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 3.3.31 and 3.3.32). In order to explain them, we need a few lemmas.

Lemma 3.3.28. *Given H and C both finite and a risk function $R : C \times H \to \mathbb{R}$ and an associated probability set $\mathbb{P}_C$ on $(\Omega, \mathcal{F})$, Ω finite, if the function*

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

is linear and

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

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

Proof. By linearity of

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

we also have linearity of

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

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. $\qquad\square$

Lemma 3.3.29. *For a see-do model $(\mathbb{P}_{C \times H}, X, Y, D, H)$ with utility $u : X \times Y \to \mathbb{R}$, the map*

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

is linear.

Proof. By definition,

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

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

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

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

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

Lemma 3.3.30. *For finite X and D, the set of all Markov kernels $X \nrightarrow D$ is convex closed.*

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

Example 3.3.31. Suppose we have a conditionally independent see-do model $(\mathbb{P}_C, \mathsf{X}, \mathsf{Y}, \mathsf{D}, \mathsf{H})$ along with a bounded utility $u : Y \to \mathbb{R}$ where H, D, X and Y are all finite, and $\{\mathbb{P}_\alpha^{\mathsf{D}|\mathsf{X}} | \alpha \in C\}$ is the set of all Markov kernels $X \rightsquigarrow D$. 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 3.3.28, 3.3.29 and 3.3.30.

Example 3.3.32. As before, but suppose we have the see-do model is not conditionally independent. Because none of the lemmas 3.3.28, 3.3.29 and 3.3.30 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.

3.4 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.

Chapter 4

Models of repeatable decision problems

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 3.3.5 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 1.1.1). 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" and the decision maker has 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. If the person previously discussed frequently has headaches, then whenever they have a headache they might sometimes take the medicine and sometimes avoid it. Under the assumption that the function that maps inputs (medicine) to outputs (headaches) is repeated, they can infer that the way their headaches respond to medicine in the future will be the same as the way they have responded to it in the past.

However, it's not entirely clear what this assumption – that the decision maker's choice determines inputs to a fixed response function – actually means. The 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. Existing causal inference frameworks rely on some version of this assumption to licence inference from a sequence of observations to the consequences of an action. The possibility to identify causal effects given blocked backdoor paths in the structural intervention framework (Pearl, 2009, Ch. 1) depends on the fact that the distribution of the effect variable conditional on the cause and

the variables blocking the backdoor paths is unchanged after intervention. The assumption of conditional ignorability (Rubin, 2005) similarly implies that the distribution of the output conditional on the input and the covariates does not depend on which counterfactual is "chosen". This point is further 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). However, it is hard to make a positive case for this assumption and in many common cases it is clearly violated. The purpose of this chapter is to explore this issue in detail: when, and to what degree is the assumption at least reasonable, even if not justified in any absolute sense. 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 – 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 inappropriate[1].

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 showed 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 assigned[2].

---

[1]It'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).

[2]Note 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

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 for 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

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, at least when applied to observed variables. 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 weaker assumption of *precedented responses*.

This chapter also has a preliminary investigation into repeated response functions in the case of data-dependent models, where inputs are allows 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

---

distributed variables, see Walley (1991, pg. 463). We consider the "Bayesian" setting here of a single stochastic function because it is simpler.

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* (CIIR sequences), a precise term for what we refer to above as "repeated response functions". The key theorem in this chapter, Theorem 4.3.21, 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 4.5.15 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 4.3.21. Section 4.4 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 4.6.

4.1 Previous work on causal symmetries

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 (Diaconis and Freedman, 1980; Kerns and Székely, 2006) and for partially exchangeable arrays (Aldous, 1981). A comprehensive overview of results is presented in Kallenberg (2005a). 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) discussed 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 discussed the application of this model to questions of causation, but did 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) made use of the assumption of exchangeable potential outcomes to prove several identification results. Saarela et al. (2020), used structural causal models to propose *conditional exchangeability*, defined as the exchangeability of the non-intervened causal parents of a target variable under intervention on some of its parents. Sareela et. al. suggested 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: Banerjee, Chassang and Snowberg (2017); Dawid (2020); Greenland and Robins (1986); Hernán (2012); Hernán and Robins (2006) all discuss assumptions of this type. Section **??** 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 4.3.3), which adds the assumption of *locality*. This additional assumption has 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 tranformation 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 **??**, 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 $(\mathsf{D}_i, \mathsf{Y}_i)$ 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 id_C, then they may reduce the overall problem of choosing id_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 id_C given the input D_i is important for this reduction; otherwise the decision maker needs to consider how Y_i depends on id_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 $(\mathsf{D}_i, \mathsf{H})$ 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 $(\Omega, \mathcal{F})$ with variables $\mathsf{Y} := (\mathsf{Y}_i)_{i \in \mathbb{N}}$ and $\mathsf{D} := (\mathsf{D}_i)_{i \in \mathbb{N}}$ has *independent and identical response functions conditional on* H if for all i, $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{D}_{[1,i)}, \mathsf{Y}_{[1,i)})|(\mathsf{H}, \mathrm{id}_C)$ and $\mathbb{P}_\alpha^{\mathsf{Y}_i | \mathsf{D}_i \mathsf{H}} = \mathbb{P}_\alpha^{\mathsf{Y}_j | \mathsf{D}_j \mathsf{H}}$ 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 determin-istically 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) $\mathbb{P}_C^{Y \wr D|W}$ 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 4.3.21 and 4.3.26.

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 ever sampling the food from either.

The assumptions of exchange commutativity and locality together make input-output con-tractibility, 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, $\mathbb{K} : X^2 \to Y^2$), 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 $((\mathbb{P}_\cdot, (\Omega, \mathcal{F}), (C, \mathcal{C})), \mathsf{D}, \mathsf{Y})$ where $(\mathbb{P}_\cdot, (\Omega, \mathcal{F}), (C, \mathcal{C}))$ is a decision model, D is a sequence of "inputs" $\mathsf{D} := (\mathsf{D}_i)_{i \in \mathbb{N}}$ and Y is a corresponding sequence of "outputs" $\mathsf{Y} = (\mathsf{Y}_i)_{i \in \mathbb{N}}$ where $\mathsf{D}_i : \Omega \to D$ and $\mathsf{Y}_i : \Omega \to Y$.

Nxample 4.3.2. We use the shorthand $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ to refer to a sequential input-output model $((\mathbb{P}_\cdot, (\Omega, \mathcal{F}), (C, \mathcal{C})), \mathsf{D}, \mathsf{Y})$, with $(\Omega, \mathcal{F})$ implicit.

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

$$
\begin{array}{cc}
\mathsf{W} \\
\mathsf{D}_{[n]} \;\;\boxed{\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}}\;\; \mathsf{Y}_{[n]} & \quad \mathsf{W} \\
\mathsf{D}_{(n,\infty)} \qquad\qquad\qquad \ast & \mathsf{D}_{[n]} \;\;\boxed{\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}_{[n]}}}\;\; \mathsf{Y}_{[n]} \\
& \mathsf{D}_{(n,\infty)} \qquad\qquad\qquad \ast
\end{array}
$$

$$=$$

$$\iff$$

$$
\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}\Big(\bigtimes_{i \in [n]} A_i \times Y^{\mathbb{N}} \,\Big|\, w, x_{[n]}, x_{[n]^{\complement}} \Big) = \mathbb{P}_C^{\mathsf{Y}_{[n]}|\mathsf{WD}_{[n]}}\Big(\bigtimes_{i \in [n]} A_i \,\Big|\, w, x_{[n]} \Big)
$$

$$\forall A_i \in \mathcal{Y}, (x_{[n]}, x_{[n]^{\complement}}) \in \mathbb{N}, w \in W$$

That is, $\mathsf{Y}_i \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{X}_{(i,\infty)} | (\mathsf{W}, \mathsf{X}_i, \mathrm{id}_C)$.

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.

Nxample 4.3.4. Given a sequence $\mathsf{Y} := (\mathsf{Y}_i)_{i \in \mathbb{N}}$ and a permutation $\rho : \mathbb{N} \to \mathbb{N}$, the permuted sequence Y_ρ is defined to be $(\mathsf{Y}_{\rho(i)})_{i \in \mathbb{N}}$.

Definition 4.3.5 (Exchange commutativity)**.** Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ along with some $\mathsf{W} : \Omega \to W$, $\alpha \in C$ we say $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ *commutes with exchange* over W if for all finite permutations $\rho : \mathbb{N} \to \mathbb{N}$

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

IO contractibility is the conjunction of both previous assumptions.

Definition 4.3.6 (IO contractibility)**.** Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ along with some $\mathsf{W} : \Omega \to W$, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ is *IO contractible* over W if it is local and commutes with exchange.

If $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ IO contractible over W, then for any subsequences $A, B \subset \mathbb{N}$ with $|A| = |B|$, we have $\mathbb{P}_\alpha^{\mathsf{Y}_A|\mathsf{WD}_A} = \mathbb{P}_\alpha^{\mathsf{Y}_B|\mathsf{WD}_B}$. In fact, Theorem 4.3.7 shows a stronger condition: if we take $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ and multiply it by $\mathrm{del}_{A^{\complement}}$ which erases all the indices in $A^{\complement}$, we get the same result as

multiplying $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ by $del_{B^\complement}$. The fact that arbitrary contractions of the sequence yield the same result motivates the name *IO contractibility*.

Theorem 4.3.7 (Equality of equally sized contractions)**.** *Given a sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *and some* W, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ *is IO contractible over* W *if and only if for all subsequences* $A, B \subset \mathbb{N}^{|A|}$ *and for every* α

$$\mathbb{P}_\alpha^{\mathsf{Y}_A|\mathsf{WD}_{A,\mathbb{N}\setminus A}} = \mathbb{P}_\alpha^{\mathsf{Y}_B|\mathsf{WD}_{B,\mathbb{N}\setminus B}}$$
$$= \mathbb{P}_\alpha^{\mathsf{Y}_A|\mathsf{WD}_A} \otimes del_{D^{|\mathbb{N}\setminus A|}}$$

Proof. Appendix B.1 $\qquad\qquad\square$

Theorem 4.3.8 shows that neither locality nor exchange commutativity is implied by the other.

Theorem 4.3.8. *Exchange commutativity does not imply locality or vice versa.*

Proof. We prove the claim by way of presenting counterexamples.

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

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

for some sequence $(d_i)_{i\in\mathbb{N}}$ such that this limit exists. Then for any finite permutation ρ

$$\mathbb{P}_C^{\mathsf{Y}_\rho|\mathsf{D}_\rho}\left(\bigtimes_{i\in\mathbb{N}} A_i \Big| (d_i)_{i\in\mathbb{N}}\right) = \prod_{i\in\mathbb{N}} \delta_{\lim_{n\to\infty}\sum_{i\in\mathbb{N}} \frac{d_{\rho^{-1}(i)}}{n}}(A_{\rho^{-1}(i)})$$
$$= \mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}\left(\bigtimes_{i\in\mathbb{N}} A_i \Big| (d_i)_{i\in\mathbb{N}}\right)$$

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

$$\mathbb{P}_C^{\mathsf{Y}_1|\mathsf{D}}(A_1|0, 1, 1, 1....) = \delta_1(A_1)$$
$$\mathbb{P}_C^{\mathsf{Y}_1|\mathsf{D}}(A_1|0, 0, 0, 0....) = \delta_0(A_1)$$

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

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

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

then

$$\mathbb{P}_C^{\mathsf{Y}_\rho|\mathsf{D}_\rho}\left(\bigtimes_{i\in\mathbb{N}} A_i \Big| (d_i)_{i\in\mathbb{N}}\right) = \prod_{i\in\mathbb{N}} \delta_i(A_{\rho^{-1}(i)})$$
$$\neq \prod_{i\in\mathbb{N}} \delta_i(A_i)$$
$$= \mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}\left(\bigtimes_{i\in\mathbb{N}} A_i \Big| (d_i)_{i\in\mathbb{N}}\right)$$

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

$$\mathbb{P}_C^{\mathsf{Y}_{[n]}|\mathsf{D}}(\underset{i\in[n]}{\times} A_i|(d_i)_{i\in\mathbb{N}}) = \prod_{i\in[n]} \delta_i(A_{\rho^{-1}(i)})$$
$$= \mathbb{P}_C^{\mathsf{Y}_{[n]}|\mathsf{D}}(\underset{i\in[n]}{\times} A_i|(0)_{i\in\mathbb{N}})$$

so $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ is local. $\qquad\square$

Theorem 4.3.8 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 ith 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 outputs. 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 $\mathsf{D}_1 = 0$), or flip a second coin use this as the input for step 2 (this is the input $\mathsf{D}_1 = 1$). D_2 is an arbitrary single-valued variable. Then for all d_1, d_2

$$\mathbb{P}^{\mathsf{Y}_1|\mathsf{D}}(y_1|d_1, d_2) = 0.5$$
$$\mathbb{P}^{\mathsf{Y}_2|\mathsf{D}}(y_2|d_1, d_2) = 0.5$$

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

In this section, we prove the main Theorem 4.3.21, which shows that a sequence of inputs and corresponding outputs features conditionally independent and identical responses if and only if the conditional distribution of the outputs given the inputs is IO contractible over some variable W.

We make use of a number of concepts in the following work: models with sequences of inputs and outputs, "tabulated" representations of conditional probabilities and "hypotheses" or "directing measures" defined as the limit of relative frequencies. These are all defined below.

Definition 4.3.9 (Count of input values). Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ on $(\Omega, \mathcal{F})$ with countable D, $\#_j^k$ is the variable

$$\#_j^k := \sum_{i=1}^{k-1} [\![\mathsf{D}_i = j]\!]$$

In particular, $\#_j^k$ is equal to the number of times $\mathsf{D}_i = j$ over all $i < k$.

Definition 4.3.10 (Tabulated conditional distribution). Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ on $(\Omega, \mathcal{F})$, define the tabulated conditional distribution $\mathsf{Y}^D : \Omega \to \mathsf{Y}^{\mathbb{N} \times D}$ by

$$\mathsf{Y}_{ij}^D = \sum_{k=1}^{\infty} [\![\#_j^k = i - 1]\!] [\![\mathsf{D}_k = j]\!] \mathsf{Y}_k$$

That is, the (i, j)-th coordinate of $\mathsf{Y}^D(\omega)$ is equal to the coordinate $\mathsf{Y}_k(\omega)$ for which the corresponding $\mathsf{D}_k(\omega)$ is the ith instance of the value j in the sequence $(\mathsf{D}_1(\omega), \mathsf{D}_2(\omega), ...)$, or 0 if there are fewer than i instances of j in this sequence.

Definition 4.3.11 (Measurable set of probability distributions). Given a measurable set $(\Omega, \mathcal{F})$, 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 map from a set to the limit of normalised partial sums of indicator functions over that set, where that limit exists. We refer to directing random measures with the letter H by default, and treat it like a hypothesis – under appropriate conditions, the directing random measure H is almost surely equal to the distribution of any variable in the sequence conditional on H. We also define H in the case that the relevant limit does not exist for completeness, although we are only interested in cases where the limit does exist. Definition 4.3.12 reduces to the definition of a directing random measure given in Kallenberg (2005b) when we consider a probability space instead of a probability set.

Definition 4.3.12 (Directing random measure). Given a probability set $(\mathbb{P}_C, \Omega, \mathcal{F})$ and a sequence $\mathsf{X} := (\mathsf{X}_i)_{i \in \mathbb{N}}$, the directing random measure of X written $\mathsf{H} : \Omega \to \mathcal{M}_1(X)$ is the function

$$\mathsf{H} := A \mapsto \begin{cases} \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{\infty} \mathbb{1}_A(\mathsf{X}_i) & \text{this limit exists for all } \alpha \in C \\ [\![A = X]\!] & \text{otherwise} \end{cases}$$

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 $((\mathsf{Y}_{1j}^D)_{j \in D}, (\mathsf{Y}_{2j}^D)_{j \in D}, ...)$.

Definition 4.3.13 (Directing random conditional)**.** Given a sequential input-output model $(\mathbb{P}_C, D, Y)$, we will say the directing random measure $H : \Omega \to \mathcal{M}_1(Y^D)$ is the function

$$H := \underset{j \in D}{\bigtimes} A_j \mapsto \begin{cases} \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{\infty} \prod_{j \in D} \mathbb{1}_{A_j}(Y_{ij}^D) & \text{this limit exists} \\ [\![\bigtimes_{j \in D} A_j = Y^D]\!] & \text{otherwise} \end{cases}$$

We say a model satisfies data-independence when future inputs are independent of outputs conditional on past inputs and the directing measure H. This makes the analysis much easier, but it rules out the analysis of decision rules that depend in some non-symmetric way on the input-output sequence (or on finite subsequences).

Definition 4.3.14 (Data-independent)**.** A sequential input-output model $(\mathbb{P}_C, D, Y)$ is weakly data-independent if $Y_i \perp\!\!\!\perp^e_{\mathbb{P}_C} D_{(i,\infty]} | (H, D_{[1,i]}, \mathrm{id}_C)$.

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 4.3.15 (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 $\eta : \mathbb{N} \times D \to \mathbb{N} \times D$ such that for each $j \in D$, $\eta_j := \eta(\cdot, j)$ is a finite permutation $\mathbb{N} \to \mathbb{N}$ and $\eta(i, j) = (\eta_j(i), j)$.

Lemma 4.3.17 shows that an 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 interpreting such a model as potential outcomes requires additional assumptions we don't make here. By representing a conditional probability as an exchangeable regular probability distribution, we can apply De Finetti's, which is a key step in proving the main result of Theorem 4.3.21.

To prove Lemma 4.3.17, we assume that the set of input sequences in which each value appears infinitely often has measure 1 for every option in C. Without this assumption, the tabulated conditional Y^D cannot be a function of the inputs D and outputs Y as there would be some values of the inputs which would not be seen often enough. Without this assumption, Y^D may not be uniquely defined. We call this side condition *infinite support*.

Definition 4.3.16 (Infinite support)**.** Given a sequential input-output model $(\mathbb{P}_C, D, Y)$ with D countable if, letting $E \subset D^{\mathbb{N}}$ be the set of all sequences for which each $j \in D$ occurs infinitely often, $\mathbb{P}_\alpha^{D|W}(E|w) = 1$ for all α, w, then we say D is *infinitely supported over* W.

Lemma 4.3.17. *Suppose a sequential input-output model* $(\mathbb{P}_C, D, Y)$ *is given with D countable and D infinitely supported over W. Then for some W, α, $\mathbb{P}_\alpha^{Y|WD}$ is IO contractible if and only if*

$$\mathbb{P}_\alpha^{Y|WD}(\underset{i \in \mathbb{N}}{\bigtimes} A_i | w, (d_i)_{i \in \mathbb{N}}) = \mathbb{P}_\alpha^{(Y_{id_i}^D)_{i \in \mathbb{N}} | W}(\underset{i \in \mathbb{N}}{\bigtimes} A_i | w) \qquad \forall A_i \in \mathcal{Y}^D, w \in W, d_i \in D$$

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

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

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

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

Proof. Only if: We define a random invertible function $R : \Omega \times \mathbb{N} \to \mathbb{N} \times D$ that reorders the indicies so that, for $i \in \mathbb{N}, j \in D$, $D_{R^{-1}(i,j)} = j$ almost surely. We then use IO contractibility to show that $\mathbb{P}_\alpha^{Y|D}(\cdot|d)$ is equal to the distribution of the elements of Y^D selected according to $d \in D^{\mathbb{N}}$.

If: We construct a conditional probability according to Definition 4.3.10 and verify that it satisfies IO contractibility.

The full proof can be found in Appendix B.2. $\qquad \square$

As a consequence of Lemma 4.3.17 along with De Finetti's representation theorem, we can say that given $(\mathbb{P}_C, D, Y)$ IO contractible, conditioning on H renders the columns of Y^D independent and identically distributed.

Lemma 4.3.18. *Suppose a sequential input-output model $(\mathbb{P}_C, D, Y)$ is given with D countable, D infinitely supported over W and for some W, $\mathbb{P}_\alpha^{Y|WD}$ is IO contractible for all α. Then, letting H be the directing random conditional of $(\mathbb{P}_C, D, Y)$ (Definition 4.3.13) and $Y_{iD}^D := (Y_{ij}^D)_{j \in D}$, we have for all $i \in \mathbb{N}$, $Y_{iD}^D \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{\mathbb{N} \setminus \{i\}D}^D, W)|(H, id_C)$ and $\mathbb{P}_\alpha^{Y_{iD}^D} = \mathbb{P}_\alpha^{Y_{kD}^D}$ and*

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

Proof. This follows directly from applying De Finetti's representation theorem to Y^D, see Appendix B.2. $\qquad \square$

If the conditions of Lemma 4.3.17 are satisfied, we do not need the full sequence of pairs (D, Y) to calculate H; any subsequence $A \subset \mathbb{N}$ that satisfies the condition that D_A is infinitely supported over W is sufficient.

Theorem 4.3.19. *Suppose a sequential input-output model $(\mathbb{P}_C, D, Y)$ is given with D countable, D infinitely supported over W and for some W, $\mathbb{P}_\alpha^{Y|WD}$ is IO contractible for all α. Consider an infinite set $A \subset \mathbb{N}$, and let $D_A := (D_i)_{i \in A}$ and $Y_A := (Y_i)_{i \in A}$ such that D_A is also infinitely supported over W. 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 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.

The proof is in Appendix B.2. $\qquad \square$

The following is a technical lemma that will be used in Theorem 4.3.21.

Lemma 4.3.20. *Suppose a sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *is given with* D *countable,* D *infinitely supported over* W, *for some* W, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ *is IO contractible for all* α *and for all* α

$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}} = \begin{array}{c} \mathsf{W} - \boxed{\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}} \\ \mathsf{D} \underline{\hspace{2cm}} \end{array} \boxed{\mathbb{F}_{\text{lu}}} \underline{\hspace{1cm}} \mathsf{Y}$$

then $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathsf{W}|(\mathsf{H}, \mathsf{D}, id_C)$ *and for all* α

$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{HD}} = \begin{array}{c} \mathsf{H} - \boxed{\mathbb{P}_C^{\mathsf{Y}^D|\mathsf{H}}} \\ \mathsf{D} \underline{\hspace{2cm}} \end{array} \boxed{\mathbb{F}_{\text{lu}}} \underline{\hspace{1cm}} \mathsf{Y}$$

Proof. We show that the function that maps the variables Y and D to H also maps Y^D and the constant $e \in D^{\mathbb{N}}$ to H' with $\mathsf{H}' \overset{\mathbb{P}_C}{\cong} \mathsf{H}$, and the result follows from disintegration along with a conditional independence given by Lemma 4.3.17.

The full proof is in Appendix B.3. $\qquad\square$

Theorem 4.3.21 is the main result of this section. It shows that sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ is IO contractible over some W if and only if there is some hypothesis H such that the Y_is are related to the D_is by conditionally independent and identical response functions (subject to a support assumption).

In the following theorem, property (2) is equivalent to the conjunction of conditionally independent and identical response functions (Def 4.2.1) and weak data-independence (Def 4.3.14).

Theorem 4.3.21 (Representation of IO contractible models)**.** *Suppose a sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *with sample space* $(\Omega, \mathcal{F})$ *is given with* D *countable and* D *infinitely supported over* W. *Then the following are equivalent:*

1. *There is some* W *such that* $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ *is IO contractible for all* α

2. *For all* i, $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (\mathsf{Y}_{\neq i}, \mathsf{D}_{\neq i}, id_C)|(\mathsf{H}, \mathsf{D}_i)$ *and for all* i, j, α

$$\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{HD}_i} = \mathbb{P}_\alpha^{\mathsf{Y}_j|\mathsf{HD}_j}$$

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

$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{HD}} = \begin{array}{c} \mathsf{H} \underline{\hspace{0.5cm}} \bullet \underline{\hspace{1cm}} \boxed{\mathbb{L}} \underline{\hspace{0.3cm}} \mathsf{Y}_i \\ \mathsf{D}_i \underline{\hspace{2cm}} \\ \hspace{3cm} i \in \mathbb{N} \end{array}$$

Proof. (1) $\implies$ (3): We apply Lemma 4.3.17 followed by Lemma 4.3.18 followed by Lemma 4.3.20.

(3) $\implies$ (2): We verify that the required conditional independences hold assuming (3).

(2) $\implies$ (1): We show that, assuming (2), then $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ is IO contractible over W for all α.

See Appendix B.4 for the full proof. $\qquad\square$

As a consequence of Theorem 4.3.21, if a sequence of input and output pairs features independent and identical responses conditional on some arbitrary variable, then we can without loss of generality consider the conditioning variable to be the directing random conditional defined over the same sequence of input-output pairs.

Corollary 4.3.22. *If a sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *has independent and identical response functions conditional on some variable* G *and* D *has infinite support over* W, *then letting* H *be the directing random conditional with respect to inputs* D *and outputs* Y, *it follows that for for all* i, $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathsf{G}|(\mathsf{D}_i, \mathsf{H}, id_C)$ *and for all* α, i, j, $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}} = \mathbb{P}_\alpha^{\mathsf{Y}_j|\mathsf{D}_j\mathsf{H}}$.

Proof. We have $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{GD}}$ is IO contractible over G. The conclusion follows by applying Theorem 4.3.21. □

4.3.2 Symmetries of sequences with conditionally independent and identical responses

Theorem 4.3.21 says that a data independent sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ features conditionally independent and identical response functions $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{HD}_i}$ for all α if and only if there is some W such that $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ is IO contractible over W for all α. The variable W is something of a nuisance; rather than thinking only about whether IO contractibility holds, we must consider whether there's *any* variable that licenses the assumption of IO contractibility.

A simple special case to consider is when W is single valued – that is, when $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}}$ is IO contractible. As Theorem 4.3.23 shows, this corresponds to the CIIR sequence models where the inputs D are unconditionally data-independent and independent of the hypothesis H. We can also consider the case where $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ 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).

More generally, we can observe that by Lemma 4.3.22, any sequence of inputs and outputs with conditionally independent and identical responses with infinitely supported inputs must be causally contractible over H. Furthermore, by Theorem 4.3.19, we can take H to be a function of any sequence of inputs and outputs for which the inputs have infinite support. Together, these imply IO contractibility over arbitrary subsequences with infinite support. This observation is the basis of Theorem 4.3.26. Applying Theorem 4.3.7 allows us to state this implication in a manner that we feel is more intuitive: informally speaking, if we want to predict any output from its corresponding input and some infinitely supported subsequence of data, then it doesn't matter which output we pick or which subsequence of data we pick, the problem remains essentially the same in every case.

Theorem 4.3.23 (Data-independent IO contractibility). *Suppose a sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *with sample space* $(\Omega, \mathcal{F})$ *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,* $\mathbb{P}_\alpha^{\mathsf{D}}(E) = 1$ *for all* α. *Then the following are equivalent:*

1. $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}}$ *is IO contractible for all* α

2. *For all* i, $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (\mathsf{Y}_{\neq i}, \mathsf{D}_{\neq i}, id_C)|(\mathsf{H}, \mathsf{D}_i)$, *for all* i, j, α

$$\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{HD}_i} = \mathbb{P}_\alpha^{\mathsf{Y}_j|\mathsf{HD}_j}$$

, $\mathsf{H} \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathsf{D}|id_C$ *and for all* i $\mathsf{D}_i \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathsf{D}_{(i,\infty]})|(\mathsf{D}_{[1,i)}, id_C)$

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

$$\mathbb{P}_\alpha^{\mathsf{YH|D}} =$$
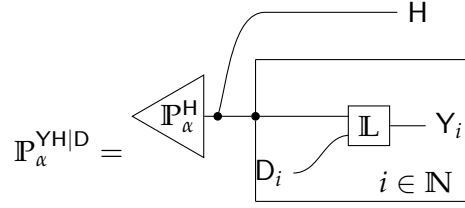
Proof. See Appendix B.4. □

While $\mathbb{P}_\mathsf{C}^{\mathsf{Y|D}}$ exchange commutative is not necessarily IO contractible, exchange commutativity of this conditional implies IO contractibility over the directing random conditional H, and thus is sufficient for conditionally independent and identical responses.

Theorem 4.3.24. *If* $\mathbb{P}_\mathsf{C}^{\mathsf{Y|D}}$ *is exchange commutative, and for each* α $\mathbb{P}_\alpha^{\mathsf{D}}$ *is absolutely continuous with respect to some exchangeable distribution* $\mathbb{Q}_\alpha^{\mathsf{D}}$ *in* $\Delta(D^\mathbb{N})$ *with directing random measure* F *and* D *infinitely supported over* F *with respect to* $\mathbb{Q}_\alpha$ *, then* $\mathbb{P}_\alpha^{\mathsf{Y|HD}}$ *is IO contractible, where* H *is the directing random conditional for* $\mathbb{P}_\alpha^{\mathsf{Y|D}}$.

Proof. We show that there is an exchangeable distribution for which the relevant conditional automatically satisfies IO contractibility and is almost surely equal to $\mathbb{P}_\alpha^{\mathsf{Y|GD}}$ for some G. □

Corollary 4.3.25. *If* $(\mathbb{P}_\mathsf{C}, \mathsf{D}, \mathsf{Y})$ *is exchange commutative over* $*$, *and for each* α $\mathbb{P}_\alpha^{\mathsf{D}}$ *is absolutely continuous with respect to some exchangeable distribution in* $\Delta(D^\mathbb{N})$ *then*

$$\mathbb{P}_\alpha^{\mathsf{Y|HD}} =$$
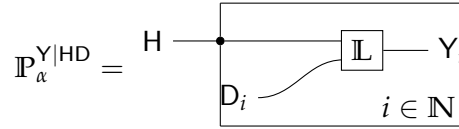
Proof. By Theorem 4.3.24, $\mathbb{P}_\alpha^{\mathsf{Y|WD}}$ is IO contractible over some W for all α, so the result follows immediately from Theorem 4.3.21. □

Theorem 4.3.26 shows that the interchangeability of infinitely supported subsequences of data is a necessary condition for CIIR sequences. We only show sufficiency under the assumption that the input-output sequence is exchangeably dominated, and it is an open question if sufficiency holds in the general case.

Theorem 4.3.26. *A data-independent sequential input-output model* $(\mathbb{P}_\mathsf{C}, \mathsf{D}, \mathsf{Y})$ *features conditionally independent and identical response functions* $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{G}}$ *with* D *infinitely supported over* G *only if for any sets* $A, B \subset \mathbb{N}$ *such that* D_A *and* D_B *are also infinitely supported over* G *and any* $i, j \in \mathbb{N}$ *such that* $i \notin A$, $j \notin B$,

$$\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{Y}_A,\mathsf{D}_A} = \mathbb{P}_\alpha^{\mathsf{Y}_j|\mathsf{D}_j|RV\mathsf{Y}_B\mathsf{D}_B}$$

. If in addition each $\mathbb{P}_\alpha^{\mathsf{YD}}$ *is dominated by some* $\mathbb{Q}_\alpha$ *such that* $\mathbb{Q}_\alpha^{\mathsf{YD}}$ *is exchangeable, then the reverse implication also holds.*

Proof. See Appendix B.4. □

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 4.3.21 shows that for weakly data-independent models, conditionally independent and identical responses is equivalent to IO contractibility over the directing random conditional H. Where $\mathbb{P}_\alpha^{\mathsf{YD}}$ is dominated by an exchangeable measure for every α, Theorem 4.3.26 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 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 have exchange commutativity (or IO contractibility) over the trivial variable $\mathsf{W} = *$ in Definitions 4.3.5 and 4.3.6 respectively. Theorem 4.3.21 and Corollary 4.3.25 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 $\mathsf{YD} := (\mathsf{D}_i, \mathsf{Y}_i)_{i \in \mathbb{N}}$ has $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ exchange commutative over $*$. This follows from the fact that

$$\mathbb{P}_C^{\mathsf{YD}} = \mathbb{P}_C^{(\mathsf{YD})_\rho}$$
$$\implies \mathbb{P}_C^{\mathsf{Y}|\mathsf{D}} = \mathbb{P}_C^{\mathsf{Y}_\rho|\mathsf{D}_\rho}$$

thus by Corollary 4.3.25, $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ features conditionally independent and identical response functions. Note that $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ 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 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

$$\mathbb{P}_C^{\mathsf{Y}_2|\mathsf{D}_2}(100|1) = \sum_{0,1} \mathbb{P}_C^{\mathsf{Y}_2|\mathsf{D}_2\mathsf{H}}(100|1, 0)\mathbb{P}_C^{\mathsf{H}|\mathsf{D}_2}(0|1) + \mathbb{P}_C^{\mathsf{Y}_2|\mathsf{D}_2\mathsf{H}}(100|1, 1)\mathbb{P}_C^{\mathsf{H}|\mathsf{D}_2}(1|1)$$
$$= 1 - \epsilon$$

but

$$\mathbb{P}_C^{Y_2|D_1D_2}(100|0,1) = \sum_{0,1} \mathbb{P}_C^{Y_2|D_1D_2H}(100|0,1,0)\mathbb{P}_C^{H|D_1D_2}(0|0,1) + \mathbb{P}_C^{Y_2|D_1D_2H}(100|0,1,1)\mathbb{P}_C^{H|D_1D_2}(1|0,1)$$

$$= 0.5$$

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 the choose the sequence D, and so H has no dependence on D.

$$\mathbb{P}_C^{Y_2|D_2}(100|1) = \sum_{0,1} \mathbb{P}_C^{Y_2|D_2H}(100|1,0)\mathbb{P}_C^H(0) + \mathbb{P}_C^{Y_2|D_2H}(100|1,1)\mathbb{P}_C^H(1)$$

$$= 0.5$$

$$\mathbb{P}_C^{Y_2|D_1D_2}(100|0,1) = \sum_{0,1} \mathbb{P}_C^{Y_2|D_1D_2H}(100|0,1,0)\mathbb{P}_C^H(0) + \mathbb{P}_C^{Y_2|D_1D_2H}(100|0,1,1)\mathbb{P}_C^H(1)$$

$$= 0.5$$

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 $\mathbb{P}_\alpha^{YZ|DE}$ is IO contractible over $*$ for all α, understanding (D, E, Y, Z) to be a single sequence of pairs, then Theorem 4.3.7 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 DMs 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,

$$\mathbb{P}_\alpha^{Z_i|E_iD_{[2]}Y_{[2]}}(100|1,1,1,0,100) > \mathbb{P}_\alpha^{Y_i|E_iE_{[2]}Z_{[2]}}(100|1,1,1,0,100)$$

In this case, the DMs model is not even exchange commutative over $*$.

4.4.2 Implications of IO contractibility

Theorem 4.3.26 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.

We are of the view that it is not simply "hard to know" when this condition is reasonable – in fact, it's often easy to know that it is unreasonable. One might respond that we might still accept that the condition is close to holding, and in this case it 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?

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_is represent whether patient i was given a particular medicine. The D_is 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_is are covariates, and the Y_is 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 4.3.26, the assumption of conditionally independent and identical responses applied to $((D, X), Y)$ implies

$$
\begin{aligned}
\mathbb{P}_\alpha^{Y_0 | D_0 X_0 D_{\text{odds}} X_{\text{odds}} Y_{\text{odds}}} &= \mathbb{P}_\alpha^{Y_0 | D_0 D_{\text{evens}\backslash\{0\}} X_{\text{evens}\backslash\{0\}} Y_{\text{evens}\backslash\{0\}}} \\
&= \mathbb{P}_\alpha^{Y_2 | D_2 X_2 X_{\text{evens}\backslash\{0,2\}} Y_{\text{evens}\backslash\{0,2\}}} \\
&= \mathbb{P}_\alpha^{Y_2 | D_2 X_2 X_{\text{odds}} Y_{\text{odds}}}
\end{aligned}
$$

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, Moakler et al. (2022); Gordon, Zettelmeyer et al. (2018).

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 $(\mathbb{P}_C, (\mathsf{D}, \mathsf{X}), \mathsf{Y})$ where $(\mathbb{P}^{\mathsf{Y}|\mathsf{WDX}}$ is IO contractible, and:

- $i > n \implies \mathsf{X}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{D}_i | (\mathsf{H}, \mathrm{id}_C)$

- $\mathbb{P}_\alpha^{\mathsf{X}_i|\mathsf{H}} \cong \mathbb{P}_\alpha^{\mathsf{X}_1|\mathsf{H}}$ for all α

Then the model exhibits a kind of "backdoor adjustment". Specifically, for $i > n$

$$
\begin{aligned}
\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}}(A|d,h) &= \int_X \mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{D}_i\mathsf{H}}(A|d,x,h)\mathbb{P}_\alpha^{\mathsf{X}_i|\mathsf{D}_i\mathsf{H}}(\mathrm{d}x|d,h) \\
&= \int_X \mathbb{P}_\alpha^{\mathsf{Y}_1|\mathsf{X}_1\mathsf{D}_1\mathsf{H}}(A|d,x,h)\mathbb{P}_\alpha^{\mathsf{X}_i|\mathsf{H}}(\mathrm{d}x|h) \\
&= \int_X \mathbb{P}_\alpha^{\mathsf{Y}_1|\mathsf{X}_1\mathsf{D}_1\mathsf{H}}(A|d,x,h)\mathbb{P}_\alpha^{\mathsf{X}_1|\mathsf{H}}(\mathrm{d}x|h) \quad\quad (4.1)
\end{aligned}
$$

Equation (4.1) 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.

4.4.3 Causal assumptions and distinguishing inputs from outputs

Example 1 above shows that the assumption of a CIIR sequence applies to an exchangeable probability model with no real alternative options or anything else "causal". As a result, one might wonder whether the assumption of CIIR sequences really is a "causal" assumption (we have no clear idea what does or does not constitute a causal assumption here, we're merely noting that an assumption that applies to regular probability distributions seems like it might not fit the bill). The apparent claim that CIIR sequences nonetheless allow us to draw causal conclusions might them (seemingly) run afoul of Cartwright's adage "no causes in, no causes out" (Cartwright, 1994).

We have also supposed, in the background at least, that a relation between the chosen option α and the distribution of (some) inputs $\mathbb{P}_\alpha^{\mathsf{D}}$ was known in advance. This seems much more like a causal assumption – it really doesn't make sense to try to apply this assumption to a probability model with no option set.

There is no obvious reason why the problem has to be this way, but it is certainly intuitively appealing to think about "inputs" that we already know how to control and "outputs" that we want to learn how to control from the given data. In fact, it is perhaps this intuitive appeal that led us to name the two sequences "inputs" and "outputs" respectively.

Given an assumption of this type, there seems to be a distinction we can draw between inputs and outputs beyond an arbitrary convention. In particular, this assumption suggests that that D_i is distinguished from Y_i by the fact that we know how to control D_i *before* we've seen any data, but in the best case we will only know how to control Y_i *after* reviewing the data. This distinction does not obviously always correspond to a "causal" direction. For example, we might have a clearer idea of how to prompt someone to go and exercise than we have for

how to help them feel energetic, yet it's plausible that (under normal circumstances) feeling energetic causes people to engage in exercise.

Nevertheless, prior knowledge of this type suggests that the inputs D_i are susceptible to influence by at least one person who doesn't know exactly how the outputs depend on the inputs – namely, the decision maker. We will investigate in the following chapter a situation where prior knowledge like this (though not this exactly) may facilitate a conclusion of a CIIR input-output from a weaker starting assumption.

4.5 Conditionally independent and identical response functions with data-dependent inputs

The results of the previous section concern data independent models; in these models, inputs cannot depend on previous data except via an "oracle" on the hypothesis H. This differs from the "see-do" models discussed in the previous chapter, where some inputs can depend in arbitrary ways on the sequence of data preceding them. Intuitively, we might expect that something similar to Theorems 4.3.21 and 4.3.26 might hold. However, the situation is complicated by the fact that we can no longer arbitrarily shuffle the order of the input-output sequence; if future inputs depend on the past data, then we could have, for example, D_2 deterministically equal to Y_1 which would, in general, may Y_1 not independent of D_2 conditional on the hypothesis H. Theorem 4.5.14 is similar to Theorem 4.3.21 except it applies to this more general setting of data-dependent inputs. However, it also doesn't lend itself to any easy interpretation that we are aware of. We consider the work in this section to be preliminary.

Example 4.5.1. Consider an input-output model $(\mathbb{P}_C, D, Y)$ 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 for some hypothesis G:

$$Y_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (Y_{<i}, D_{<i}, \mathrm{id}_C)|GD_i \qquad\qquad \forall i \in \mathbb{N}$$
$$\wedge \mathbb{P}_\alpha^{Y_i|HD_i} = \mathbb{P}_\alpha^{Y_0|GD_0} \qquad\qquad \forall \alpha \in C, i \in \mathbb{N}$$

and, for simplicity, assume $G \perp\!\!\!\perp^e_{\mathbb{P}_C} (D, \mathrm{id}_C)$ also.

Then, for arbitrary $\alpha \in C$
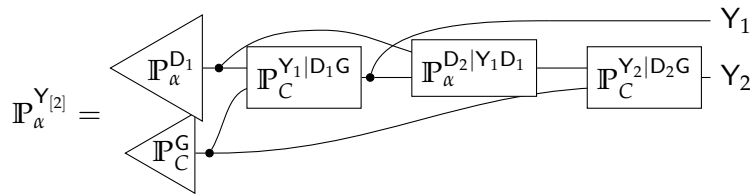
note that D_2 depends on Y_1 and D_1. $\mathbb{P}_\alpha^{D_2|Y_1D_1}$ has been "inserted" between the response conditionals $\mathbb{P}_C^{Y_1|D_1H}$ and $\mathbb{P}_C^{Y_2|D_2H}$.

Given $\mathbb{P}_C^{Y_1|D_1H}$ and $\mathbb{P}_C^{Y_2|D_2H}$, define the comb

$$\mathbb{P}_C^{Y_{[2]}\wr D_{[2]}} := \begin{array}{c}\text{diagram}\end{array}$$
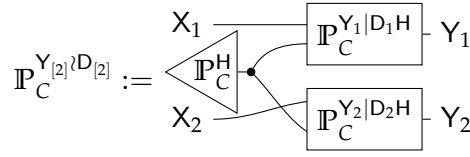
then $\mathbb{P}_C^{Y_{[2]}\wr D_{[2]}}$ is IO contractible. $\mathbb{P}_C^{Y_{[2]}\wr D_{[2]}}$ is *not* a uniform conditional probability; in general

$$\mathbb{P}_\alpha^{D_1D_2}\mathbb{P}_C^{Y_{[2]}\wr D_{[2]}} \neq \mathbb{P}_\alpha^{Y_1Y_2}$$

4.5.1 Combs

Combs are a generalisation of conditional distributions that support the "insert" operation that appears in Example 4.5.1. Speaking very roughly, where conditional distributions are Markov kernels missing a part "on the left", combs are Markov kernels that may be missing one or more parts "in the middle". If we provide a conditional distribution with its missing part – that is, if we compute the semidirect product of the conditional distribution and the appropriate marginal – we get the joint distribution of the variables that appear on the left and right sides of the conditional distribution. Similarly, if we "insert" all of the missing parts into a comb, we get the joint distribution of all of the incoming and outgoing variables that appear in the comb.

A graphical depiction of the "insert" operation gives some intuition for why it is called "insert":

$$\mathbb{P}_\alpha^{Y_1D_2Y_2|D_1} = \text{insert}(\mathbb{P}_\alpha^{D_2|D_1Y_1}, \mathbb{P}_C^{Y_{[2]}\wr D_{[2]}})$$

$$(4.2)$$

$$(4.3)$$
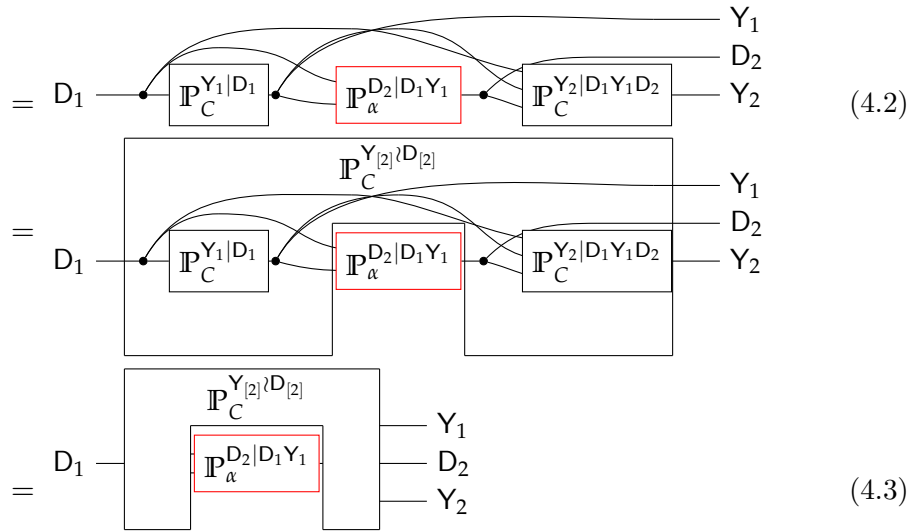
While Equation (4.2) is a well-formed string diagram in the category of Markov kernels, Equation (4.3) is not. In the case that all the underlying sets are discrete, Equation (4.3) can be defined using an extended string diagram notation appropriate for the category of real-valued matrices (Jacobs et al., 2019), though we do not introduce this extension here.

Formal definitions of combs follow. As with conditional probabilities, a *uniform n-comb* $\mathbb{P}_C^{Y_{[n]}\wr X_{[n]}}$ is a Markov kernel that satisfies the definition of an *n*-comb for each $\alpha \in C$.

Definition 4.5.2 (*n*-Comb)**.** Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ with variables $Y_i : \Omega \to Y$, $D_i : \Omega \to D$ for $i \in [n]$ and $W : \Omega \to W$, the uniform *n*-comb $\mathbb{P}^{Y_{[n]} \wr D_{[n]} | W} : W \times D^n \rightarrow Y^n$ is the Markov kernel given by the recursive definition

$$\mathbb{P}^{Y_1 \wr D_1 | W} = \mathbb{P}^{Y_1 | D_1 W}$$

Definition 4.5.3 ($\mathbb{N}$-comb)**.** Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ with variables $Y_i : \Omega \to Y$ and $D_i : \Omega \to D$, for $i \in \mathbb{N}$ and $W : \Omega \to W$, the $\mathbb{N}$-comb $\mathbb{P}^{Y_{\mathbb{N}} \wr D_{\mathbb{N}} | W} : W \times D^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ is the Markov kernel such that for all $n \in \mathbb{N}$

$$\mathbb{P}^{Y_{\mathbb{N}} \wr D_{\mathbb{N}} | W}[\mathrm{id}_{Y^n} \otimes \mathrm{del}_{Y^{\mathbb{N}}}] = \mathbb{P}^{Y_{[n]} \wr D_{[n]} | W} \otimes \mathrm{del}_{Y^{\mathbb{N}}}$$

Theorem 4.5.4 (Existence of $\mathbb{N}$-combs)**.** *Given a probability set* $\mathbb{P}$ *with variables* $Y_i : \Omega \to Y$ *and* $D_i : \Omega \to D$ *for* $i \in \mathbb{N}$ *and* $W : \Omega \to W$, 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

$$\mathbb{P}^{Y_{[n]} \wr D_{[n]} | W}[\mathrm{id}_{Y^{n-m}} \otimes \mathrm{del}_{Y^m}] = \mathbb{P}^{Y_{[n-m]} \wr D_{[n-m]}} \otimes \mathrm{del}_{Y^m}$$

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 B.1.2. $\qquad\square$

For discrete sets, the insert operation has a compact definition:

Definition 4.5.5 (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 $d_i \in D$

$$\mathrm{insert}(\mathbb{P}_\alpha^{D_{[2,n]} \wr Y_{[n-1]} | D_1}, \mathbb{P}_\alpha^{Y_{[n]} \wr D_{[n]}})(y_{[n]}, d_{[2,n]} | d_1)$$
$$= \mathbb{P}_\alpha^{Y_{[n]} \wr D_{[n]}}(y_{[n]} | d_{[n]}) \mathbb{P}_\alpha^{D_{[2,n]} \wr Y_{[n-1]} | D_1}(d_{[n]} | d_1, y_{[n-1]})$$

Inserting a comb into a comb (of appropriate dimensions) yields a conditional probability.

Theorem 4.5.6. *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,*

$$\mathrm{insert}(\mathbb{P}_\alpha^{D_{[2,n]} \wr Y_{[n-1]} | D_1}, \mathbb{P}_\alpha^{Y_{[n]} \wr D_{[n]}})$$
$$= \mathbb{P}_\alpha^{Y_{[n]} D_{[2,n]} | D_1}$$

Proof. Take $\mathsf{Y}_{[0]} = \mathsf{D}_{n+1} = *$, and

$$
\begin{aligned}
&\mathrm{insert}(\mathbb{P}_\alpha^{\mathsf{D}_{[2,n]} \wr \mathsf{Y}_{[n-1]} | \mathsf{D}_1}, \mathbb{P}_\alpha^{\mathsf{Y}_{[n]} \wr \mathsf{D}_{[n]}})(y_{[n]}, d_{[2,n]} | d_1) \\
&= \mathbb{P}_\alpha^{\mathsf{Y}_{[n]} \wr \mathsf{D}_{[n]}}(y_{[n]} | d_{[n]}) \mathbb{P}_\alpha^{\mathsf{D}_{[n]} \wr \mathsf{Y}_{[n-1]} | \mathsf{D}_1}(d_{[2,n]} | d_1, y_{[n-1]}) \\
&= \prod_{i=1}^{n} \mathbb{P}_\alpha^{\mathsf{Y}_{[i]} | \mathsf{D}_{[i]} \mathsf{Y}_{[i-1]}}(y_i | d_{[i]}, y_{[i-1]}) \mathbb{P}_\alpha^{\mathsf{D}_{i+1} | \mathsf{D}_{[i]} \mathsf{Y}_{[i-1]}}(d_i | d_{[i-1]}, y_{[i-1]}) \\
&= \mathbb{P}^{\mathsf{Y}_{[n]} \mathsf{D}_{[2,n]}}(y_{[n]}, d_{[n]} | d_1)
\end{aligned}
$$

$\square$

Aside: combs are the output of the "fix" operation

There is a relationship between combs and the "fix" operation defined in Richardson, Evans et al. (2017). In particular, suppose we have a probability $\mathbb{P}_\alpha$ and a comb $\mathbb{P}_\alpha^{\mathsf{Y}_{[2]} | \mathsf{D}_{[2]}}$. Then (assuming discrete sets)

$$
\begin{aligned}
\mathbb{P}_\alpha^{\mathsf{Y}_{[2]} \wr \mathsf{D}_{[2]}}(y_1, y_2 | d_1, d_2) &= \mathbb{P}_\alpha^{\mathsf{Y}_1 | \mathsf{D}_1}(y_1 | d_1) \mathbb{P}_\alpha^{\mathsf{Y}_2 | \mathsf{D}_2}(y_2 | d_2) \\
&= \frac{\mathbb{P}_\alpha^{\mathsf{Y}_1 | \mathsf{D}_1}(y_1 | d_1) \mathbb{P}_\alpha^{\mathsf{D}_2 | \mathsf{Y}_1 \mathsf{D}_1}(d_2 | y_1, d_1) \mathbb{P}_\alpha^{\mathsf{Y}_2 | \mathsf{D}_2}(y_2 | d_2)}{\mathbb{P}_\alpha^{\mathsf{D}_2 | \mathsf{Y}_1 \mathsf{D}_1}(d_2 | y_1, d_1)} \\
&= \frac{\mathbb{P}_\alpha^{\mathsf{Y}_{[2]} \mathsf{D}_2 | \mathsf{D}_1}(y_1, y_2, d_2 | d_1)}{\mathbb{P}_\alpha^{\mathsf{D}_2 | \mathsf{Y}_1 \mathsf{D}_1}(d_2 | y_1, d_1)}
\end{aligned}
$$

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 $\mathbb{P}_\alpha^{\mathsf{Y} \wr \mathsf{D}}$ than about the conditional probability $\mathbb{P}_\alpha^{\mathsf{Y} | \mathsf{D}}$. 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)

$$
\mathrm{Swap}_\rho \mathbb{P}_C^{\mathsf{Y} \wr \mathsf{D}} \mathrm{Swap}_{\rho-1} \neq \mathbb{P}_C^{\mathsf{Y}_\rho \wr \mathsf{D}_\rho}
$$

We can generalise IO contractibility to a notion that applies to generic Markov kernels, and do so in Section 4.5.2. 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 4.5.15 is an analogue of Theorem 4.3.21 for the case of a data-dependent model. There are two crucial differences between these theorems. First, while Theorem 4.3.21 constructs the hypothesis H as a function of the given variables, Theorem

4.5.15 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.

Secondly, Theorem 4.5.15 considers the only unconditional IO contractibility, without the "auxiliary" variable W. As a result it, it is restricted to the special case where $\mathsf{G} \perp\!\!\!\perp_{\mathbb{P}_C} (\mathsf{X}, \mathrm{id}_C)$.

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 4.5.7 (kernel locality)**.** A Markov kernel $\mathbb{K} : X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ is *local* if for all $n \in \mathbb{N}$, $A_i \in \mathcal{Y}$, $(x_{[n]}, x_{[n]^{\complement}}) \in \mathbb{N}$ there exists $\mathbb{L} : X^n \rightarrow Y^n$ such that

$$
\begin{array}{ccc}
\begin{array}{c}
\mathsf{W} \\
\mathsf{D}_{[n]} \quad \boxed{\mathbb{P}_{\alpha}^{\mathsf{Y}|\mathsf{WD}}} \quad \mathsf{Y}_{[n]} \\
\mathsf{D}_{(n,\infty)} \quad \ast
\end{array}
&
&
\begin{array}{c}
\mathsf{W} \\
\mathsf{D}_{[n]} \quad \boxed{\mathbb{P}_{\alpha}^{\mathsf{Y}|\mathsf{WD}_{[n]}}} \quad \mathsf{Y}_{[n]} \\
\mathsf{D}_{(n,\infty)} \quad \ast
\end{array}
\end{array}
$$

$$
=
$$

$$
\Longleftrightarrow
$$

$$
\mathbb{K}\Big(\bigtimes_{i \in [n]} A_i \times Y^{\mathbb{N}} \big| x_{[n]}, x_{[n]^{\complement}}\Big) = \mathbb{L}\Big(\bigtimes_{i \in [n]} A_i \big| x_{[n]}\Big)
$$

Definition 4.5.8 (kernel exchange commutativity)**.** A Markov kernel $\mathbb{K} : X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ *commutes with exchange* if for all finite permutations $\rho : \mathbb{N} \rightarrow \mathbb{N}$, $A_i \in \mathcal{Y}$, $(x_{[n]}, x_{[n]^{\complement}}) \in \mathbb{N}$

$$
\mathbb{K}\mathrm{swap}_{\rho,Y} = \mathrm{swap}_{\rho,X}\mathbb{K}
$$

$$
\Longleftrightarrow
$$

$$
\mathbb{K}\Big(\bigtimes_{i \in \mathbb{N}} A_{\rho(i)} \big| (x_i)_{i \in \mathbb{N}}\Big) = \mathbb{K}\Big(\bigtimes_{i \in \mathbb{N}} A_i \big| (x_{\rho(i)})_{i \in \mathbb{N}}\Big)
$$

IO contractibility is the conjunction of both assumptions.

Definition 4.5.9 (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.

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 4.5.11 is similar to Theorem 4.3.7, except it is stated in terms of transformations of a Markov kernel instead of in terms of conditional probabilities of variables.

Definition 4.5.10 (Marginalisation map)**.** Given a sequence $\mathsf{Y} := (\mathsf{Y}_i)_{i \in \mathbb{N}}$ and $A \subset \mathbb{N}$, marg_A is the Markov kernel associated with the function $\mathsf{Y} \mapsto (\mathsf{Y}_i)_{i \in A}$.

Theorem 4.5.11 (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*

$$
\mathbb{K}\mathrm{marg}_A = \mathrm{swap}_{A \rightarrow [n]}\mathbb{L} \otimes \mathit{del}_{X^{\mathbb{N}}}
$$

Proof. Appendix B.5 □

Lemma 4.5.12 is similar to Lemma 4.3.17, except the latter uses a variable Y^D explicitly defined on the sample space, while Lemma 4.5.12 simply says an appropriate probability distribution exists, but may not be the distribution of any variable on the given sample space.

Lemma 4.5.12. *A Markov kernel* $\mathbb{K} : X^{\mathbb{N}} \nrightarrow Y^{\mathbb{N}}$ *is IO contractible if and only if there exists a column exchangeable probability distribution* $\mu \in \Delta(Y^{|X| \times \mathbb{N}})$ *such that*

$$\mathbb{K} = \begin{array}{c} \text{W} - \boxed{\mathbb{P}_\alpha^{Y^D | W}} \\ \text{D} \underline{\hspace{3cm}} \end{array} \boxed{\mathbb{F}_{lu}} \underline{\hspace{1cm}} Y$$

$$\Longleftrightarrow$$

$$\mathbb{K}(\underset{i \in \mathbb{N}}{\bigtimes} A_i | (x_i)_{i \in \mathbb{N}}) = \mu(\underset{i \in \mathbb{N}}{\bigtimes} Y^{x_i - 1} \times A \times Y^{|X| - (x_i + 1)}) \forall A_i \in \mathcal{Y}$$

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

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

Proof. Appendix B.5. □

Lemma 4.5.13 (Exchangeable table to response functions)**.** *Given* $\mathbb{K} : X^{\mathbb{N}} \nrightarrow Y^{\mathbb{N}}$, X *and* Y *standard measurable, if*

$$\mathbb{K} = \begin{array}{c} \triangleleft\!\!\boxed{\mu} \\ \text{D} \underline{\hspace{1.5cm}} \end{array} \boxed{\mathbb{F}_{lu}} \underline{\hspace{1cm}} Y$$

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* $\mathsf{H} : Y^{X \times \mathbb{N}} \to H$ *and* $\mathbb{L} : H \times X \nrightarrow Y$ *such that*

Proof. Appendix B.5. □

Theorem 4.5.14 is similar to Theorem 4.3.21, but it is stated without the use of variables. It shows that a IO contractible Markov kernel $X^{\mathbb{N}} \nrightarrow Y^{\mathbb{N}}$ is representable as a "prior" $\mu \in \Delta(H)$ and a "parallel product" of Markov kernels $H \times X \nrightarrow Y$, which the response conditionals.

Theorem 4.5.14. *Given a kernel* $\mathbb{K} : X^{\mathbb{N}} \nrightarrow Y^{\mathbb{N}}$, *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*

$$\mathbb{K}(\underset{i\in\mathbb{N}}{\bigtimes} A_i | (x_i)_{i\in\mathbb{N}}) = \int_H \prod_{i\in\mathbb{N}} \mathbb{L}(A_i | h, x_i) \mu(\mathrm{d}h)$$

Proof. Appendix B.5. $\qquad\square$

Theorem 4.5.15 introduces a latent hypothesis variable G to allow a statement of Theorem 4.5.14 in terms of conditional probabilities instead of "anonymous" Markov kernels.

Theorem 4.5.15. *Given a sequential input-output model* $(\mathbb{P}'_C, \mathsf{D}', \mathsf{Y}')$ *on* $(\Omega, \mathcal{F})$, *then* $\mathbb{P}'^{\mathsf{Y}'\wr\mathsf{D}'}_C$ *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* $\mathsf{G} : \Omega \times G \to G$ *such that* $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{<i}, \mathsf{X}_{<i}, \mathsf{C}) | (\mathsf{X}_i, \mathsf{G})$ *and* $\mathbb{P}^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{G}}_C = \mathbb{P}^{\mathsf{Y}_j|\mathsf{X}_j\mathsf{G}}_C$ *for all* $i, j \in \mathbb{N}$ *and* $\mathsf{G} \perp\!\!\!\perp_{\mathbb{P}_C} (\mathsf{X}, id_C)$.

Proof. Appendix B.5. $\qquad\square$

4.6 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 form 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 outcomes 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, we explore a few candidates for such weaker assumptions.

While this chapter also includes a discussion of data-dependent models (Theorem 4.5.15), this work has not yet offered any easy-to-interpret equivalences between repeated response functions and model symmetries. Notably, combs play a key role in this analysis as well as the analysis of the seemingly unrelated question of identification of marginal graphical models.

Chapter 5

Causal modelling with decision models

In previous chapters we've proposed a general type for decision models and examined a particular assumption – the assumption of CIIR sequences – that allows one to predict future consequences of actions based on previously observed data. We noted that this assumption is often unreasonable. In this chapter, we consider the problem of constructing more broadly applicable decision models that license some kind of causal inference.

We show that causal Bayesian networks and potential outcomes models can both be translated to decision models. Thus any causal model constructed using either of these frameworks can be translated into a decision model which will yield the same conclusions. These are both widely used frameworks with many established results for causal identifiability and causal learning. The translations given in this chapter offer a means of importing such results to the decision modelling framework. However, we don't just want compatibility with past work – we want a framework that enables us to discover new insights in causal inference. That is the question we address in the second part of this chapter.

When translating causal Bayesian networks and potential outcomes models to decision models, there are some degrees of freedom that need to be filled with additional assumptions. Causal Bayesian networks are "rolled up" and an assumption must be made about how exactly to "unroll" it to a sequential model. Some of the additional assumptions that we regard to be required to ensure that the resulting unrolled model is faithful to the original intent of the rolled-up model are nontrivial. Potential outcomes models, on the other hand, do not offer a notion of "options", and a judgement must be made about how a potential outcomes model induces a collection of options and their associated consequences. These gaps are minor, but they do reflect the different commitments of the different approaches. These differences are summarised in the table below:

	Must	Optionally
Potential outcomes	Represent potential outcomes of every "output" variable	Represent consequences of choosing different options
Causal Bayesian networks	Represent consequences of choosing different options for "generic variables", options are interventions	Specify exactly which variables are affected by the choice
Decision models	Represent consequences of choosing different options	Options are interventions, represent potential outcomes of output variables

Section 5.1 sets out the kinds of decision models induced by causal Bayesian networks, and Section 5.2 sets out the decision models induced by potential outcomes.

The assumption of *precedent*, explored in Section 5.1.5, is motivated by the general idea that in some cases a decision maker might guess that every option they have available has already been tried by someone, somewhere, though they do not know exactly when. On its own, it doesn't allow a decision maker to conclude very much (although it does allow them to surmise that, given enough data, the consequences of their choice must have occurred at least once). However, in conjunction with the assumption of *generic relationships between conditionals*, it allows a decision maker to conclude from a conditional independence that a certain sequence of variable pairs features conditionally independent and identical responses. This result is similar in spirit to the principle underlying causal inference by invariant prediction (Peters, Bühlmann and Meinshausen, 2016), which identifies causal parents on the basis of groups of variables that render a target variable conditionally independent of an "environment" variable. However, we offer substantially different assumptions – instead of an assumption of generic relationships between conditionals, invariant prediction places restrictions on the kinds of "interventions" that can happen in each environment. The assumption of generic relationships between conditionals is itself similar to assumptions that have been used to justify inferences of causal direction (Lemeire and Janzing, 2013; Meek, 1995). Our result suggests that the assumption of precedent might facilitate understanding of the relationship between this notion of causal direction and invariance based causal inference without appealing to any notion of intervention. This is explored further in Section 5.1.5

We also investigate alternative justifications of the assumption of CIIR sequences motivated by these causal modelling frameworks. In the previous chapter, we focussed on the equivalence of certain prediction problems. In Section 5.2.1 we consider that, under some conditions, CIIR sequences can be alternatively justified by considering how different options might induce symmetric consequences. We also consider how might treat potential outcomes as "pseudo-observable" variables that represent certain things that we are practically unable to measure, but that we nonetheless might impose constraints on as if they were observed variables with certain properties.

5.1 Causal Bayesian networks

Causal Bayesian networks are a family of structural interventional models. In the form presented here, they provide a "hard intervention" operation, which offers a rule for transforming a joint probability distribution $\mathbb{P}^{\mathsf{V}}$ over a sequence of variables V according to a directed acyclic graph $\mathbf{G}$ with nodes V corresponding 1-to-1 to the variables in V. There are many variations of causal Bayesian networks with more general classes of intervention (Yang et al., 2018).

Structural Causal Models (SCMs) are an alternative class of structural interventional models. Such models distinguish a set of noises U from observed variables V, and provide a collection of functions relating the observed variables to the noises and other observed variables. The noises are allowed to be stochastic. Functional relationships can be interpreted as deterministic conditional probabilities, and it is in principle possible to formulate SCMs using our approach. However SCMs are often extended in ways that causal Bayesian networks are not, for example with counterfactual operations (Bareinboim et al., 2020), or to include cyclic causal relationships (Bongers et al., 2016; Forré and Mooij, 2020), and because addressing these extensions in detail is beyond the scope of this work, we focus on causal Bayesian networks.

In order to represent causal Bayesian networks as decision models, we have to make two generalisations. First, the variables that appear in a causal Bayesian network are "generic variables". This is different to the "observed random variables" discussed in Section 2.5.1 which are associated with measurement procedures. Generic variables are instead associated with a *type* of measurement procedure. We can have a measurement procedure to measure *your* height, and a type of measurement procedure that measures people's height. A sequence of measurements of different people's heights could be associated with a single type of measurement procedure. To make predictions about observed random variables, we need to generalise causal Bayesian networks to sequential models – this is the first generalisation.

Causal Bayesian networks are specified by a directed acyclic graph G and a probability distribution $\mathbb{P}$. However, implicit in their use is the fact that the probability distribution $\mathbb{P}$ and maybe even the graph G (or parts of it) should be learned from the given data. A decision model makes the dependence of the model on the data explicit via a hypothesis variable H. A causal Bayesian network might be thought of as a representation of a hypothesis, rather than a decision model in its own right. Constructing decision models with hypotheses corresponding to causal Bayesian networks is the second generalisation we make.

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 5.1.11 and 5.1.12). These changes don't make a meaningful difference to the content of the theory.

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 $G = (V, E)$ is a set of nodes V and edges, which are ordered pairs of nodes $E \subset V \times V$. Nodes are written using the font 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 $G = (V, E)$ and $V_i \subset V$, the parents of V_i are $\mathrm{Pa}_G(V_i) := \{V_j \in V | (V_j, V_i) \in E\}$.

We offer a recursive definition of *descendants*.

Definition 5.1.3 (Non-descendants)**.** Given a directed graph $G = (V, E)$, a descendant of V_i is a node V_j such that V_i is a parent of V_j, or there is some parent V_k of V_j that is a descendant of V_i. Any node that is a not a descendant of V_i is a non-descendant of V_i, and the set of non-descendants of V_i is denoted $\mathrm{ND}(V_i)$

A path is a sequence of edges such that the ith edge and the $i+1$th edge share exactly one node.

Definition 5.1.4 (Path)**.** Given a directed graph $G = (V, E)$, 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 ith edge is the beginning of the $i+1$th edge.

Definition 5.1.5 (Directed path)**.** Given a directed graph $G = (V, E)$, 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, $E_i = (V_k, V_l)$ implies $E_{i+1} = (V_l, V_m)$ for some $V_m \in V$.

In an acyclic graph, directed paths never reach to the same node more than once.

Definition 5.1.6 (Directed acyclic graph)**.** A directed graph $G = (V, E)$ 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 5.1.7 (Blocked path)**.** Given a DAG $G = (V, E)$, a path p is blocked by $V_A \subset V$ iff

    1. $(V_i, V_j) \in p$ and $(V_j, V_k) \in p$ for all $V_j \in V_A$

    2. $(V_j, V_i) \in p$ and $(V_j, V_k) \in p$ for all $V_j \in V_A$

    3. $(V_i, V_j) \in p$ and $(V_k, V_j) \in p$ for all $V_j \cup \mathrm{De}_{(}()V_j) \cap V_A = \varnothing$

Definition 5.1.8 (d-separation)**.** Given a DAG $G = (V, E)$, V_A is *d*-separated from V_B by V_C (all subsets of V) if V_C blocks every path starting at V_A and ending at V_B. This is written $V_A \perp_{\mathcal{G}} V_B | V_C$.

Definition 5.1.9 (Variable-node association)**.** Given a graph $G = (V, E)$ and a sequence of variables $\mathsf{V}_A := (\mathsf{V}_i)_{i \in A}$, if $|A| = |V|$ we can associate a variable with each node of the graph with an invertible map $m : \{\mathsf{V}_i | i \in A\} \to V$ that sends $\mathsf{V}_i \mapsto V)_i$. By convention, we give associated variables and nodes corresponding indices, and graphical operations are defined on variables through m, i.e. $\mathrm{Pa}(\mathsf{V}_i) := m(\mathrm{Pa}(m^{-1}(\mathsf{V}_i)))$.

Definition 5.1.10 (Compatibility)**.** Given a measurable space $(\Omega, \mathcal{F})$, a Markov kernel $\mathbb{P}. : C \rightarrowtail \Omega$ and a sequence of variables $(\mathsf{V}_i)_{i \in A}$ with $\mathsf{V}_i : \Omega \to V_i$ and a DAG $\mathcal{G}$ with nodes $\{V_i\}_{i \in A}$ and the variable-node association m, $\mathbb{P}.$ is compatible with $\mathcal{G}$ relative to m if for all $I, J, K \subset A$, $V_I \perp_{\mathcal{G}} V_J | V_K$ implies $\mathsf{V}_I \perp\!\!\!\perp_{\mathbb{P}.}^e \mathsf{V}_J | (\mathsf{V}_K, \mathrm{id}_C)$.

The following definition is reproduced from Pearl (2009) with the differences previously mentioned: notation has been matched to ours, the interventional model is assumed to be measurable and the interventional distributions are assumed to be defined on a common sample space.

Definition 5.1.11 (Interventional model)**.** An interventional model is a tuple $(\mathbb{P}_C, \Omega, (\mathsf{V}_A)$ where $(\Omega, \mathcal{F})$ is a measurable space, $\mathsf{V} := (\mathsf{V}_i)_{i \in A}$, $A \subset \mathbb{N}$ a sequence of variables with $\mathsf{V}_i : \Omega \to V_i$, and where the option set C given by

$$C := \{\mathrm{do}_\varnothing\} \cup \{(\mathrm{do}_B, v_B) | B \subset A, v_B \in \mathrm{Range}(V_B)\}$$

That is, we take every subsequence V_B of V_A and for each subsequence add to C every element of the range of V_B, each labeled with the symbol do_B.

Definition 5.1.12 (Causal Bayesian network)**.** Given an interventional model $(\mathbb{P}_C, \Omega, \mathsf{V}_A)$ and a directed acyclic graph G with nodes V, $(\mathbb{P}_C, \Omega, \mathsf{V}_A, G)$ is a *causal Bayesian network* with respect the node-variable association m if:

    1. $\mathbb{P}.$ is compatible with G with respect to m

    2. $B \neq \varnothing \implies \mathbb{P}_{(\mathrm{do}_B, v_B)}^{\mathsf{V}_B} = \delta_{v_B}$

3. $\mathbb{P}^{V_i|\mathrm{Pa}(V_i)}_{(\mathrm{do}_B, v_B)} \overset{\mathbb{P}_{(\mathrm{do}_B, v_B)}}{\cong} \mathbb{P}^{V_i|\mathrm{Pa}(V_i)}_{\mathrm{do}_\varnothing}$ for all $i \notin B$

This definition of a causal Bayesian network is agnostic about how one actually goes about constructing a model of this type. As a practical matter, if in the course of trying to to construct a causal Bayesian network one selects a set of variables V_A because they are convenient or for some other reason that's not downstream of the chosen modelling, one has to be careful the chosen variables V_A are "interventionally compatible". In particular, we require $v_{\mathrm{ND}(V_i)} \mapsto \delta_{v_i}$ be a $V_i|\mathrm{ND}(V_i)$-valid conditional for all i (Definition 2.4.2, or else the probability set for some interventions on V_i may be empty. For a contrived example, the sequence $(V_i, 2 * V_i)$ is not interventionally compatible, as at least one variable must be a non-descendant of the other, but it is not possible to set the value of one independently of the other. See also the discussion of body mass index in Example 2.4.8 for a real world example arising from a failure to perform this check.

For continuously valued variables, the fact that this definition offers the ability to pick a version of the conditional probability for each intervention is problematic. Suppose V_i is a parent of V_j, and the associated variable V_i is continuously valued and $\mathbb{P}^{V_i}_{\mathrm{do}_\varnothing}(\{v_i\}) = 0$ for all singletons $v_i \in V_i$. Then for every intervention $\mathrm{do}_{\{i\}}(v_i)$, we can choose a version of $\mathbb{P}^{V_j|V_i}_{\mathrm{do}_\varnothing}$ that takes an arbitrary value at the point v_i (because this point has measure 0), so property (3) is satisfied trivially. Some additional consistency condition seems to be required for this case, but we do not explore what it is here.

The freedom to choose versions of the conditional distributions where the "passive" distribution has no support might actually be a feature that distinguishes causal Bayesian networks from SCMs, but we don't investigate this point in detail.

5.1.2 Unrolled causal Bayesian networks

Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ and an independent and identically distributed (IID) sequence $X := (X_i)_{i \in [n]}$, it is common to "roll up" the joint distribution $\mathbb{P}^X \in \Delta(X^n)$ to a single representative distribution $\mathbb{P}^{X_0} \in \Delta(X)$ and say something like "the X_i are IID according to $\mathbb{P}^{X_0}$". Because of the IID assumption, the full joint distribution $\mathbb{P}^X \in \Delta(X^n)$ 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, though the ambiguity doesn't seem to be especially problematic for ordinary causal Bayesian networks. In Definition 5.1.13, we propose a canonical form of an unrolled causal Bayesian network.

Definition 5.1.13 (Unrolled causal Bayesian network)**.** Given an interventional model $(\mathbb{P}_., C, \Omega, V_{A \times \mathbb{N}})$ and a directed acyclic graph $\boldsymbol{G}$ with nodes V_A, $(\mathbb{P}_., C, \Omega, V_{A \times \mathbb{N}})$ is an *unrolled causal Bayesian network* with respect to the node-variable association maps $m_j : V_{ij} \mapsto V_i$ if, for all $j, k \in \mathbb{N}$:

1* $\mathbb{P}^{V_{Aj}}_.$ is compatible with $\boldsymbol{G}$ with respect to m_j for all $j \in \mathbb{N}$

2* $\pi_j(\alpha) = (\mathrm{do}_{B_j}, v_{B_j})$ and $B_j \neq \varnothing$ implies $\mathbb{P}^{V_{Bj}}_{(\mathrm{do}_{B_j}, v_{B_j})} = \delta_{v_{B_j}}$

3* If $\pi_j(\alpha) = (\mathrm{do}_{B_j}, v_{B_j})$, $\pi_k(\alpha) = (\mathrm{do}_{B_k}, v_{B_k})$ and $i \notin B_j \cup B_k$ then $\mathbb{P}^{V_{ij}|\mathrm{Pa}(V_{ij})}_\alpha \overset{\mathbb{P}_\alpha}{\cong} \mathbb{P}_{\mathrm{do}_\varnothing} V_{ik}|\mathrm{Pa}(V_{ik})$

4* $\pi_j(\alpha) = \pi_j(\alpha')$ implies $\mathbb{P}^{V_{Aj}}_\alpha = \mathbb{P}^{V_{Aj}}_{\alpha'}$

5* $V_{Aj} \perp\!\!\!\perp^e_{\mathbb{P}_C} V_{A\mathbb{N}\setminus\{j\}}|\mathrm{id}_C$

Explaining the definition of unrolled causal Bayesian networks

Suppose we have a causal Bayesian network $(\mathbb{P}_C, \Omega, \mathsf{V}_A)$ that we want to extend to a sequence of variables $\mathsf{V} := (\mathsf{V}_{ij})_{i \in A, j \in \mathbb{N}}$. We need to propose a set of interventions for the unrolled model, and ensure that we have appropriate analogues of all of the causal Bayesian network compatibility conditions that hold for each element of the sequence. There is a little ambiguity in the choice of an extended set of interventions (though it may not be problematic in practice). Unrolled causal Bayesian networks have not been defined in the literature, and there are other formal ambiguities – for example, we could omit condition 4*, but omitting this condition would allow for unrolled models that roll up to different causal Bayesian networks depending on the precise extended intervention chosen, which clashes with our understanding of what a causal Bayesian network is "supposed" to model.

First, we extend the set C to be the set of sequences of interventions

$$\{(\mathrm{do}_{B_j j}(v_B))_{j \in \mathbb{N}} | \forall j : B_j \subset A, v_{B_j} \in \mathrm{Range}(\mathsf{V}_{B_j})\}$$

i.e. C now consists of all sequences of separate interventions to each subsequence of variables $\mathsf{V}_{Aj} := (\mathsf{V}_{ij})_{i \in A}$, understood to refer to variables arising from a particular iteration of the decision procedure.

This specification of interventions in an unrolled causal Bayesian network differs slightly from the method explored in Lattimore and Rohde (2019a) which unrolls a causal Bayesian network to a length 2 sequence and forces the intervention on the first element of the sequence to be the passive intervention. This difference seems like it would usually be unproblematic insofar as the specification of which elements of the sequence might be influenced by the choices of the decision maker may often be pretty obvious.

Given a graph $G = (V, E)$, we now have a collection of variable-node association maps $m_j : \{\mathsf{V}_{ij} | i \in A\} \to V$ such that $m_j(\mathsf{V}_{ij}) = V_i$.

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 $\mathsf{V}_{Aj} := (\mathsf{V}_{ij})_{i \in A}$ are mutually independent given any particular sequence of interventions. Furthermore, Definition 5.1.12 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 jth intervention in the sequence. One might posit the following analogue of condition (3):

3' $\pi_j(\alpha) = (\mathrm{do}_{B_j}, v_{B_j})$ implies $\mathbb{P}_\alpha^{\mathsf{V}_{ij} | \mathrm{Pa}(\mathsf{V}_{ij})} \stackrel{\mathbb{P}_\alpha}{\cong} \mathbb{P}_{\mathrm{do}_\varnothing}^{\mathsf{V}_{i1} | \mathrm{Pa}(\mathsf{V}_{i1})}$ for all $i \notin B$

Where $\mathrm{do}_\varnothing^n$ is a sequence of n $\mathrm{do}_\varnothing$ interventions. This is a combination of an assumption that variables in the sequence are conditionally independent and identically distributed given appropriate interventions and condition (3) from Definition 5.1.12. However, it's not quite satisfactory. Take $B_j := \mathrm{Pa}(\mathsf{V}_{ij})$, and suppose $\mathbb{P}_{\mathrm{do}_\varnothing^n}^{B_1}(\{x\}) = 0$. Then (3') would be satisfied by a model for which

$$\mathbb{P}_{(\mathrm{do}_{B_1}, x, \mathrm{do}_{B_2}, x)}^{\mathsf{V}_{i1} | \mathsf{V}_{B_1}}(U|x) = \delta_0(U)$$
$$\mathbb{P}_{(\mathrm{do}_{B_1}, x, \mathrm{do}_{B_2}, x)}^{\mathsf{V}_{i2} | \mathsf{V}_{B_2}}(U|x) = \delta_1(U)$$

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 interventions force the variable into the region that originally had no support.

We propose instead a restricted assumption of identical response functions: for any pair V_{ij} and 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 V_{ij} given its parents is almost surely equal (with respect to $\mathbb{P}_\alpha$) to the conditional probability of 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 jth intervention does not depend on any of the interventions other than the jth. If this were not the case, then even if the restricted assumption of identical response functions were satisfied, different sequences of interventions might "roll up" to different interventional models. In particular, consider $\mathbb{P}^{B_1}_{\mathrm{do}^n_\varnothing}(\{x\}) = 0$ again, B_j as before. Then we might have, consistently with 1*-3*,

$$\mathbb{P}^{V_{i1}|V_{B_1}}_{(\mathrm{do}_{B_1},x,\mathrm{do}_{B_2},y)}(U|x) = \delta_0(U)$$

$$\mathbb{P}^{V_{i1}|V_{B_1}}_{(\mathrm{do}_{B_1},x,\mathrm{do}_{B_2},y')}(U|x) = \delta_1(U)$$

There is some freedom in choosing the conditional distribution of V_i given its parents because it has no support under the passive intervention, and without 4* this freedom allows us to make different choices when we intervene in different ways on unrelated elements of the sequence.

Condition 5* is the requirement that observations are mutually independent.

5.1.3   Causal Bayesian networks with uncertain joint distributions

Condition 3* of Definition 5.1.13 establishes that in a causal Bayesian network we can identify some input and output sequences with identical responses, with exactly which ones we can identify depending on the interventions chosen. In Chapter 4, we considered response functions that were identical *conditional on some hypothesis* H. The joint distribution $\mathbb{P}$ that appears in the previous definition of a causal Bayesian network is implicitly understood to be unknown. We can make this explicit by adding a hypothesis T and requiring that the causal Bayesian network properties hold for each value of T. We could have T index both joint distributions and directed graphs, but here we will consider the case where the graph is known in advance and only the joint distribution is not.

Definition 5.1.14 (Uncertain unrolled causal Bayesian network)**.** Given an interventional model $(\mathbb{P}_\cdot, C, \Omega, (V_{ij})_{i \in A, j \in \mathbb{N}})$ and a directed acyclic graph G, $(\mathbb{P}_\cdot, C, \Omega, (V_{ij})_{i \in A, j \in \mathbb{N}}, T, G)$ is an *uncertain unrolled causal Bayesian network* with respect to some hypothesis $T : \Omega \to T$ if for each $h \in H$, defining $\mathbb{P}_{\cdot,t} := \alpha \mapsto \mathbb{P}^{\mathrm{id}_\Omega | T}_\alpha(\cdot|t)$, $(\mathbb{P}_{\cdot,t}, C, \Omega, (V_i)_{i \in A}, G)$ is an unrolled causal Bayesian network.

Recalling the discussion in Section 3.2.1, Definition 5.1.14 associates each intervention with a unique probability distribution. One could suggest that uncertain unrolled causal Bayesian networks should therefore be called "Bayesian causal Bayesian networks". We could also consider models with a non-stochastic hypothesis H which we might call "non-Bayesian causal Bayesian networks".

An uncertain unrolled causal Bayesian network *almost* features a number of CIIR input-output sequences. Due to 3*, such a model features conditionally independent and identical

response functions with inputs $\mathsf{Pa}(\mathsf{V}_i)$ and outputs V_i wherever α consists of a sequence of interventions none of which target V_{ij} for any j. 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 parameter H) by application of Theorem 4.3.21.

Theorem 5.1.15 (IO contractibility of CBNs). *Given an uncertain unrolled causal Bayesian network $(\mathbb{P}_\cdot, C, \Omega, (\mathsf{V}_{ij})_{i \in A, j \in \mathbb{N}}, \mathsf{H}, \boldsymbol{G})$, take $C' \subset C$ to be sequences of interventions that, for some $i \in A$, do not target a particular V_{ij} for any $j \in \mathbb{N}$ and ensure every sequence $\mathsf{V}_{j\mathbb{N}}$ has infinite support. Then $\mathsf{V}_{i\mathbb{N}} \perp\!\!\!\perp^e_{\mathbb{P}_{C'}} id_C | (\mathsf{H}_i, \mathsf{Pa}(\mathsf{V}_{i\mathbb{N}}))$ and $\mathbb{P}_C^{\mathsf{V}_{i\mathbb{N}} | \mathsf{H}_i \mathsf{Pa}(\mathsf{V}_{i\mathbb{N}})}$ is IO contractible over H_i where H_i is the directing random conditional with respect to $(\mathbb{P}_C, \mathsf{Pa}(\mathsf{V}_{i\mathbb{N}}), \mathsf{V}_{i\mathbb{N}})$.*

Proof. Appendix C.1. □

5.1.4 Probabilistic Graphical Models

Lattimore and Rohde (2019a,b) have demonstrated how to "unroll" causal Bayesian networks into what they call "Probabilistic Graphical Models". Their construction is very similar to ours, as we will show here, and for readers interested in how identifiability results from regular causal Bayesian networks translate to our scheme, Lattimore and Rohde's work provides several examples.

Precisely, a probabilistic graphical model is a map $\mathbb{P}_\cdot$ from the set of single-node interventions C to probability distributions $\mathbb{P}_\alpha$ defined on $(\Omega, \mathcal{F})$. A probabilistic graphical model is typically associated with a causal Bayesian network $(\mathbb{Q}_\cdot, C, \Omega', (\mathsf{V}_i)_{i \in A}, \boldsymbol{G})$ where, for each $\mathsf{V}_i : \Omega \to V_i$ in the original causal Bayesian network, two variables V_i and V_i^* are defined on $(\Omega, \mathcal{F})$.

The probabilistic graphical model also adds a "parameter" W_i for each variable pair $(\mathsf{V}_i, \mathsf{V}_i^*)$ such that, taking C' to be interventions not targeting V_i^*, for any $\alpha \in C'$, $\mathbb{P}_\alpha^{\mathsf{V}_i^* | \mathsf{W}_i \mathsf{Pa}(\mathsf{V}_i)} = \mathbb{P}_\alpha^{\mathsf{V}_i^* | \mathsf{W}_i \mathsf{Pa}(\mathsf{V}_i^*)}$ and $\mathsf{V}_i \perp\!\!\!\perp^e_{\mathbb{P}_{C'}} (\mathsf{V}_A^*, id_C) | (\mathsf{W}_i)$ (where parents are assessed relative to the graph $\boldsymbol{G}$). W_i serves precisely the same role as H_i in Theorem 5.1.15, except it is not defined in terms of the directing random conditional.

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 $\boldsymbol{G}$ associated with $(\mathbb{Q}_\cdot, C, \Omega', (\mathsf{V}_i)_{i \in A}, \boldsymbol{G})$:
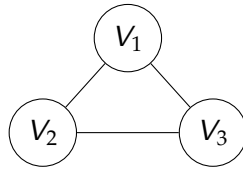
Here is the probabilistic graphical model associated with the intervention (do_2, v_2)
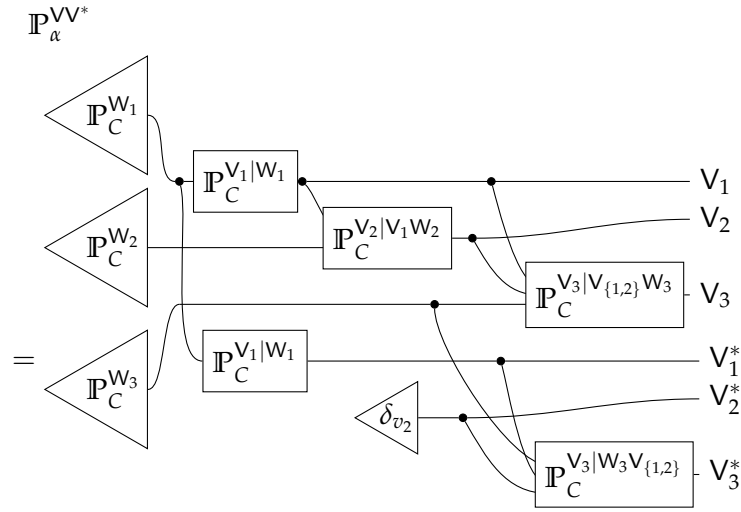
and here is the uncertain unrolled CBN associated with the restricted set of interventions C' that consists of, for each element of the sequence, either the empty intervention or some intervention targeting V_2
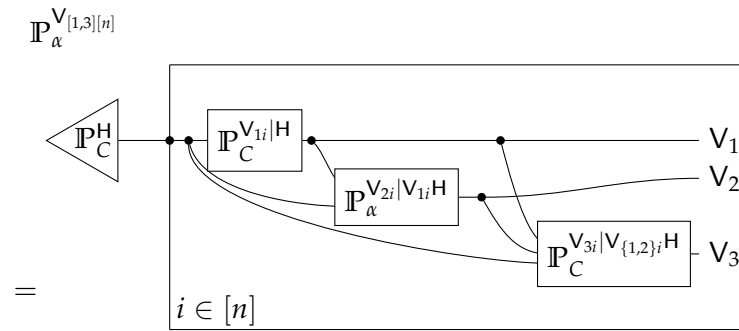
where

$$\mathbb{P}_\alpha^{V_{2i}|V_{1i}H} = \begin{cases} \delta_v & \pi_i(\alpha) = (\mathrm{do}_2, v) \\ \mathbb{P}_{mathrmdo_\varnothing}^{V_{2i}|V_{1i}H} & \text{otherwise} \end{cases}$$

5.1.5 Unobserved confounders and precedent

As we have pointed out, the assumption of CIIR sequences is often unreasonable. Causal Bayesian networks "almost" make this assumption with respect to inputs $\mathrm{Pa}(V_{i\mathbb{N}})$ and outputs $V_{i\mathbb{N}}$. One of the ways that this appraoch gets around the fact that the assumption is unreasonable is to assume that some elements of $\mathrm{Pa}(V_{i\mathbb{N}})$ are unobserved. In this case, the "interchangeable conditioning sequences" are never actually observed, and so the question of whether or not they are interchangeable is moot.

We are limiting our attention to data-independent models (recall Definition 4.3.14), which means that unobservability of some variable does not have any implications within the model

– only that it isn't attached to a measurement procedure. If we were considering a data-dependent variation of causal Bayesian networks, the fact that $V_{1\mathbb{N}}$ is unobserved may have implications within the model – for example, that input D_i may not depend on V_{1j} for any j.

Unobserved variables in the set of parents of a particular variable of interest may be called *unobserved confounders*[1]. Suppose we have an uncertain unrolled causal Bayesian network $(\mathbb{P}., C, \Omega, (V_{ij})_{i \in [3], j \in \mathbb{N}}, H, \boldsymbol{G})$ where the graph $\boldsymbol{G}$ is as follows:
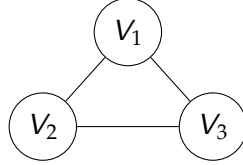
and we consider the subset $C' \subset C$ of interventions that are either empty or target V_2 only. We note that Theorem 5.1.15 implies that $\mathbb{P}_C^{V_{3\mathbb{N}}|H_3 V_{1\mathbb{N}} V_{2[n]}}$ is IO contractible, but not $\mathbb{P}_C^{V_{3\mathbb{N}}|H_3 V_{2\mathbb{N}}}$. We declare that $V_{1[n]}$ is not observed – that is, it is not associated with a measurement procedure.

We observe in Theorem 5.1.16 that the IO contractibility of $\mathbb{P}_C^{V_{3\mathbb{N}}|H V_{1\mathbb{N}} V_{2\mathbb{N}}}$ implies that $\mathbb{P}_C^{V_{3\mathbb{N}}|H V_{2\mathbb{N}}}$ is unchanged by any swaps that leave the value of the full vector $V_{1\mathbb{N}}$ unchanged. That is, the assumption of an unobserved confounder in this case implies a kind of "partial IO contractibility".

Theorem 5.1.16. *Given* $(\mathbb{P}., C, \Omega, (V_{ij})_{i \in [3], j \in \mathbb{N}}, H, \boldsymbol{G})$ *with* $\mathbb{P}_C^{V_{3\mathbb{N}}|H_3 V_{1\mathbb{N}} V_{2\mathbb{N}}}$ *IO contractible over* H_3, V_i *discrete for all* $i \in [3]$ *and* $(V_{1\mathbb{N}}, V_{2\mathbb{N}})$ *infinitely supported over* H_3, *let* $Q : \Omega \to \mathbb{N}^{\mathbb{N}}$ *be a random finite permutation of* $\mathbb{N}$ *such that* $V_{1\mathbb{N}} \overset{\mathbb{P}_C}{\cong} V_{1Q(\mathbb{N})}$. *Then*

$$\mathbb{P}_C^{V_{3\mathbb{N}}|H_3 V_{2\mathbb{N}}} \overset{\mathbb{P}_C}{\cong} \mathbb{P}_C^{V_{3Q(\mathbb{N})}|H_3 V_{2Q(\mathbb{N})}}$$

Proof. By IO contractibility of $\mathbb{P}_C^{V_{3\mathbb{N}}|H V_{1\mathbb{N}} V_{2\mathbb{N}}}$ over H_3

$$\mathbb{P}_C^{V_{3\mathbb{N}}|H_3 V_{1\mathbb{N}} V_{2\mathbb{N}}} = \mathbb{P}_C^{V_{3Q(\mathbb{N})}|H_3 V_{1Q(\mathbb{N})} V_{2Q(\mathbb{N})}}$$
$$= \mathbb{P}_C^{V_{3Q(\mathbb{N})}|H_3 V_{1\mathbb{N}} V_{2Q(\mathbb{N})}}$$

---

[1]To be a confounder, we also require an observed parent to be a descendant of the unobserved parent, but this detail isn't important for this discussion

Thus

$$\mathbb{P}_C^{V_{3\mathbb{N}}|HV_{2\mathbb{N}}}$$

$$= \mathbb{P}_C^{V_{3Q(\mathbb{N})}|HV_{2Q(\mathbb{N})}}$$

□

Theorem 5.1.16 says that IO contractibility of $\mathbb{P}_C^{V_{3\mathbb{N}}|HV_{1\mathbb{N}}V_{2\mathbb{N}}}$ implies the "exchange commutativity" of $\mathbb{P}_C^{V_{3\mathbb{N}}|HV_{2\mathbb{N}}}$ with respect to the random permutation Q of the indices of $\mathbb{N}$ – in fact, for all such random permutations that preserve $V_{1\mathbb{N}}$. Note that Q generally cannot be deduced from observations, because $V_{1\mathbb{N}}$ cannot be deduced from observations. That is, $V_{1\mathbb{N}}$ determines some subset of indices that are interchangeable, even though not *all* indices are interchangeable.

Suppose $V_{[3]n}$ is the result of some intervention targeting V_{2n}, and $V_{[3]\{n\}^\complement}$ are all the result of passive observation. We can interpret Theorem 5.1.16 as expressing the idea that there is some (unknown) random subset of $\{n\}^\complement$ that is "just like" the intervened sample $V_{[3]n}$. We could say that this random subset is a *precedent* for the intervention.

Thus, using causal Bayesian networks we can express the idea that some intervention has precedent in the observed data, and indeed such an assumption appears to be implicit in the notion of unobserved parents in a causal Bayesian network. This raises the question: do we *need* causal Bayesian networks to express this notion? The sentiment of this assumption long predates formal theories of causation. In fact, we can find it in the Bible:

> Everything that happens has happened before; nothing is new, nothing under the sun. Ecclesiastes 1:9, Contemporary English Version

We offer a simple definition of precedent. The idea is that we're given some sequence $Z := (Z_i)_{i\in\mathbb{N}}$ and our choice influences Z_n and, no matter what we do, there is some subsequence Z_A of $Z_{\{n\}^\complement}$ such that $Z_{A\cup\{n\}}$ is exchangeable (A may, in general, depend on the option chosen). We propose the following properties to precisely characterise precedent:

Definition 5.1.17 (Precedent). A decision model $(\mathbb{P}., (\Omega, \mathcal{F}), (C, \mathcal{C}))$ along with a sequence of variables $Z := (Z_i)_{i\in\mathbb{N}}$ is *n-precedented* if

1. The sample space can be extended with a discrete variable $T := (T_i)_{i\in\mathbb{N}}$ such that for every $\alpha \in C$, each T_i indexes an exchangeable subsequence to which the corresponding Z_i belongs; i.e. $\mathbb{P}_\alpha^{Z|T}$ is exchange commutative for all $\alpha \in C$ (over $*$)

2. Neither $Z_{\{n\}^\complement}$ nor $T_{\{n\}^\complement}$ depend on the choice α; $(Z_{\{n\}^\complement}, T_{\{n\}^\complement}) \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathrm{id}_C$

3. $(Z_i, T_i)_{i\in\{n\}^\complement}$ is exchangeable

4. T is infinitely supported over to the $(\mathbb{P}_C, \mathsf{T}, \mathsf{Z})$ directing conditional H

5. $\mathbb{P}_\alpha^{\mathsf{T}_n}(t) > 0$ for $t \in T$

Property (1) guarantees that for every available choice, Z can be partitioned into exchangeable subsequences, and we identify which exchangeable subsequence Z_i belongs to with T_i. Property (2) identifies $(\mathsf{Z}_{\{n\}^{\complement}}, \mathsf{T}_{\{n\}^{\complement}})$ as passive observations that the decision maker can't affect. Property (3) assumes for convenience that these observations are exchangeable, a common assumption to make of observations. Properties (4) and (5) are convenient and (as far as we can tell) do no harm.

From this, it follows that, for H the directing random conditional of $(\mathbb{P}_C, \mathsf{T}, \mathsf{Z})$, $\mathbb{P}_C^{\mathsf{Z}|\mathsf{HT}}$ is IO contractible.

Theorem 5.1.18 (IO contractibility of precedented decision models)**.** *Given* $(\mathbb{P}_\cdot, (\Omega, \mathcal{F}), (C, \mathcal{C}))$, *if* $\mathsf{Z} := (\mathsf{Z}_i)_{i \in \mathbb{N}}$ *is n-precedented with indexing variable* T, *then taking* H *to be the directing random conditional of* $(\mathbb{P}_C, \mathsf{T}, \mathsf{Z})$, $\mathsf{Z} \perp\!\!\!\perp_{\mathbb{P}_C}^e id_C | (\mathsf{H}, \mathsf{T})$ *and* $\mathbb{P}_C^{\mathsf{Z}|\mathsf{HT}}$ *is IO contractible.*

Proof. By assumption, $\mathbb{P}_\alpha^{\mathsf{Z}|\mathsf{T}}$ is exchange commutative. In addition, also by assumption, $\mathbb{P}_\alpha^{\mathsf{T}_{\{n\}^{\complement}}}$ is exchangeable. But then $\mathbb{P}_\alpha^{\mathsf{T}}$ is exchangeably dominated, and the result follows from Theorem 4.3.24. $\qquad\square$

Thus precedent implies "garden variety" IO contractibility of $\mathbb{P}_C^{\mathsf{Z}|\mathsf{HT}}$. Apart from the stronger assumptions about observations, the only difference to the discussion of IO contractibility in Chapter 4 is in our interpretation; we don't assume that T is observed, and we don't assume that we know anything in advance about how T responds to α.

The assumption of precedent as given in Definition 5.1.17 can, under some side conditions, yield nontrivial conclusions, as we show in Theorem 5.1.21. Before we state this theorem, it is worth going through an example first.

Example 5.1.19. Suppose we have a collection of doctors who each see a series of patients, 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 these decisions could be stochastic or deterministic. The decision maker is *also* a doctor, and is reviewing the data contained in the sequences $(\mathsf{Z}_i, \mathsf{X}_i, \mathsf{Y}_i)_{i \in [n]}$, where Z_i identifies the doctor involved in the ith treatment interaction. The decision maker supposes that whatever overall treatment plan the decision maker will adopted (which could and probably does also involve features not listed in this set of variables), the same thing has probably been done at least sometimes by some of these other doctors – that is, their treatment protocol is precedented in the data. Because the other doctors have some variation in their treatment behaviour, it stands to reason that different doctors making the same prescription decisions should see different results *if the different unobserved treatment plans actually lead to different results*. Conversely, if there is *no* variation in results different doctors obtain, then whether or not treatment occurred is the *only* important feature of any treatment plan.

This story might fail if the doctors all knew exactly the long-run probabilistic outcomes of different treatment plans and coordinated with one another to mask any variation they induced. For example, doctor 1 picks a medium effectiveness unobserved plan 100% of the time, while doctor 2 picks a highly effective unobserved plan 50% of the time and a low effectiveness unobserved plan 50% of the time, leading to the same distribution over outcomes. Our decision maker also has to assume that there is some "randomness" in each doctor's choice

of unobserved treatment plan so that they cannot precisely coordinate in this manner. In particular, one doctor's hidden treatment plan cannot be a deterministic function of any other doctor's hidden treatment plan and the "true" effectiveness of the treatment plans.

Nxample 5.1.20 (Matrix notation)**.** Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ with D, Y discrete, the directing random conditional H takes values in the set of Markov kernels $D \to Y$, which can be identified with a subset of matrices in $\mathbb{R}^{|D| \times |Y|}$. We can therefore refer to elements of H as matrix elements $(\mathsf{H}_d^y)_{d \in D, y \in Y}$ with $\sum_{y \in Y} \mathsf{H}_d^y \overset{\mathbb{P}_C}{\cong} 1$ for all d, and $\mathsf{H}_d^y \overset{\mathbb{P}_\alpha}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{HD}}(y|h,d)$.

Theorem 5.1.21. *Suppose we have a probability set* $\mathbb{P}_C$ *on* $(\Omega, \mathcal{F})$ *with variables* $\mathsf{Y} :=$ $(\mathsf{Y}_i)_{i \in \mathbb{N}}$, $\mathsf{X} := (\mathsf{X}_i)_{i \in \mathbb{N}}$ *and* $\mathsf{Z}_i := (\mathsf{Z}_i)_{i \in \mathbb{N}}$, *with* X, Y, Z *discrete and n-precedented with indexing variable* T. *Let* G *be the directing random conditional of* $(\mathsf{P}_C, \mathsf{T}, \mathsf{Z})$ *and* H *the directing random conditional of* $(\mathbb{P}_C, (\mathsf{T}, \mathsf{Z}), (\mathsf{X}, \mathsf{Y}))$. *Suppose there is some* $N \in \mathcal{H}$ *such that* $\mathbb{P}_\alpha^{\mathsf{H}}(N) > 0$ *for all* α, *and define the probability conditioned on* $\mathsf{H} \in N$:

$$probP_{\alpha, \mathsf{H} \in N}(A) := \frac{\mathbb{P}_\alpha^{id_\Omega \mathsf{H}}(A \times N)}{\mathbb{P}_\alpha^{\mathsf{H}}(N)}$$

Suppose for all i, $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_{C, \mathsf{H} \in N}}^e \mathsf{Z}_i | (\mathsf{H}, \mathsf{X}_i)$, *and for all* $t \in T$ *and* $z, z' \in Z$,

$$\mathbb{P}_{\alpha, \mathsf{H} \in N}^{\mathsf{G}_z^d | \mathsf{HG}_{z'}^d}(\cdot | h, g_{z'}^d) \ll U_{[0,1]} \tag{5.1}$$

where $U_{[0,1]}$ *is the uniform probability measure on* $([0,1], \mathcal{B}([0,1]))$.

Then $\mathbb{P}_{\alpha, \mathsf{H} \in N}^{\mathsf{Y}|\mathsf{XH}}$ *is IO contractible over* H.

Proof. We show that the assumption of conditional independence imposes a polynomial constraint on G_z^d which is nontrivial unless $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_{C, \mathsf{H} \in \mathbb{N}}}^e (\mathsf{Z}_i, \mathsf{T}_i, \mathrm{id}_C) | (\mathsf{H}, \mathsf{X}_i)$, and hence the solution set S for this constraint is empty when this conditional independence does not hold.

Full proof in Appendix C.2. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\Box$

Theorem 5.1.21 provides some support for the kind of inference explained in Example 5.1.19. We offer the assumption of precedent as a formalisation of the idea that "whatever the decision maker does, it's been done before". If the decision maker can conclude that the hypothesis H lies in a set N that implies the outcome is independent of the doctor identifiers Z_i given the treatment decisions X_i, and further believes that within the set N, there is nevertheless zero probability that doctors coordinate precisely enough to mask any variation due to unobserved differences in treatment protocol, then the decision maker can conclude that the observed treatment-outcome map is identical to the treatment-outcome map they should use to assess the decision of whether or not to treat.

A substantial shortcoming of Theorem 5.1.21 is the fact that, if the distribution of H is dominated by the uniform measure on $[0,1]^{|T||Z||X||Y|}$, then $\mathsf{H} \in N$ will also be a set of measure 0, as the conditional independence condition required by N is itself a nontrivial polynomial restriction on values of H. Thus assuming some continuous distribution of H does not satisfy the conditions required by the theorem.

This is reminiscent of problems encountered by other conditional independence based approaches to causal inference. In particular, we draw conclusions from conditional independence – a measure 0 event under a continuous distribution of H – but also from the assumption

that *other* measure 0 events under continuous distributions can be neglected. One way that this has been addressed is to treat approximate conditional independence as if it was conditional independence (Zhang and Spirtes, 2003).

The appropriate strengthening of the assumptions in Example 5.1.19 seems to require not only that the other doctor's actions aren't perfectly coordinated, but that we can further place a limit on how precisely they might be approximately coordinated. In this case, insufficient variation in overall outcomes between doctors might still license the conclusion that the unobserved treatment variations *probably* don't have *much* influence. Strengthening the required conditions like this has substantial practical implications; our decision maker is interested in Y presumably because they are interested in influencing Y. However, for the same reason, all of the other doctors may well be interested in influencing Y in exactly the same way. Thus, while it's unlikely they can all coordinate *perfectly*, they may well achieve a good approximation of coordination, which, unless the decision maker has a very large amount of data, might be enough to prevent them from drawing conclusions from an apparent lack of variation in outcomes between doctors.

Technically, we could consider an ϵ-enlargement of the set of solutions S constructed in the proof of Theorem 5.1.21, and place an upper limit on the probability of parameters falling into this set. It may then be possible to conclude a similar result from a suitably precise approximation of the assumed conditional independence, but demonstrating this remains an open question.

Assumptions similar to Eq. (5.1) appear in two other capacities in the causal inference literature. First, in justifications for *causal faithfulness* it is assumed that, if we have variables ordered according to their structural causal relationships then an assumption of this type holds Meek (1995). Secondly, according to the principle of *independence of conditionals*, researchers have proposed a number of measures of "genericity" that are maximised when conditionals are computed in a manner compatible with their structural causal relationships (Lemeire and Janzing, 2013). The common thread here is that an assumption like Eq. (5.1) is considered to be implied by the direction of causal relationships in a structural model. Our result is somewhat different – we start with the assumption of a generic relationship between conditionals, rather than a causal direction, and draw a conclusion of identical responses. These approaches might be combined – under the right conditions, assuming or inferring a causal direction might be sufficient to apply Theorem 5.1.21, though exactly when such reasoning is valid is an open question.

5.2   Potential Outcomes models

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 sequential and they feature no notion of "intervention". A third difference relates to the possibility of expressing "counterfactual" statements, (although this difference may be contingent on the particular manner we use to unroll a causal Bayesian network).

To formulate a decision making model from a potential outcomes model, we do have to make a judgement about what the "options" are (CBNs provide the notion of "intervention" for this role), but we do not need to make any judgements about how to unroll a potential outcomes model. For the following, we rely on Rubin (2005) for the definition of a potential outcomes model.

Our definition of potential outcomes is formally similar to the tabulated conditional distribution (Definition 4.3.10), but not quite identical: in particular, for any $d \in D$, a potential outcomes model holds that $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{Y}_i^D \mathsf{D}_i}(y^d|y^D, d) = 1$, but this is generally false for the tabulated conditionals in Definition 4.3.10.

Definition 5.2.1 (Potential outcomes). Given $(\mathbb{P}_C, \Omega, \mathcal{F})$ and, for some i, variables $\mathsf{D}_i : \Omega \to D$ (D denumerable), $\mathsf{Y}_i : \Omega \to Y$ and $\mathsf{Y}_i^D : \Omega \to Y^D$, Y_i^D is a vector of *potential outcomes* with respect to D_i for all α

$$\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{Y}_i^D \mathsf{D}_i} = \begin{array}{l} \mathsf{Y}_i^D \\ \mathsf{D}_i \end{array} \!\!\!\!\!\!\!\!\!\!\!\! \boxed{\mathbb{F}_{\text{lus}}} \!\!\!\!\!\! \mathsf{Y}_i$$

Where $\mathbb{F}_{\text{lus}}$ is the Markov kernel associated with the single-shot lookup map

$$\text{lus} : Y^D \times D \to Y$$
$$(d, (y_i)_{i \in D}) \mapsto y_d$$

Note that $|D|$ copies of Y_i ($\mathsf{Y}_i, \mathsf{Y}_i, \mathsf{Y}_i, ... \mathsf{Y}_i$) 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 some probabilistic model with potential outcomes. Traditionally, potential outcomes models did not feature any set of options, and so are modeled by a single probability distribution. That is, a "traditional" potential outcomes model is a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ rather than a probability function, but our definition also allows for decision models with potential outcomes.

Definition 5.2.2 (Potential outcomes model). $(\mathbb{P}_C, \Omega, \mathcal{F})$ is a potential outcomes model with respect to $\mathsf{Y}^D := (\mathsf{Y}_i^D)_{i \in A}$, $\mathsf{Y} := (\mathsf{Y}_i)_{i \in A}$ and $(\mathsf{D}_i)_{i \in A}$ if Y_i^D is a vector of potential outcomes with respect to D_i and Y_i for all $i \in A$.

Theorem 5.2.3. *A potential outcomes model $(\mathbb{P}_C, \Omega, \mathcal{F})$ with respect to $\mathsf{D}_i : \Omega \to D$, $\mathsf{Y}_i : \Omega \to Y$ and $\mathsf{Y}_i^D : \Omega \to Y^D$, Y_i^D features the conditional independence $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_C}^e id_C|(\mathsf{D}, \mathsf{Y}^D)$, and $\mathbb{P}_C^{\mathsf{Y}|\mathsf{Y}^D \mathsf{D}}$ is IO contractible (with respect to $*$).*

Proof. IO contractibility of follows from the fact that Y_i is deterministic given Y_i^D and D_i, and thus $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (\mathsf{D}_{\{i\}^\complement}, \mathsf{Y}_{\{i\}^\complement}, id_C)|(\mathsf{Y}_i^D, \mathsf{D}_i)$. Furthermore, for all i, j, α

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

hence for all α, (D, Y) is a CIIR sequence of inputs and outputs, and hence $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{Y}^D \mathsf{D}}$ is IO contractible for all α.

Furthermore, from Definition 5.2.1, $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{Y}_i^D \mathsf{D}_i}$ is the same for all $\alpha \in C$, and by the argument above,

$$\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{Y}_i^D \mathsf{D}_i \mathsf{Y}_{\{i\}^\complement}^D \mathsf{D}_{\{i\}^\complement}} = \mathbb{P}_C^{\mathsf{Y}_i|\mathsf{Y}_i^D \mathsf{D}_i} \otimes \text{Del}_{Y^{|D \times A \setminus \{i\}|} \times D^{|A|}}$$

hence

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{Y}^D \mathsf{D}} = \bigotimes_{i \in A} \mathbb{P}_C^{\mathsf{Y}_1|\mathsf{Y}_1^D \mathsf{D}_1}$$

hence $Y \perp\!\!\!\perp^e_{\mathbb{P}_C} \text{id}_C | (D, Y^D)$. □

A key theorem of potential outcomes is that, if D is "strongly ignorable" with respect to Y^D, then the average treatment effect is identified. "Strong ignorability" here means that the probability $\mathbb{P}^{D_i}_\alpha(d) > 0$ for each d and for each choice α the inputs D are independent of the potential outcomes 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 our generalisation of potential outcomes to decision models, not only to probability distributions featuring potential outcomes.

Theorem 5.2.4 (Potential outcomes identifiability). *Suppose $(\mathbb{P}_C, \Omega, \mathcal{F})$ is a potential outcomes model with respect to $Y^D := (Y^D_i)_{i \in \mathbb{N}}$, $Y := (Y_i)_{i \in \mathbb{N}}$ and $D := (D_i)_{i \in \mathbb{N}}$, and further suppose there is some $X := (X_i)_{i \in \mathbb{N}}$ such that $\mathbb{P}^{Y^D X}_\alpha$ is exchangeable for all α, $D \perp\!\!\!\perp^e_{\mathbb{P}_C} Y^D | (X, Y, \text{id}_C)$, (D, X) is infinitely supported and for each α $\mathbb{P}^{DX}_\alpha$ is absolutely continuous with respect to some exchangeable distribution in $\Delta((D \times X)^{\mathbb{N}})$. Then there is some W such that for all α $\mathbb{P}^{Y|WXD}_\alpha$ is IO contractible over W.*

Proof. By exchangeability of $\mathbb{P}^{Y^D X}_\alpha$, $\mathbb{P}^{Y^D|X}_\alpha$ commutes with exchange. Because Y is deterministic given D and Y^D, $Y \perp\!\!\!\perp^e_{\mathbb{P}_C} (X, \text{id}_C) | (Y^D, D)$ Thus, for some finite permutation ρ, by IO contractibility of $\mathbb{P}^{Y|Y^D D}_C$

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

IO contractibility of $\mathbb{P}^{Y|WXD}_\alpha$ over some W follows from Theorem 4.3.24. □

5.2.1 Justifying identifiability assumptions

In the previous chapter we investigated some justifications for IO contractibility (which, roughly speaking, is the same as "identifiability" in other causal modelling frameworks). Theorem 5.2.4 offers an alternative collection of assumptions and yields a similar conclusion. An obvious question to ask is what the justifications for these assumptions are, and how they relate to the justifications in the previous chapter.

We think the justifications given for these assumptions are typically quite different to those discussed in Section 4.4.2. We focus on the two assumptions of exchangeability of potential outcomes ($\mathbb{P}^{Y^D X}_\alpha$ is exchangeable in Theorem 5.2.4), and "ignorability" ($D \perp\!\!\!\perp^e_{\mathbb{P}_C} Y^D | (X, Y, \text{id}_C)$ in Theorem 5.2.4). Two means of justifying these assumptions involve considering changes to an experiment that one believes would yield identical outcome distributions, or treating the potential outcomes as if they were variables that we could measure in principle, and asking what properties these variables might have.

Different actions that yield the same outcome distribution

Greenland and Robins (1986) explain an assumption akin to the conjunction of the exchangeability of potential outcomes and ignorability:

> 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.

This assumption is evocative of the assumption of commutativity of exchange (Definition 4.3.5). The latter *seems* to say something like "swapping the inputs leaves the outputs unchanged up to permutation". However, strictly speaking all it asserts is that a conditional probability is unchanged under swaps of the labels, and says nothing about swapping "individuals".

One way we could interpret Greenland and Robins' assumption (and by no means do we think this interpretation is obligatory) is to suppose that the decision maker's option set C genuinely affords them the opportunity to swap treatments. We might consider, more generally, that there is some α, α' such that for some finite permutation ρ

$$\mathbb{P}_\alpha^{\mathsf{D}} = \mathbb{P}_{\alpha'}^{\mathsf{D}_\rho}$$

and, furthermore, these permuted inputs lead to permuted outputs

$$\mathbb{P}_\alpha^{\mathsf{D}} \odot \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}} = \mathbb{P}_{\alpha'}^{\mathsf{D}_\rho} \odot \mathbb{P}_{\alpha'}^{\mathsf{Y}_\rho|\mathsf{D}_\rho} \tag{5.2}$$

This condition is actually independent of commutativity of exchange. For simplicity we'll consider deterministic examples. First, suppose $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}}$ maps everything to an infinite sequence of 1s, while $\mathbb{P}_{\alpha'}^{\mathsf{Y}|\mathsf{D}}$ maps everything to an infinite sequence of 2s. Both are clearly exchange commutative, but, if the marginals of D were appropriately related, would not satisfy Equation (5.2).

On the other hand, suppose $\mathbb{P}_\alpha^{\mathsf{Y}_1|\mathsf{D}_1}$ sends $d_1 \mapsto -d_1$ and $\mathbb{P}_\alpha^{\mathsf{Y}_{N\setminus\{1\}}|\mathsf{D}_{N\setminus\{1\}}}$ sends $d_{N\setminus\{1\}} \mapsto d_{N\setminus\{1\}}$. On the other hand $\mathbb{P}_{\alpha'}^{\mathsf{Y}_2|\mathsf{D}_2}$ sends $d_2 \mapsto -d_2$ and $\mathbb{P}_{\alpha'}^{\mathsf{Y}_{N\setminus\{2\}}|\mathsf{D}_{N\setminus\{2\}}}$ sends $d_{N\setminus\{2\}} \mapsto d_{N\setminus\{2\}}$. If $\mathbb{P}_\alpha^{\mathsf{D}_1\mathsf{D}_{N\setminus\{1\}}} = \mathbb{P}_{\alpha'}^{\mathsf{D}_2\mathsf{D}_{N\setminus\{2\}}}$ then Equation (5.2) will be satisfied but exchange commutativity generally will not.

It's hard to think of a practical example of the first possibility, but for the second possibility we can imagine an experiment where we can either let the first patient choose their treatment and compel the second patient, or vise-versa. Up to permutation, both yield the same joint distribution of treatments and outcomes, but this joint distribution will generally not be exchangeable. Note that Greenland and Robins do not consider stochastic swaps of individuals' treatments, so this objection doesn't apply to their original scenario. Having said that, if stochastically influencing treatment is all we can actually do, then their original assumption requires some alternative interpretation to the one we have provided here.

The two conditions under discussion coincide if we additionally assume $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathrm{id}_C|\mathsf{D}$. In that case, we have

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

which, under a support condition together with Eq. (5.2) implies

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

The independence $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathrm{id}_C | D$ allows us to conclude exchange commutativity from the symmetries between the consequences brought about by different options.

Potential outcomes as "pseudo-observables"

Rubin (2005) writes in defense of the exchangeability of potential outcomes:

> 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)[...]the indexing of the units is, by definition, a random permutation of $1, ..., N$, and thus any distribution on the science must be row-exchangeable

We interpret Rubin to mean something like this:

1. The properties of "unit i" – some real-world thing – determine the values of all of the variables indexed by i, and the potential outcomes ("the science") of unit i is one of its properties

2. The assignment of indices to units was either literally performed by some physically random process, or at least if it were the model in question would be exactly the same

3. Thus we should adopt a model that is invariant to the operation of shuffling all of the indices assigned to the units, or equivalently invariant to the operation of shuffling the indices assigned to sets of variables

This argument doesn't quite go through. Recall how, in Section 4.4.3, we mentioned that a decision maker may often know the impact of their choice on the inputs D before examining the data. Note also in Definition 5.1.13 (and it's descendants) that a CBN model permits us to choose interventions that have different effects on variables of the same type with different indices. Perhaps a decision maker knows a priori how to fix the value of D_5, but has no influence over $D_{[4]}$, or (essentially equivalently) they can intervene on D_5 but not $D_{[4]}$. Any model with this property is not invariant to swapping the index 5 with any of the prior indices. Similarly, if a decision maker can control the entire sequence $D_{[5]}$, they may in general choose different distributions to set each D_i to, which will again make the model change under permutations of $D_{[5]}$.

One could replace claim 3 with something like:

3' The distribution of potential outcomes is the same no matter what choice the decision maker makes

4 The distribution of variables that are insensitive to the decision maker's choices should be invariant to the operation of shuffling indices

To the extent that potential outcomes model counterfactual relationships, 3' is a strong claim about the nature of counterfactuals. It would be surprising if such a relation between counterfactual quantities and the consequences of making choices were true unless counterfactuals were defined in such a manner as to make it true.

Proposition 4 together with 2 is suggestive of a two-step process of assessing symmetry in decision models: first, we identify some apparent symmetry of the decision procedure in the real world, and second we consider this symmetry to imply exchangeability only of those variables associated with features of the world that do not change in response to our choices.

Example 5.2.5. A sequential experiment is modeled by a probability set $\mathbb{P}_C$ with binary treatments $\mathsf{D} := (\mathsf{D}_i)_{i \in \mathbb{N}}$ and binary outcomes $\mathsf{Y} := (\mathsf{Y}_i)_{i \in \mathbb{N}}$. The set of choices C is the set of all probability distributions$\Delta(D^{\mathbb{N}})$.

The treatments are decided as follows: the decision maker 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 decision maker will assign patient i the treatment d_i. Otherwise, if μ_i is nondeterministic, the decision maker 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.

If the inputs were always an invertible function of the choice, we would have $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathrm{id}_C | \mathsf{D}$. We assume that the randomisation procedure does not change this fact.

How might the decision maker justify an assumption of IO contractibility of $\mathbb{P}_C^{\mathsf{Y}|\mathsf{HD}}$ in this context? A number of options are:

- They might assess that given any two choices α and α' that yield treatment distributions that are equivalent up to a permutation of one another, the corresponding outcomes will also be equivalent up to the same permutation, and conclude exchange commutativity from this (see Section 5.2.1)

- They might assess that the problem of predicting an output from its input and some infinitely supported input-output sequence is essentially the same as the problem of predicting any other output from its corresponding input and any other infinitely supported input-output sequence (see Section 4.4.2)

- They might speculate that each patient at each point in time contains, in some manner obscure to practically possible measurements, the outcome they would experience under any input they are given, something that is not influenced in any way by the decision maker's choice. They might further speculate that, owing to their ignorance about relevant differences between patients, their choices should be modeled as probabilistically independent of these "potential outcomes" *and* these potential outcomes should be modeled exchangeably (Section 5.2.1)

- They might assess that their choices are well-modeled by interventions in an uncertain unrolled causal Bayesian network (Definition 5.1.14, see also Pearl and Mackenzie (2018, Ch. 4))

We are also interested in the question of how the decision maker might conclude that the assumption of IO contractibility of $\mathbb{P}_C^{\mathsf{Y}|\mathsf{HD}}$ is false. To this end, we modify the example to consider an alternative experiment where IO contractibility probably should not be assumed. The construction is deliberately chosen to be somewhat awkward for the decision modelling approach – we suppose that α is "mostly chosen" by someone other than the decision maker. In particular, the decision maker is able to choose μ_n only for some $n \in \mathbb{N}$, and someone else subsequently determines the rest of the "choice". How might the justifications in the above example change?

Example 5.2.6 (I choose vs you choose)**.** Consider the previous experiment, except the sequence α of (possibly stochastic) treatment assignments is provided by the decision maker's assistant with the exception of μ_n. The decision maker considers it possible that the assistant has made this assignment with a view to promoting certain possible sequences of outputs Y over others.

Note that, for the deterministic subset C' of C, D is still an invertible function of id_C. Calling id_C a "choice" is now not so well justified.

- The decision maker considers μ_n to be importantly different from the other features of the choice, and so α and α' that differ by a swap involving μ_n to lead to outcomes that differ by more than the corresponding swap

- The decision maker consideres the index n to be importantly different from other indices, and so assesses that the problem of predicting Y_n given D_n and an arbitrary infinitely supported input-output sequence differs in important respects from the problem of predicting Y_m given D_m ($m \neq n$) given a similar infinitely supported sequence

- The decision maker speculates that their assistant might have some knowledge of the potential outcomes of the involved individuals, and so cannot conclude that the inputs are independent of these potential outcomes

- The decision maker assesses that their assistant might make the inputs depend on some unobserved confounder, and so any index m under the assistant's control is not modelled by an intervention on D_m

Perhaps it's too much to ask for any justification to be completely satisfactory. The two options that invoke symmetries feel like they want some unambiguously symmetric thing to justify their less transparent claims of symmetry. The invocation of potential outcomes "written within" each patient seems like it requires these potential outcomes to somehow be physically real, and it's plausible that they are not physically real. The approach based on interventions seems to beg the question of why the decision maker's actions are interventions but the assistant's are not. All four approaches seem to offer some insight into the question, and to us at least none are obvious implications of any of the other ones.

In this example, the choice id_C "depends on something unknown" to the decision maker. Back in Section 2.5.5 we opted not to consider models where the choice is a random variable because this introduced difficulties for modelling decision problems and didn't bring obvious benefits. However, this decision relied on the fact that id_C was in fact the result of a decision maker's deliberation involving the model $\mathbb{P}_C$. A second purpose of Example 5.2.6 is to make the point that calling some arbitrary variable a "choice" (or "id_C") and choosing not model its distribution does not make its distribution irrelevant.

Kasy (2016) has argued that "randomised controlled trials are not needed for causal identifiability, only controlled trials", and suggested 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, Chassang, Montero et al. (2020) suggested that deterministic rules might falter when an experimenter can't pick a function to balance covariates in a way that satisfies everyone in a panel of reviewers. As a final comment on these examples, we think the issue of randomised vs deterministic assignments in controlled experiments is more subtle than these authors claim. As these examples argue, there is a difference between a deterministic assignment selected by "me" according to my model and a deterministic assignment selected by "you" (or even one selected by "me" but with some opaque selection procedure). It may be possible to deal with this issue by specifying the model and assignment principle used to come up with a deterministic assignment, especially if it could be made clear that the procedure used cannot be easily modified post-hoc to alter conclusions.

5.3 Conclusion

We've shown how models constructed according to the causal Bayesian network and potential outcomes approaches can be translated to decision models. Both translations feature CIIR sequences, though in general these sequences involve unobserved inputs. In order to make this translation, we strengthened the assumptions underpinning causal Bayesian networks to rule out some possibilities we deemed undesirable that could arise when certain inputs have no support in the observational distribution. We also offered an interpretation for how the effects of choices could be represented in a potential outcomes model. We don't claim that either of these translations is the only possible way to translate models from either family to decision models, but we believe that the translations we offer are reasonable.

We only considered the simplest kind of interventions in casual Bayesian networks – perfect or "hard" interventions. Decision models may offer a more precise formalisation of interventions whose effects are in some manner unknown, owing to their explicit representation of uncertainty over they hypothesis H. Recalling the suggestion in Section 3.1 that counterfactuals may also be modeled by considering them to be the consequences of certain (physically implausible) actions, it may also be possible to extend the translation we have presented to encompass "counterfactual interventions".

We showed in Theorem 5.1.21 that the assumption of *precedent* can, under a side condition of *generic relationships* between certain conditionals, yield the conclusion that some sequence of inputs and corresponding outputs are related by conditionally independent and identical responses. The side condition we consider is similar to assumptions considered to be implied by causal direction, though the precise correspondence between notions of causal direction and the kind of generic relationship between conditionals we require is an open question.

Recalling the discussion in Section 4.4.3, the conclusion of CIIR sequences from Theorem 5.1.21 is not enough *on its own* to be useful to a decision maker. The decision maker must *also* bring some prior knowledge such as how their options affect the inputs in the sequence in question. As we point out Section 4.4.3, this kind of prior knowledge itself can suggest that some "input" variables are susceptible to influence by individuals who are ignorant about the conditional distribution of "outputs" given these inputs. We raise as a further open question when, if ever, reasoning of this type might support an assumption of the right kind of "generic relationship" between conditionals required by Theorem 5.1.21.

Chapter 6

Conclusion

In this thesis, we considered how idealised decision problems could be used as a basis for causal modelling. Decision making models, as we defined them, are similar to classical statistical models with sample spaces, random variables and probability measures defined on the sample space, but they add to this an option set, with each option associated with a probability measure on the sample space. The point of this is that, in general, a decision maker knows something a "classical statistician" does not – she knows what her set of options is. Decision making models also differ from the more common kinds of causal models – potential outcomes and structural interventional models – in that we suggest the former is interpreted alongside a *decision procedure* (Section 2.5.5), while the interpretation of potential outcomes and structural interventions is somewhat more free-floating.

Our reason for pursuing this approach is that often (though not always), causal modelling is done to assist a decision maker to make a good choice from among their options. In this situation, we suggest that a decision maker already has a decision procedure on hand and in that case our approach demands less additional theory than the potential outcomes or structural interventional approach. A key question is whether this "potential reduction in assumptions" actually enables the construction of practically useful causal models using fewer assumptions.

A simple assumption a decision maker with access to data could make when assessing their options is that they can identify corresponding sequences of inputs and outputs that are related by a fixed stochastic function. These sequences span the given data as well as observations that arise as a result of the choice made by the decision maker. They can therefore determine how they should expect their choice to influence the world by assessing the relationship between inputs and outputs they observed in the data. We proposed that this idea of a fixed but unknown functional relationship can be formalised using a sequential model that features conditionally independent and identical responses (CIIRs).

When a decision maker assumes that a model features CIIRs, the functional relationships themselves are unobserved. The decision maker doesn't necessarily have a clear means of interpreting arbitrary unobserved variables. We showed that the assumption of CIIRs is equivalent to a two symmetries of a decision model: firstly, it is equivalent to the *IO contractible* of this model over some auxiliary variable, and secondly (under some side conditions) it is equivalent to the interchangeability of infinite conditioning sequences (Theorems 4.3.23 and 4.3.26 respectively). We pointed out the second condition is often unreasonable.

The identified symmetries offer decision makers an alternative means of assessing unconfoundedness-like assumptions. In the potential outcomes framework, unconfoundedness is given as an assumption of independence between the potential outcomes and the corresponding inputs,

and in the structural graphical models framework it is given by the assumption that all back-door paths between the input and output variables are blocked by an observed variable. We showed that a decision maker might instead consider whether a set of problems are identical – for example, predicting the consequences of their actions given the data and predicting held out observations given the same data. A decision maker might form an opinion about these questions without appealing to theories of counterfactuals or structural interventions.

On the other hand, the fact that this assumption of interchangeable data sequences seems to be mostly unreasonable means that it doesn't quite deliver on the "useful" front. People genuinely do make the assumption of unconfoundedness, but perhaps this could be interpreted as an assumption that the condition of interchangeable data sequences holds (in some sense) approximately. We identified the assumption of *precedent* as a possible weakening of this condition. This assumption holds that, rather than *every* input-output pair in the sequence of observations obey the same relation as the input-output pairs produced as a result of the decision maker's actions, only some unknown (but non-negligible) fraction obeys the same relation. This assumption is motivated to some extent by the observation that this is precisely what is implied in the structural interventional setting by the assumption of *hidden confounders*.

The assumption of precedent has significant implications for inference when it is combined with an assumption of *generic relations between conditionals*. Theorem 5.1.21 showed that, given these assumptions, a conditional independence observed in the data implies an identical response function between inputs and outputs that arise as consequences of the decision maker's action. A weakness of this theorem is that the assumptions of precedent and generic relations between conditionals are themselves expressed in terms of an unobserved variable, and so we can't assume that a decision maker has any clear way of interpreting these assumptions. On top of this, the assumption of generic relations between conditionals is not particularly easy to understand by itself.

We offered the example of medical practitioners making prescriptions and observing patient outcomes to illustrate how this theorem works. In this case, one can informally understand the assumption of precedent as the assumption that, whatever the decision maker ends up doing, other medical practitioner have already sometimes acted in just the same way as the decision maker. There is an intuitive sense in which the choices made by practitioners are "causally prior" to the treatment and subsequent observation of patient outcomes (and also temporally prior).

Separately, a number of authors have suggested that generic relations between conditionals usually hold if the conditionals correspond to causal relations (Lemeire and Janzing, 2013; Meek, 1995). Our result suggests that such generic relations might, along with the assumption of precedent, be sufficient to support inference of the consequences of decisions from data. We have not established a perfect correspondence between the kinds of "generic relations" required by Theorem 5.1.21 and the kinds of generic relations investigated by researchers studying the principle of independent causes and mechanisms. However, if such a correspondence holds, then we would have an understanding of why independent causes and mechanisms are important to the practice of data-driven decision making that does not depend on a structural theory of causation. Of particular interest is the question of whether the various methods developed to assess causal direction based on the principle of independent causes and mechanisms provides any reason to believe the required type of relation for Theorem 5.1.21 holds – see (Mooij, J.M. et al., 2016) for an overview of some of these methods.

Another extension of this line of work is to consider finite sample performance of inference based on the assumption of precedent. Intuitively, one might expect that inference based on

the assumption of precedent is hard when the actions of interest are taken infrequently. Furthermore, we speculate that inference is also hard when the relation between conditionals is almost non-generic, which might happen when the data is produced by individuals controlling inputs in order to keep outputs in a desired range.

To express this theory, we made use of a string diagram notation for writing out some proofs and as a visual aid to understanding different kinds of decision models. String diagrams are simply a notation for reasoning using probability theory, and as such are a convenience, not a critical piece of the 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 without ambiguity. DAGs have an advantage over string diagrams in that correspondences between many structural properties of the diagram and properties of the model have been worked out – for example, the correspondence between d-separation and conditional independence, as well as more sophisticated properties necessary and sufficient for identifiability (Shpitser and Pearl, 2008; Tian and Pearl, 2002). A string diagram analogue of d-separation and its relation to different notions of conditional independence postulated by Fritz (2020) would further facilitate the use of string diagrams in causal reasoning.

The standard approaches to causal inference depend on a theory of causation – this may be a theory of structural interventions, a vague notion of counterfactuals or something else. Such theories can be helpful to the extent that they make contact with intuitive ideas we have about causation and counterfactuals, but such intuitions are only really relevant to a small class of data-driven decision making problems. As we have shown, theories of causation are not needed to formally represent data-driven decision problems, nor are they needed to formulate substantive assumptions that license a decision maker to draw conclusions about the consequences of their choices from the data they have available. Theories of causation are troublesome to researchers who want to study problems of data-driven decision making in diverse contexts; it simply isn't clear how to generalise the relevant causal intuitions beyond problems in which they are apparently reliable. Our work calls into question whether relying on theories of causation is really necessary. We have shown that causal inference problems can be formalised, and in some cases solved, without them, which allows one thus to sidestep the vexed question of exactly what they mean.

Bibliography

Aldous, D. J. (1981). 'Representations for partially exchangeable arrays of random variables'. *Journal of Multivariate Analysis*, 11(4), pp. 581–598. DOI: 10.1016/0047-259X(81)90099-3 (cited on p. 64).

Banerjee, A. V., Chassang, S. and Snowberg, E. (2017). 'Chapter 4 - Decision Theoretic Approaches to Experiment Design and External Validity'. In: *Handbook of Economic Field Experiments*. Ed. by A. V. Banerjee and E. Duflo. Vol. 1. Handbook of Field Experiments. North-Holland, pp. 141–174. DOI: 10.1016/bs.hefe.2016.08.005 (cited on p. 65).

Banerjee, A. V., Chassang, S., Montero, S. and Snowberg, E. (2020). 'A Theory of Experimenters: Robustness, Randomization, and Balance'. *American Economic Review*, 110(4), pp. 1206–1230. DOI: 10.1257/aer.20171634 (cited on p. 107).

Bareinboim, E., Correa, J. D., Ibeling, D. and Icard, T. (2020). *On Pearl's Hierarchy and the Foundations of Causal Inference*. Tech. rep. ⬀ (visited on 20 April 2022) (cited on pp. 46, 89).

Barto, A. G. and Sutton, R. S. (1998). *Reinforcement Learning: An Introduction*. text. ⬀ (visited on 30 April 2018) (cited on p. 14).

Berk, R. (2010). 'What You Can and Can't Properly Do with Regression'. *Journal of Quantitative Criminology*, 26(4), pp. 481–487. DOI: 10.1007/s10940-010-9116-4 (cited on p. 62).

Blackwell, D. A. (1979). *Theory of Games and Statistical Decisions*. New York: Dover Publications. ISBN: 978-0-486-63831-7 (cited on p. 59).

Bogachev, V. and Malofeev, I. (2020). 'Kantorovich problems and conditional measures depending on a parameter'. *Journal of Mathematical Analysis and Applications*, 486, p. 123883. DOI: 10.1016/j.jmaa.2020.123883 (cited on pp. 28, 31).

Bolker, E. D. (1966). 'Functions Resembling Quotients of Measures'. *Transactions of the American Mathematical Society*, 124(2), pp. 292–312. DOI: 10.2307/1994401 (cited on p. 52).

Bolker, E. D. (1967). 'A Simultaneous Axiomatization of Utility and Subjective Probability'. *Philosophy of Science*, 34(4), pp. 333–340. ⬀ (visited on 6 April 2022) (cited on p. 52).

Bongers, S., Peters, J., Schölkopf, B. and Mooij, J. M. (2016). 'Theoretical Aspects of Cyclic Structural Causal Models'. *arXiv:1611.06221 [cs, stat]*. arXiv: 1611.06221. ⬀ (visited on 6 February 2019) (cited on pp. 3, 11, 35, 89).

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S. and Drouin, A. (2020). 'Differentiable Causal Discovery from Interventional Data'. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 21865–21877. ⬀ (visited on 9 September 2022) (cited on p. 12).

Cartwright, N. (1994). *No Causes in, No Causes out*. Oxford University Press (cited on p. 80).

Cartwright, N. (2001). 'Modularity: It Can - and Generally Does - Fail'. *Institute of Philosophy*. ⬀ (visited on 21 July 2020) (cited on p. 7).

Chickering, D. M. (2002). 'Learning Equivalence Classes of Bayesian-Network Structures'. *Journal of Machine Learning Research*, 2(Feb), pp. 445–498. ⬀ (visited on 14 October 2018) (cited on p. 8).

Chickering, D. M. (2003). 'Optimal Structure Identification with Greedy Search'. *J. Mach. Learn. Res.*, 3, pp. 507–554. DOI: `10.1162/153244303321897717` (cited on pp. 8, 12).

Chiribella, G., D'Ariano, G. and Perinotti, P. (2008). 'Quantum Circuit Architecture'. *Physical review letters*, 101, p. 060401. DOI: `10.1103/PhysRevLett.101.060401` (cited on p. 63).

Cho, K. and Jacobs, B. (2019). 'Disintegration and Bayesian inversion via string diagrams'. *Mathematical Structures in Computer Science*, 29(7), pp. 938–971. DOI: `10.1017/S0960129518000488` (cited on pp. 21, 22, 24, 32).

Çinlar, E. (2011). *Probability and Stochastics*. Springer (cited on pp. 18, 20, 31, 123).

Constantinou, P. and Dawid, A. P. (2017). 'Extended Conditional Independence and Applications in Causal Inference'. *The Annals of Statistics*, 45(6), pp. 2618–2653. ⬀ (visited on 28 September 2021) (cited on pp. 17, 31–33).

Correa, J. and Bareinboim, E. (2020). 'A Calculus for Stochastic Interventions:Causal Effect Identification and Surrogate Experiments'. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(06). Number: 06, pp. 10093–10100. DOI: `10.1609/aaai.v34i06.6567` (cited on p. 12).

Daniusis, P., Janzing, D., Mooij, J., Zscheischler, J., Steudel, B., Zhang, K. and Schoelkopf, B. (2012). *Inferring deterministic causal relations*. arXiv:1203.3475 [cs, stat]. DOI: `10.48550/arXiv.1203.3475` (cited on p. 8).

Dawid, A. P. (2000). 'Causal Inference without Counterfactuals'. *Journal of the American Statistical Association*, 95(450), pp. 407–424. DOI: `10.1080/01621459.2000.10474210` (cited on pp. 13, 70).

Dawid, A. P. (2002). 'Influence Diagrams for Causal Modelling and Inference'. *International Statistical Review*, 70(2), pp. 161–189. DOI: `10.1111/j.1751-5823.2002.tb00354.x` (cited on p. 13).

Dawid, A. P. (2020). 'Decision-theoretic foundations for statistical causality'. *arXiv:2004.12493 [math, stat]*. arXiv: 2004.12493. ⬀ (visited on 23 September 2020) (cited on pp. 13, 65).

Dawid, P. (2012). 'The Decision-Theoretic Approach to Causal Inference'. In: *Causality*. John Wiley & Sons, Ltd, pp. 25–42. ISBN: 978-1-119-94571-0. DOI: `10.1002/9781119945710.ch4` (cited on p. 13).

Diaconis, P. (1988). 'Recent progress on de Finetti's notions of exchangeability'. *Bayesian Statistics*, 3, pp. 111–125 (cited on p. 64).

Diaconis, P. and Freedman, D. (1980). 'Finite Exchangeable Sequences'. *The Annals of Probability*, 8(4), pp. 745–764. DOI: `10.1214/aop/1176994663` (cited on p. 64).

Eberhardt, F., Hoyer, P. and Scheines, R. (2010). 'Combining Experiments to Discover Linear Cyclic Models with Latent Variables'. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. ISSN: 1938-7228. JMLR Workshop and Conference Proceedings, pp. 185–192. ⬀ (visited on 8 September 2022) (cited on p. 11).

Eberhardt, F. and Scheines, R. (2007). 'Interventions and Causal Inference'. *Philos. Sci.*, 74. DOI: `10.1086/525638` (cited on p. 12).

Eckles, D. and Bakshy, E. (2021). 'Bias and High-Dimensional Adjustment in Observational Studies of Peer Effects'. *Journal of the American Statistical Association*, 116(534), pp. 507–517. DOI: 10.1080/01621459.2020.1796393 (cited on p. 80).

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press. ISBN: 978-1-4832-2123-6 (cited on p. 58).

Feynman, R. (1979). *The Feynman lectures on physics*. Le cours de physique de Feynman. INIS Reference Number: 13648743. France: Intereditions. ISBN: 978-2-7296-0030-3 (cited on p. 40).

Finetti, B. de ([1937] 1992). 'Foresight: Its Logical Laws, Its Subjective Sources'. In: *Breakthroughs in Statistics: Foundations and Basic Theory*. Ed. by S. Kotz and N. L. Johnson. Springer Series in Statistics. New York, NY: Springer, pp. 134–174. ISBN: 978-1-4612-0919-5. DOI: 10.1007/978-1-4612-0919-5_10 (cited on pp. 48, 64).

Fisher, R. A. (1958). 'Cancer and Smoking'. *Nature*, 182(4635), pp. 596–596. DOI: 10.1038/182596a0 (cited on p. 1).

Fong, B. (2013). 'Causal Theories: A Categorical Perspective on Bayesian Networks'. *arXiv:1301.6201 [math]*. arXiv: 1301.6201. (visited on 8 August 2018) (cited on pp. 15, 24, 26).

Forré, P. and Mooij, J. M. (2017). 'Markov Properties for Graphical Models with Cycles and Latent Variables'. *arXiv:1710.08775 [math, stat]*. arXiv: 1710.08775. (visited on 22 July 2020) (cited on p. 3).

Forré, P. and Mooij, J. M. (2018). 'Constraint-based Causal Discovery for Non-Linear Structural Causal Models with Cycles and Latent Confounders'. *arXiv:1807.03024 [cs, stat]*. arXiv: 1807.03024. (visited on 22 July 2020) (cited on p. 12).

Forré, P. and Mooij, J. M. (2020). 'Causal Calculus in the Presence of Cycles, Latent Confounders and Selection Bias'. In: *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*. ISSN: 2640-3498. PMLR, pp. 71–80. (visited on 8 September 2022) (cited on pp. 11, 35, 89).

Fritz, T. (2020). 'A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics'. *Advances in Mathematics*, 370, p. 107239. DOI: 10.1016/j.aim.2020.107239 (cited on pp. 22, 24, 111).

Ghassami, A. (2020). 'Causal discovery beyond Markov equivalence'. Thesis. University of Illinois. (visited on 9 September 2022) (cited on p. 11).

Glymour, M. M. and Spiegelman, D. (2017). 'Evaluating Public Health Interventions: 5. Causal Inference in Public Health Research—Do Sex, Race, and Biological Factors Cause Health Outcomes?' *American Journal of Public Health*, 107(1), pp. 81–85. DOI: 10.2105/AJPH.2016.303539 (cited on p. 12).

Gordon, B. R., Moakler, R. and Zettelmeyer, F. (2022). 'Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement'. *arXiv:2201.07055 [econ]*. arXiv: 2201.07055. (visited on 30 April 2022) (cited on p. 80).

Gordon, B. R., Zettelmeyer, F., Bhargava, N. and Chapsky, D. (2018). *A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook*. SSRN Scholarly Paper ID 3033144. Rochester, NY: Social Science Research Network. (visited on 7 February 2019) (cited on p. 80).

Greenland, S. and Robins, J. M. (1986). 'Identifiability, Exchangeability, and Epidemiological Confounding'. *International Journal of Epidemiology*, 15(3), pp. 413–419. DOI: 10.1093/ije/15.3.413 (cited on pp. 65, 104).

Haavelmo, T. (1943). 'The Statistical Implications of a System of Simultaneous Equations'. *Econometrica*, 11(1). Publisher: [Wiley, Econometric Society], pp. 1–12. DOI: `10.2307/1905714` (cited on p. 5).

Halpern, J. Y. (1999). 'A Counter Example to Theorems of Cox and Fine'. *Journal of Artificial Intelligence Research*, 10, pp. 67–85. DOI: `10.1613/jair.536` (cited on p. 48).

Hauser, A. and Bühlmann, P. (2012). 'Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs'. *Journal of Machine Learning Research*, 13(79), pp. 2409–2464. ⇗ (visited on 8 September 2022) (cited on p. 12).

Heckerman, D. and Shachter, R. (1995). 'Decision-Theoretic Foundations for Causal Reasoning'. *Journal of Artificial Intelligence Research*, 3, pp. 405–430. DOI: `10.1613/jair.202` (cited on p. 13).

Hernán, M. A. and Taubman, S. L. (2008). 'Does obesity shorten life? The importance of well-defined interventions to answer causal questions'. *International Journal of Obesity*, 32(S3), S8–S14. DOI: `10.1038/ijo.2008.82` (cited on p. 4).

Hernán, M. A. (2012). 'Beyond exchangeability: The other conditions for causal inference in medical research'. *Statistical Methods in Medical Research*, 21(1), pp. 3–5. DOI: `10.1177/0962280211398037` (cited on p. 65).

Hernán, M. A. and Robins, J. M. (2006). 'Estimating causal effects from epidemiological data'. *Journal of Epidemiology and Community Health*, 60(7), pp. 578–586. DOI: `10.1136/jech.2004.029496` (cited on pp. 62, 65).

Hernán, M. A. (2016). 'Does water kill? A call for less casual causal inferences'. *Annals of Epidemiology*, 26(10), pp. 674–680. DOI: `10.1016/j.annepidem.2016.08.016` (cited on p. 4).

Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If.* Chapman & Hall/CRC. ⇗ (visited on 22 April 2022) (cited on p. 46).

Horvitz, E., Heckerman, D. and Langlotz, C. (1986). 'A Framework for Comparing Alternative Formalisms for Plausible Reasoning'. In: ⇗ (visited on 19 April 2022) (cited on p. 48).

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J. and Schölkopf, B. (2009). 'Nonlinear causal discovery with additive noise models'. In: *Advances in Neural Information Processing Systems 21*. Ed. by D. Koller, D. Schuurmans, Y. Bengio and L. Bottou. Curran Associates, Inc., pp. 689–696. ⇗ (visited on 6 February 2019) (cited on p. 8).

Imbens, G. W. and Angrist, J. D. (1994). 'Identification and Estimation of Local Average Treatment Effects'. *Econometrica*, 62(2), pp. 467–475. DOI: `10.2307/2951620` (cited on p. 6).

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge: Cambridge University Press. ISBN: 978-0-521-88588-1. DOI: `10.1017/CBO9781139025751` (cited on pp. 5, 6, 46, 64).

Jacobs, B., Kissinger, A. and Zanasi, F. (2019). 'Causal Inference by String Diagram Surgery'. In: *Foundations of Software Science and Computation Structures*. Ed. by M. Bojańczyk and A. Simpson. Lecture Notes in Computer Science. Springer International Publishing, pp. 313–329. ISBN: 978-3-030-17127-8 (cited on pp. 64, 82).

Jeffrey, R. C. (1965). *The Logic of Decision.* University of Chicago Press. ISBN: 978-0-226-39582-1 (cited on p. 52).

Kallenberg, O. (2005a). *Probabilistic Symmetries and Invariance Principles*. Probability and Its Applications. New York: Springer-Verlag. ISBN: 978-0-387-25115-8. DOI: `10.1007/0-387-28861-9` (cited on p. 64).

Kallenberg, O. (2005b). 'The Basic Symmetries'. In: *Probabilistic Symmetries and Invariance Principles*. Probability and Its Applications. New York, NY: Springer, pp. 24–68. ISBN: 978-0-387-28861-1. DOI: `10.1007/0-387-28861-9_2` (cited on pp. 71, 128, 129, 138, 144).

Kasy, M. (2016). 'Why Experimenters Might Not Always Want to Randomize, and What They Could Do Instead'. *Political Analysis*, 24(3), pp. 324–338. DOI: `10.1093/pan/mpw012` (cited on p. 107).

Kerns, G. J. and Székely, G. J. (2006). 'Definetti's Theorem for Abstract Finite Exchangeable Sequences'. *Journal of Theoretical Probability*, 19(3), pp. 589–608. DOI: `10.1007/s10959-006-0028-z` (cited on p. 64).

Kohavi, R. and Thomke, S. (2017). 'The Surprising Power of Online Experiments'. *Harvard Business Review*. Section: Experimentation. ⬀ (visited on 27 June 2022) (cited on p. 2).

Lattimore, F. and Rohde, D. (2019a). 'Causal inference with Bayes rule'. *arXiv:1910.01510 [cs, stat]*. arXiv: 1910.01510. ⬀ (visited on 30 September 2021) (cited on pp. 93, 95).

Lattimore, F. and Rohde, D. (2019b). 'Replacing the do-calculus with Bayes rule'. *arXiv:1906.07125 [cs, stat]*. arXiv: 1906.07125. ⬀ (visited on 23 September 2020) (cited on p. 95).

Lattimore, F. R. (2017). 'Learning how to act: making good decisions with machine learning'. Accepted: 2018-06-27T06:17:13Z Last Modified: 2020-05-19 Publisher: The Australian National University. DOI: `10.25911/5d67b766194ec` (cited on p. 14).

Lemeire, J. and Janzing, D. (2013). 'Replacing Causal Faithfulness with Algorithmic Independence of Conditionals'. *Minds and Machines*, 23(2), pp. 227–249. DOI: `10.1007/s11023-012-9283-1` (cited on pp. 7, 14, 89, 101, 110).

Lewis, D. (1981). 'Causal decision theory'. *Australasian Journal of Philosophy*, 59(1), pp. 5–30. DOI: `10.1080/00048408112340011` (cited on p. 53).

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J. and Moher, D. (2009). '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. DOI: `10.1371/journal.pmed.1000100` (cited on p. 2).

Lindley, D. V. and Novick, M. R. (1981). 'The Role of Exchangeability in Inference'. *The Annals of Statistics*, 9(1), pp. 45–58. ⬀ (visited on 4 May 2022) (cited on p. 64).

Meek, C. (1995). 'Strong Completeness and Faithfulness in Bayesian Networks'. In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. UAI'95. event-place: Montréal, Qué, Canada. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 411–418. ISBN: 978-1-55860-385-1. ⬀ (visited on 19 March 2019) (cited on pp. 14, 89, 101, 110).

Menger, K. (2003). 'Random Variables from the Point of View of a General Theory of Variables'. In: *Selecta Mathematica: Volume 2*. Ed. by K. Menger, B. Schweizer, A. Sklar, K. Sigmund, P. Gruber, E. Hlawka, L. Reich and L. Schmetterer. Vienna: Springer, pp. 367–381. ISBN: 978-3-7091-6045-9. DOI: `10.1007/978-3-7091-6045-9_31` (cited on p. 41).

Mooij, J.M., Peters, J., Janzing, D., Zscheischler, J., Schölkopf, B. and Amsterdam Machine Learning lab (IVI, FNWI) (2016). 'Distinguishing Cause from Effect Using Observational

Data: Methods and Benchmarks'. *Journal of Machine Learning Research*, 17. ↗ (visited on 7 February 2019) (cited on pp. 8, 110).

Muller, S. M. (2015). 'Causal Interaction and External Validity: Obstacles to the Policy Relevance of Randomized Evaluations'. *The World Bank Economic Review*, 29(suppl_1), S217–S225. DOI: `10.1093/wber/lhv027` (cited on p. 62).

Ng, I., Zhu, S., Chen, Z. and Fang, Z. (2019). *A Graph Autoencoder Approach to Causal Structure Learning.* Number: arXiv:1911.07420 arXiv:1911.07420 [cs, stat]. DOI: `10.48550/arXiv.1911.07420` (cited on pp. 8, 12).

Nosek, B. A., Ebersole, C. R., DeHaven, A. C. and Mellor, D. T. (2018). 'The preregistration revolution'. *Proceedings of the National Academy of Sciences*, 115(11), pp. 2600–2606. DOI: `10.1073/pnas.1708274114` (cited on p. 2).

Ogata, K. (1995). *Discrete-Time Control Systems.* 2 edition. Englewood Cliffs, N.J: Pearson. ISBN: 978-0-13-034281-2 (cited on p. 14).

Okamoto, M. (1973). 'Distinctness of the Eigenvalues of a Quadratic form in a Multivariate Sample'. *The Annals of Statistics*, 1(4). Publisher: Institute of Mathematical Statistics, pp. 763–765. ↗ (visited on 5 July 2022) (cited on p. 150).

Open Science Collaboration (2015). 'Estimating the reproducibility of psychological science'. *Science*, 349(6251), aac4716. DOI: `10.1126/science.aac4716` (cited on p. 2).

Oreskes, N. and Conway, E. M. (2011). *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.* New York, NY: Bloomsbury Press. ISBN: 978-1-60819-394-3 (cited on p. 1).

Pearl, J. (2009). *Causality: Models, Reasoning and Inference.* 2nd ed. Cambridge University Press (cited on pp. 7–9, 12, 36, 40, 61, 62, 80, 90, 91).

Pearl, J. (2015). 'Causes of Effects and Effects of Causes'. *Sociological Methods & Research*, 44(1). Publisher: SAGE Publications Inc, pp. 149–164. DOI: `10.1177/0049124114562614` (cited on p. 7).

Pearl, J. (2018). 'Does Obesity Shorten Life? Or is it the Soda? On Non-manipulable Causes'. *Journal of Causal Inference*, 6(2). DOI: `10.1515/jci-2018-2001` (cited on pp. 4, 44).

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect.* 1 edition. New York: Basic Books. ISBN: 978-0-465-09760-9 (cited on pp. 7, 46, 106).

Peters, J. and Bühlmann, P. (2015). 'Structural Intervention Distance for Evaluating Causal Graphs'. *Neural Computation*, 27(3), pp. 771–799. DOI: `10.1162/NECO_a_00708` (cited on p. 12).

Peters, J., Bühlmann, P. and Meinshausen, N. (2016). 'Causal inference by using invariant prediction: identification and confidence intervals'. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5), pp. 947–1012. DOI: `10.1111/rssb.12167` (cited on p. 89).

Peters, J., Janzing, D. and Schölkopf, B. (2017). *Elements of Causal Inference.* MIT Press (cited on p. 8).

Proctor, R. N. (2012). 'The history of the discovery of the cigarette–lung cancer link: evidentiary traditions, corporate denial, global toll'. *Tobacco Control*, 21(2), pp. 87–91. DOI: `10.1136/tobaccocontrol-2011-050338` (cited on p. 1).

Richardson, T. S. and Robins, J. M. (2013). '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), p. 2013 (cited on p. 36).

Richardson, T. S., Evans, R. J., Robins, J. M. and Shpitser, I. (2017). 'Nested Markov Properties for Acyclic Directed Mixed Graphs'. *arXiv:1701.06686 [stat].* arXiv: 1701.06686. 🔗 (visited on 8 April 2022) (cited on pp. 9, 84).

Rubin, D. B. (2005). 'Causal Inference Using Potential Outcomes'. *Journal of the American Statistical Association*, 100(469), pp. 322–331. DOI: `10.1198/016214504000001880` (cited on pp. 46, 62, 64, 65, 72, 101, 105).

Saarela, O., Stephens, D. A. and Moodie, E. E. M. (2020). 'The role of exchangeability in causal inference'. DOI: `10.48550/arXiv.2006.01799` (cited on pp. 62, 64, 65).

Savage, L. J. (1951). 'The theory of statistical decision'. *Journal of the American Statistical Association*, 46, pp. 55–67. DOI: `10.2307/2280094` (cited on p. 53).

Savage, L. J. (1954). *The Foundations of Statistics.* Wiley Publications in Statistics (cited on pp. 50, 51, 120).

Scherrer, N., Bilaniuk, O., Annadani, Y., Goyal, A., Schwab, P., Schölkopf, B., Mozer, M. C., Bengio, Y., Bauer, S. and Ke, N. R. (2022). *Learning Neural Causal Models with Active Interventions.* arXiv:2109.02429 [cs, stat]. DOI: `10.48550/arXiv.2109.02429` (cited on p. 12).

Selinger, P. (2011). 'A Survey of Graphical Languages for Monoidal Categories'. In: *New Structures for Physics.* Ed. by B. Coecke. Lecture Notes in Physics. Berlin, Heidelberg: Springer, pp. 289–355. ISBN: 978-3-642-12821-9. DOI: `10.1007/978-3-642-12821-9_4` (cited on pp. 22, 24).

Shahar, E. (2009). 'The association of body mass index with health outcomes: causal, inconsistent, or confounded?' *American Journal of Epidemiology*, 170(8), pp. 957–958. DOI: `10.1093/aje/kwp292` (cited on p. 38).

Shimizu, S., Hoyer, P. O., Hyv&#228, A., rinen and Kerminen, A. (2006). 'A Linear Non-Gaussian Acyclic Model for Causal Discovery'. *Journal of Machine Learning Research*, 7(72), pp. 2003–2030. 🔗 (visited on 12 September 2022) (cited on p. 8).

Shpitser, I. and Pearl, J. (2008). 'Complete Identification Methods for the Causal Hierarchy'. *Journal of Machine Learning Research*, 9(Sep), pp. 1941–1979. 🔗 (visited on 22 September 2020) (cited on pp. 8, 111).

Skyrms, B. (1982). 'Causal Decision Theory'. *The Journal of Philosophy*, 79(11), pp. 695–711. DOI: `10.2307/2026547` (cited on p. 53).

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G. and Richardson, T. (2000). *Causation, prediction, and search.* MIT press (cited on pp. 8, 12, 46).

Starr, W. (2021). 'Counterfactuals'. In: *The Stanford Encyclopedia of Philosophy.* Ed. by E. N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University. 🔗 (visited on 2 September 2022) (cited on p. 5).

Statista (2020). *Cigarettes - worldwide | Statista Market Forecast.* 🔗 (visited on 21 September 2020) (cited on p. 1).

Steele, K. and Stefánsson, H. O. (2020). 'Decision Theory'. In: *The Stanford Encyclopedia of Philosophy.* Ed. by E. N. Zalta. Winter 2020. Metaphysics Research Lab, Stanford University. 🔗 (visited on 9 June 2021) (cited on p. 50).

Stroebe, W. (2019). 'What Can We Learn from Many Labs Replications?' *Basic and Applied Social Psychology*, 41(2), pp. 91–103. DOI: 10.1080/01973533.2019.1577736 (cited on p. 2).

Tian, J. and Pearl, J. (2002). 'A general identification condition for causal effects'. In: *Aaai/iaai*, pp. 567–573 (cited on p. 111).

Tinbergen, J. (1930). 'Determination and Interpretation of Supply Curves: An Example (1930)'. *The foundations of econometric analysis*, p. 233 (cited on p. 5).

Toth, C., Lorch, L., Knoll, C., Krause, A., Pernkopf, F., Peharz, R. and Kügelgen, J. von (2022). *Active Bayesian Causal Inference.* arXiv:2206.02063 [cs, stat]. DOI: 10.48550/arXiv.2206.02063 (cited on p. 12).

Uhler, C., Raskutti, G., Bühlmann, P. and Yu, B. (2013). 'Geometry of the faithfulness assumption in causal inference'. *The Annals of Statistics*, 41(2). arXiv: 1207.0547, pp. 436–463. DOI: 10.1214/12-AOS1080 (cited on p. 8).

Von Neumann, J. and Morgenstern, O. (1944). *Theory of games and economic behavior.* Theory of games and economic behavior. Princeton, NJ, US: Princeton University Press (cited on p. 50).

Vorob'ev, N. N. (1962). 'Consistent Families of Measures and Their Extensions'. *Theory of Probability & Its Applications*, 7(2). DOI: 10.1137/1107014 (cited on p. 35).

Wald, A. (1950). *Statistical decision functions.* Statistical decision functions. Oxford, England: Wiley (cited on pp. 14, 46, 49, 53).

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities.* Chapman & Hall (cited on pp. 48, 63).

Wiblin, R. (2016). *Why smoking in the developing world is an enormous problem and how you can help save lives.* ⬀ (visited on 21 September 2020) (cited on p. 1).

Woodward, J. (2016). 'Causation and Manipulability'. In: *The Stanford Encyclopedia of Philosophy.* Ed. by E. N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University. ⬀ (visited on 20 July 2020) (cited on p. 7).

World Health Organisation (2018). *Tobacco Fact sheet no 339.* ⬀ (visited on 21 September 2020) (cited on p. 1).

Yang, K., Katoff, A. and Uhler, C. (2018). 'Characterizing and Learning Equivalence Classes of Causal DAGs under Interventions'. In: *International Conference on Machine Learning*, pp. 5537–5546. ⬀ (visited on 19 July 2018) (cited on pp. 12, 89).

Zhang, J. and Spirtes, P. (2003). 'Strong Faithfulness and Uniform Consistency in Causal Inference'. In: *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence.* UAI'03. event-place: Acapulco, Mexico. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 632–639. ISBN: 978-0-12-705664-7. ⬀ (visited on 23 May 2019) (cited on p. 101).

Zhang, L. (2014). 'The Abdul Latif Jameel poverty action lab: bringing evidence-based policy into international development'. *Harvard International Review*, 35(4), pp. 4–6. ⬀ (visited on 27 June 2022) (cited on p. 2).

Zheng, X., Aragam, B., Ravikumar, P. K. and Xing, E. P. (2018). 'DAGs with NO TEARS: Continuous Optimization for Structure Learning'. In: *Advances in Neural Information Processing Systems.* Vol. 31. Curran Associates, Inc. ⬀ (visited on 15 June 2022) (cited on pp. 8, 12).

Appendix A

Axiomatisation of decision theories

As a supplement to Chapter 3, we reproduce the axioms of Savage and Jeffrey-Bolker decision theories. We offer these for the reader's convenience with minimal commentary.

A.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 $T : S \times C \to F$, 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 $A, B \in \mathcal{S}$, $A \leqslant B$ 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 $\leqslant$ as "is less probable than"

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

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" $\leqslant$ 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.

A.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.

Appendix B

Proofs of key results in Chapter 4

B.1 IO Contractibility

B.1.1 Equality of equally sized contractions

This is the proof of Theorem 4.3.7.

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 B.1.1 (Finite transposition). Given two equally sized sequences $A, B \in \mathbb{N}^n$ with $A = (a_i)_{i \in [n]}$, $B = (b_i)_{i \in [n]}$, $A \to B : \mathbb{N} \to \mathbb{N}$ is the permutation such that

$$[A \to B](a_i) = b_i$$

that sends the ith element of A to the ith element of B and vise versa. Note that $B \to A$ is the inverse of $A \to B$.

Lemma B.1.2 is used to extend conditional probabilities of finite sequences to infinite ones.

Lemma B.1.2 (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* $j > i$

$$\mathbb{K}_j(id_{Y^i} \otimes Del_{Y^{j-i}}) = \mathbb{K}_i \otimes Del_{X^{j-i}} \tag{B.1}$$

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 $x_{|m} \in X^n$ be the first n elements of x. By Equation (B.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 $\rho : [n] \to [n]$

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

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

$$\mathbb{Q}_x(\underset{i \in [n]}{\times} A_i \times Y^{\mathbb{N}}) = \mathbb{K}_n(\underset{i \in [n]}{\times} A_{\rho(i)} | x_{[n]}) \tag{B.2}$$

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

$$\mathbb{Q}_x(\underset{i \in [n]}{\times} A_i \times Y^{\mathbb{N}}) \geq \mathbb{Q}_x(\underset{i \in [p]}{\times} A_i \times Y^{\mathbb{N}})$$

$$\geq \mathbb{Q}_x(\underset{i \in \mathbb{N}}{\times} A_i)$$

so by the Monotone convergence theorem, the sequence $\mathbb{Q}_x(\times_{i \in [n]} A_i)$ converges as $n \to \infty$ to $\mathbb{Q}_x(\times_{i \in \mathbb{N}} A_i)$. $x \mapsto \mathbb{Q}_x^{Z_n}(\times_{i \in [n]} A_i)$ is measurable for all n, $\{A_i \in \mathcal{Y} | i \in \mathbb{N}\}$ by Equation (B.2), and so $x \mapsto \mathbb{Q}_x$ is also measurable.

Thus $x \mapsto \mathbb{Q}_x$ is the desired Markov kernel $\mathbb{K}$. $\qquad\square$

Corollary B.1.3. *Given* $(\mathbb{P}_C, \Omega, \mathcal{F})$, $\mathsf{W} : \Omega \to V$ *and two pairs of sequences* $(\mathsf{V}, \mathsf{X}) := (\mathsf{V}_i, \mathsf{X}_i)_{i \in \mathbb{N}}$ *and* $(\mathsf{Y}, \mathsf{Z}) := (\mathsf{Y}_i, \mathsf{Z}_i)_{i \in \mathbb{N}}$ *with corresponding variables taking values in the same sets* $V = Y$ *and* $X = Z$, *if* $(\mathbb{P}_C, \mathsf{V}, \mathsf{X})$ *and* $(\mathbb{P}_C, \mathsf{Y}, \mathsf{Z})$ *are both local over* W *and*

$$\mathbb{P}^{\mathsf{X}_{[n]} | \mathsf{W}\mathsf{V}_{[n]}} = \mathbb{P}^{\mathsf{Z}_{[n]} | \mathsf{W}\mathsf{Y}_{[n]}}$$

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

$$\mathbb{P}^{\mathsf{X} | \mathsf{W}\mathsf{V}} = \mathbb{P}^{\mathsf{Z} | \mathsf{W}\mathsf{Y}}$$

Proof. By assumption of locality

$$\mathbb{P}^{\mathsf{X}_{[n]} | \mathsf{W}\mathsf{V}_{[n]}} \otimes \mathrm{Del}_{W^{\mathbb{N}}} = \mathbb{P}^{\mathsf{X} | \mathsf{W}\mathsf{V}}(\mathrm{id}_{X^n} \otimes \mathrm{Del}_{X^{\mathbb{N}}})$$

$$\mathbb{P}^{\mathsf{Z}_{[n]} | \mathsf{W}\mathsf{Y}_{[n]}} \otimes \mathrm{Del}_{W^{\mathbb{N}}} = \mathbb{P}^{\mathsf{Z} | \mathsf{W}\mathsf{Y}}(\mathrm{id}_{X^n} \otimes \mathrm{Del}_{X^{\mathbb{N}}})$$

hence for all $n, m > n$

$$\mathbb{P}^{\mathsf{X}_{[m]} | \mathsf{W}\mathsf{V}_{[m]}}(\mathrm{id}_{X^n} \otimes \mathrm{Del}_{X^{m-n}}) = \mathbb{P}^{\mathsf{Z}_{[m]} | \mathsf{V}\mathsf{Y}_{[m]}}(\mathrm{id}_{X^n} \otimes \mathrm{Del}_{X^{m-n}})$$

$$= \mathbb{P}^{\mathsf{X}_{[n]} | \mathsf{W}\mathsf{V}_{[n]}} \otimes \mathrm{Del}_{W^{m-n}}$$

and, in particular, by lemma B.1.2, $\mathbb{P}^{\mathsf{X} | \mathsf{W}\mathsf{V}}$ and $\mathbb{P}^{\mathsf{Z} | \mathsf{W}\mathsf{Y}}$ are the limits of the same sequence. $\quad\square$

Theorem B.1.4 (Graphical representation of exchange commutativity)**.** *A sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *along with some* $\mathsf{W} : \Omega \to W$ *commutes with exchange over* W *if and only if for every* α, *every finite permutation* $\rho : \mathbb{N} \to \mathbb{N}$ *and corresponding swap map* $\mathrm{Swap}_\rho : X^{\mathbb{N}} \rightarrow X^{\mathbb{N}}$

Proof. This follows from the fact that

$$\mathbb{P}_\alpha^{\mathsf{Y}_\rho | \mathsf{WD}_\rho} = \begin{array}{c} \mathsf{W} \\ \mathsf{D} \end{array} \boxed{\mathsf{Swap}_{\rho^{-1}}} \boxed{\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}} \boxed{\mathsf{Swap}_\rho} \vdash \mathsf{Y}$$

To see this, note that

$$\begin{array}{c} \mathsf{W} \\ \mathsf{D} \end{array} \boxed{\mathsf{Swap}_{\rho^{-1}}} \boxed{\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}} \boxed{\mathsf{Swap}_\rho} \vdash \mathsf{Y} \; (\underset{i\in\mathbb{N}}{\bigtimes} A_i | w, (d_i)_\mathbb{N})$$

$$= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}(\underset{i\in\mathbb{N}}{\bigtimes} A_{\rho^{-1}(i)} | w, (d_{\rho^{-1}(i)})_\mathbb{N})$$

$$= \mathbb{P}_\alpha^{\mathsf{Y}_\rho|\mathsf{WD}_\rho}(\underset{i\in\mathbb{N}}{\bigtimes} A_i | w, (d_i)_\mathbb{N})$$

$\square$

The main proof follows.

Theorem 4.3.7. *Given a sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *and some* W, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ *is IO contractible over* W *if and only if for all subsequences* $A, B \subset \mathbb{N}^{|A|}$ *and for every* α

$$\mathbb{P}_\alpha^{\mathsf{Y}_A|\mathsf{WD}_{A,\mathbb{N}\backslash A}} = \mathbb{P}_\alpha^{\mathsf{Y}_B|\mathsf{WD}_{B,\mathbb{N}\backslash B}}$$

$$= \mathbb{P}_\alpha^{\mathsf{Y}_A|\mathsf{WD}_A} \otimes del_{\mathsf{D}|\mathbb{N}\backslash A|}$$

Proof. Only if: For $Z \in \mathbb{N}^{|A|}$, let $del_{Z^\complement}$ be the Markov kernel associated with the map that sends Y to $\mathsf{Y}_Z := (\mathsf{Y}_i)_{i \in Z}$.

If A is finite, then let $n := |A|$ and by exchange commutativity

$$\mathbb{P}_\alpha^{\mathsf{Y}_A|\mathsf{WD}_{A,\mathbb{N}\backslash A}} = \mathbb{P}_\alpha^{\mathsf{Y}_A|\mathsf{WD}_{A\to[n]}}$$

$$= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}_{A\to[n]}} del_{A^\complement}$$

$$= \mathbb{P}_\alpha^{\mathsf{Y}_{[n]\to A}|\mathsf{WD}} del_{A^\complement}$$

Use the fact that $[n] \to A \circ B \to [n] = B \to A$ and apply exchange commutativity to get

$$\mathbb{P}_\alpha^{\mathsf{Y}_{[n]\to A}|\mathsf{WD}} \mathbb{F}_{\Pi_A} = \mathbb{P}_\alpha^{\mathsf{Y}_{B\to A}|\mathsf{WD}_{B\to[n]}} del_{A^\complement}$$

$$= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}_{B\to[n]}} del_{B^\complement}$$

$$= \mathbb{P}_\alpha^{\mathsf{Y}_B|\mathsf{WD}_{B,\mathbb{N}\backslash B}}$$

if A is infinite, then we can take finite subsequences A_m that are the first m elements of A and similarly for B_m. Then by previous reasoning

$$\mathbb{P}_\alpha^{\mathsf{Y}_{A_m}|\mathsf{WD}_{A_m\to[m]}} = \mathbb{P}_\alpha^{\mathsf{Y}_{[m]}|\mathsf{WD}}$$

$$= \mathbb{P}_\alpha^{\mathsf{Y}_{B_m}|\mathsf{WD}_{B_m\to[m]}}$$

then by Corollary B.1.3

$$\mathbb{P}_\alpha^{\mathsf{Y}_A | \mathsf{WD}_{A \to [n]}} = \mathbb{P}_\alpha^{\mathsf{Y}_{Bm} | \mathsf{WD}_{Bm \to [m]}}$$

Finally, by locality

$$\mathbb{P}_\alpha^{\mathsf{Y}_A | \mathsf{WD}_{A \to [n]}} = \mathbb{P}_\alpha^{\mathsf{Y}_A | \mathsf{WD}_A} \otimes \mathrm{Del}_{D^{|\mathbb{N} \setminus A}}$$

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. $\qquad \square$

B.2 Tabulated conditional distributions

This is the proof of Lemmas 4.3.17 and 4.3.18 and Theorem 4.3.19. The following definitions are reproduced for convenience.

Definition 4.3.9. *Given a sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *on* $(\Omega, \mathcal{F})$ *with countable* D, $\#_j^k$ *is the variable*

$$\#_j^k := \sum_{i=1}^{k-1} \llbracket \mathsf{D}_i = j \rrbracket$$

In particular, $\#_j^k$ *is equal to the number of times* $\mathsf{D}_i = j$ *over all* $i < k$.

Definition 4.3.10. *Given a sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *on* $(\Omega, \mathcal{F})$, *define the tabulated conditional distribution* $\mathsf{Y}^D : \Omega \to Y^{\mathbb{N} \times D}$ *by*

$$\mathsf{Y}_{ij}^D = \sum_{k=1}^\infty \llbracket \#_j^k = i - 1 \rrbracket \llbracket \mathsf{D}_k = j \rrbracket \mathsf{Y}_k$$

That is, the (i, j)-th *coordinate of* $\mathsf{Y}^D(\omega)$ *is equal to the coordinate* $\mathsf{Y}_k(\omega)$ *for which the corresponding* $\mathsf{D}_k(\omega)$ *is the ith instance of the value* j *in the sequence* $(\mathsf{D}_1(\omega), \mathsf{D}_2(\omega), ...)$, *or 0 if there are fewer than* i *instances of* j *in this sequence.*

The proof of the theorem follows.

Lemma 4.3.17. *Suppose a sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *is given with* D *countable and* D *infinitely supported. Then for some* W, α, $\mathbb{P}_\alpha^{\mathsf{Y} | \mathsf{WD}}$ *is IO contractible if and only if*

$$\mathbb{P}_\alpha^{\mathsf{Y} | \mathsf{WD}} \Big(\underset{i \in \mathbb{N}}{\times} A_i \big| w, (d_i)_{i \in \mathbb{N}} \Big) = \mathbb{P}_\alpha^{(\mathsf{Y}_{id_i}^D)_{i \in \mathbb{N}} | \mathsf{W}} \Big(\underset{i \in \mathbb{N}}{\times} A_i \big| w \Big) \qquad \forall A_i \in \mathcal{Y}^D, w \in W, d_i \in D$$

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

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

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

$$\mathbb{P}_\alpha^{(\mathsf{Y}_{ij}^D)_{\mathbb{N}\times D}|\mathsf{W}} = \mathbb{P}_\alpha^{(\mathsf{Y}_{\eta(i,j)}^D)_{\mathbb{N}\times D}|\mathsf{W}} \tag{B.3}$$

Proof. Only if: We define a random invertible function $\mathsf{R} : \Omega \times \mathbb{N} \to \mathbb{N} \times D$ that reorders the indicies so that, for $i \in \mathbb{N}, j \in D$, $\mathsf{D}_{\mathsf{R}^{-1}(i,j)} = j$ almost surely. We then use IO contractibility to show that $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}}(\cdot|d)$ is equal to the distribution of the elements of Y^D selected according to $d \in D^{\mathbb{N}}$.

Note that at most one of $[\![\#_j^k = i - 1]\!][\![\mathsf{D}_k = j]\!]$ and $[\![\#_j^l = i - 1]\!][\![\mathsf{D}_l = j]\!]$ can be greater than 0 for $k \neq l$ and, by assumption, $\sum_{j \in D} \sum_{k \in \mathbb{N}} [\![\#_j^k = i - 1]\!][\![\mathsf{D}_k = j]\!] = 1$ almost surely (that is, for any i, j there is some k such that D_k is the ith occurrence of j). Define $\mathsf{R}_k : \Omega \to \mathbb{N} \times D$ by $\omega \mapsto \arg\max_{i \in \mathbb{N}, j \in D} [\![\#_j^k = i - 1]\!][\![\mathsf{D}_k = j]\!](\omega)$ (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 $\mathsf{R} : \mathbb{N} \to \mathbb{N} \times D$ by $k \mapsto \mathsf{R}_k$. R is almost surely bijective and

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

By construction, $\mathsf{D}_{\mathsf{R}^{-1}(i,j)} = j$ almost surely; that is, $\mathsf{D}_{\mathsf{R}^{-1}}$ 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

$$\begin{aligned}
\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{WD}_{\mathsf{R}^{-1}}}(A|w,d) &= \mathbb{P}_\alpha^{\mathsf{Y}_{\mathsf{R}^{-1}}|\mathsf{WD}_{\mathsf{R}^{-1}}}(A|w,d) \\
&\stackrel{\mathbb{P}_{\mathcal{C}}}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}_{\mathsf{R}^{-1}}|\mathsf{WD}_{\mathsf{R}^{-1}}}(A|w,e) \\
&= \mathbb{P}_\alpha^{\mathsf{Y}^D}(A|w) \tag{B.4}
\end{aligned}$$

for any $d \in D^{\mathbb{N}}$.

Now,

$$\mathbb{P}_\alpha^{\mathsf{Y}_{\mathsf{R}^{-1}}|\mathsf{WD}_{\mathsf{R}^{-1}}}(A|w,d) = \int_R \mathbb{P}_\alpha^{\mathsf{Y}_\rho|\mathsf{WD}_\rho}(A|d) \mathbb{P}_\alpha^{\mathsf{R}^{-1}|\mathsf{WD}_{\mathsf{R}^{-1}}}(d\rho|w,d) \tag{B.5}$$

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

$$\begin{aligned}
\mathbb{P}^{\mathsf{Y}_{\rho^n([n])}|\mathsf{WD}_{\rho^n([n])}} &= \mathbb{P}^{\mathsf{Y}_{\rho([n])}|\mathsf{WD}_{\rho([n])}} \\
&= \mathbb{P}^{\mathsf{Y}_{[n]}|\mathsf{WD}_{[n]}}
\end{aligned}$$

By Corollary B.1.3, it must therefore be the case that

$$\mathbb{P}^{\mathsf{Y}|\mathsf{WD}} = \mathbb{P}^{\mathsf{Y}_\rho|\mathsf{WD}_\rho}$$

Then from Equation (B.5)

$$
\mathbb{P}_\alpha^{\mathsf{Y}_{\mathsf{R}-1}|\mathsf{WD}_{\mathsf{R}-1}}(A|w,d) \stackrel{\mathbb{P}_C}{\cong} \int_R \mathbb{P}_\alpha^{\mathsf{Y}_\rho|\mathsf{WD}_\rho}(A|d)\mathbb{P}_\alpha^{\mathsf{R}-1|\mathsf{WD}_{\mathsf{R}-1}}(\mathrm{d}\rho|w,d)
$$

$$
\stackrel{\mathbb{P}_C}{\cong} \int_R \mathbb{P}_C^{\mathsf{Y}|\mathsf{WD}}(A|w,d)\mathbb{P}_\alpha^{\mathsf{R}-1|\mathsf{WD}_{\mathsf{R}-1}}(\mathrm{d}\rho|w,d)
$$

$$
\stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{\mathsf{Y}|\mathsf{WD}}(A|w,d) \tag{B.6}
$$

for all $i,j \in \mathbb{N}$. Then by Equation (B.4) and Equation (B.6)

$$
\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}(A|w) = \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}(A|w,e) \tag{B.7}
$$

Take some $d \in D^{\mathbb{N}}$. From Equation (B.7) and IO contractibility of $\mathbb{P}_C^{\mathsf{Y}|\mathsf{WD}}(A|e)$,

$$
(\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}} \otimes \mathrm{id}_D)\mathbb{F}_{lu}(A|w,d) = \mathbb{P}_\alpha^{(\mathsf{Y}^D_{id_i})_{i\in\mathbb{N}}|\mathsf{W}}(A|d)
$$

$$
= \mathbb{P}_\alpha^{(\mathsf{Y}_{id_i})_{i\in\mathbb{N}}|\mathsf{WD}}(A|w,e)
$$

$$
= \mathbb{P}_\alpha^{(\mathsf{Y}_{id_i})_{i\in\mathbb{N}}|\mathsf{W}(\mathsf{D}_{id_i})_{\mathbb{N}}}(A|w,(e_{id_i})_{i\in\mathbb{N}})
$$

$$
= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}(A|w,(e_{id_i})_{i\in\mathbb{N}})
$$

$$
= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}(A|w,(d_i)_{i\in\mathbb{N}})
$$

It remains to be shown that Y^D is invariant to finite permutations within rows. Consider some finite permutation within columns $\eta : \mathbb{N} \times D \to \mathbb{N} \times D$, note that $e_{\eta(i,j)} = j$ and hence $(e_{\eta(i,j)})_{i\in\mathbb{N},j\in D} = e$. Thus

$$
\mathbb{P}_\alpha^{(\mathsf{Y}^D_{\eta(i,j)})_{\mathbb{N}\times D}|\mathsf{W}}(A|w) = \mathbb{P}_\alpha^{(\mathsf{Y}^D)_{\mathbb{N}\times D}|\mathsf{W}}\mathrm{Swap}_\eta(A|w)
$$

$$
= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}\mathrm{Swap}_\eta(A|w,e) \qquad\qquad \text{from Eq. (B.7)}
$$

$$
= \mathbb{P}_\alpha^{\mathsf{Y}_\eta|\mathsf{WD}}(A|w,e)
$$

$$
= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}_{\eta^{-1}}}(A|w,e) \qquad\qquad \text{by exchange commutativity}
$$

$$
= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}(A|w,(e_{\eta^{-1}(i,j)})_{i\in\mathbb{N},j\in D})
$$

$$
= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}(A|w,e)
$$

$$
= \mathbb{P}_\alpha^{(\mathsf{Y}^D_{ij})_{\mathbb{N}\times D}|\mathsf{W}}(A|w) \qquad\qquad \text{from Eq. (B.7)}
$$

If: We construct a conditional probability according to Definition 4.3.10 and verify that it satisfies IO contractibility.

Suppose

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

where $\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}$ satisfies Equation (B.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 ith element of S with the ith element of T for all i.

$$\mathbb{P}_\alpha^{\mathsf{Y}_S|\mathsf{WD}}(\underset{i\in[n]}{\times} A_i | w, d) = \mathbb{P}_\alpha^{(\mathsf{Y}_{id_i}^D)_{i\in S}|\mathsf{W}}(\underset{i\in[n]}{\times} A_i | w)$$

$$= \mathbb{P}_\alpha^{(\mathsf{Y}_{S\leftrightarrow T(i)d_i}^D)_{i\in S}|\mathsf{W}}(\underset{i\in[n]}{\times} A_i | w)$$

$$= \mathbb{P}_\alpha^{(\mathsf{Y}_{id_{S\leftrightarrow T(i)}}^D)_{i\in T}|\mathsf{W}}(\underset{i\in[n]}{\times} A_i | w)$$

$$= \mathbb{P}_\alpha^{(\mathsf{Y}_{id'_i}^D)_{i\in T}|\mathsf{W}}(\underset{i\in[n]}{\times} A_i | w)$$

$$= \mathbb{P}_\alpha^{\mathsf{Y}_T|\mathsf{WD}}(\underset{i\in[n]}{\times} A_i | w, d')$$

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

$$= \mathbb{P}_\alpha^{\mathsf{Y}_{[n]}|\mathsf{WD}}(\underset{i\in[n]}{\times} A_i | w, d')$$

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

$$\mathbb{P}_\alpha^{\mathsf{Y}_S|\mathsf{WD}}(\underset{i\in[n]}{\times} A_i | w, d) = (\mathbb{P}_\alpha^{\mathsf{Y}_{[n]}|\mathsf{WD}_{[n]}} \otimes \mathrm{Del}_{D^{\mathbb{N}}})(\underset{i\in[n]}{\times} A_i | w, d')$$

so $\mathbb{K}$ is IO contractible by Theorem 4.3.7. $\qquad\square$

Lemma 4.3.18. *Suppose a sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *is given with* D *countable,* D *infinitely supported and for some* W, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ *is IO contractible for all* α. *Then, letting* H *be the directing random conditional of* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *(Definition 4.3.13) and* $\mathsf{Y}_{iD}^D := (\mathsf{Y}_{ij}^D)_{j\in D}$, *we have for all* $i \in \mathbb{N}$, $\mathsf{Y}_{iD}^D \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{\mathbb{N}\setminus\{i\}D}^D, \mathsf{W}, id_C)|\mathsf{H}$ *and*

$$\mathbb{P}_C^{\mathsf{Y}_{iD}^D|\mathsf{H}}(A|\nu) \overset{\mathbb{P}_\alpha}{\cong} \nu(A)$$

Proof. Fix $w \in W$ and consider $\mathbb{P}_{\alpha,w}^{\mathsf{Y}^D} := \mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}(\cdot|w)$. From Lemma 4.3.17, we have the exchangeability of the sequence $(\mathsf{Y}_{1D}^D, \mathsf{Y}_{2D}^D, ...)$ with respect to $(\mathbb{P}_{\alpha,w}, \Omega, \mathcal{F})$ as a special case of the invariance of $\mathbb{P}_\alpha^{(\mathsf{Y}_{ij}^D)_{\mathbb{N}\times D}|\mathsf{W}}$ to permutations of rows. By the column exchangeability of $\mathbb{P}_{\alpha,w}^{\mathsf{Y}^D}$, from Kallenberg (2005b, Prop. 1.4) (where H is precisely what Kallenberg calls the directing random measure)

Because the right hand side does not depend on w, we can say

$$\mathbb{P}_{\alpha}^{\mathsf{Y}^D|\mathsf{HW}} = \quad \mathsf{H} \longrightarrow \boxed{\begin{array}{c} \bullet\; \boxed{\mathbb{P}^{\mathsf{Y}_{iD}^D|\mathsf{H}}} - \mathsf{S}_i \\[4pt] i \in \mathbb{N} \end{array}}$$
$$\mathsf{W} \longrightarrow *$$

and because it also does not depend on α we have $\mathsf{Y}^D \perp\!\!\!\perp_{\mathbb{P}_C}^e (\mathsf{W}, \mathrm{id}_C)|\mathsf{H}$. Further application of Kallenberg (2005b, Prop. 1.4) yields $\mathsf{Y}_{iD}^D \perp\!\!\!\perp_{\mathbb{P}_C}^e (\mathsf{Y}_{\mathbb{N}\setminus\{i\}D}^D, \mathsf{W})|(\mathsf{H}, \mathrm{id}_C)$ and

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

Again, the right hand side does not depend on α, which yields $\mathsf{Y}_{iD}^D \perp\!\!\!\perp_{\mathbb{P}_C}^e (\mathsf{Y}_{\mathbb{N}\setminus\{i\}D}^D, \mathsf{W}, \mathrm{id}_C)|\mathsf{H}$.
$\square$

Theorem 4.3.19. *Suppose a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ is given with D countable, D infinitely supported and for some W, $\mathbb{P}_{\alpha}^{\mathsf{Y}|\mathsf{WD}}$ is IO contractible for all α. Consider an infinite set $A \subset \mathbb{N}$, and let $\mathsf{D}_A := (\mathsf{D}_i)_{i \in A}$ and $\mathsf{Y}_A := (\mathsf{Y}_i)_{i \in A}$. Then H_A, the directing random conditional of $(\mathbb{P}_C, \mathsf{D}_A, \mathsf{Y}_A)$ is almost surely equal to H, the directing random conditional of $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$.*

Proof. The strategy we will pursue is to show that an arbitrary subsequence of $(\mathsf{D}_i, \mathsf{Y}_i)$ 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 $\mathsf{Y}^{D,A}$ as the tabulated conditional of $(\mathsf{D}_A, \mathsf{Y}_A)$, i.e. let $\#_j^{A,k}$ be the count restricted to A:

$$\#_j^{A,k} := \sum_{i \in A}^{k-1} [\![\mathsf{D}_i = j]\!]$$

then

$$\mathsf{Y}_{ij}^{D,A} := \sum_{k \in A} [\![\#_j^{A,k} = i-1]\!][\![\mathsf{D}_k = j]\!]\mathsf{Y}_k$$
$$= \sum_{k \in A} [\![\#_j^{A,k} = i-1]\!][\![\mathsf{D}_k = j]\!]\mathsf{Y}_{\mathsf{R}_k j}^D$$

That is, defining $\mathsf{Q} : \mathbb{N} \to \mathbb{N}$ by $i \mapsto \sum_{k \in A}[\![\#_j^{A,k} = i-1]\!][\![\mathsf{D}_k = j]\!]\mathsf{R}_k$ then

$$\mathsf{Y}_{ij}^{D,A} = \mathsf{Y}_{\mathsf{Q}(i)j}^D \tag{B.8}$$

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

Equation (B.8) is what is meant by "the subsequence $(\mathsf{D}_A, \mathsf{Y}_A)$ induces a random contraction over the rows of Y^D". We will now show that $\mathsf{Y}^{D,A}$ has the same distribution as Y^D.

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

$$
\begin{aligned}
\mathbb{P}_\alpha^{Y^{D,A}|WQ}(B|w,q) &= \mathbb{P}_\alpha^{Y^D|W}\text{con}_q(B|w) \\
&= \mathbb{P}_\alpha^{Y|WD}\text{con}_q(B|w,e) && \text{by Eq.(B.7)} \\
&= \mathbb{P}_\alpha^{Y|WD}(B|w,e) && \text{by Theorem 4.3.7} \\
&= \mathbb{P}_\alpha^{Y^D|W}(B|w) && \text{by Eq.(B.7)} && \text{(B.9)}
\end{aligned}
$$

Finally, take H_A the directing random measure of $Y^{D,A}$. We conclude from the equality Eq. (B.9) and from the fact that there is a one-to-one map from directing random measures to exchangeable distributions that $H_A \overset{\mathbb{P}_\alpha}{\cong} H$. $\qquad\square$

B.3 Representation of IO contractible models

This is the proof of Lemma 4.3.20 and Theorem 4.3.21.

Lemma 4.3.20. *Suppose a sequential input-output model* $(\mathbb{P}_C, D, Y)$ *is given with* D *countable,* D *infinitely supported, for some* W, $\mathbb{P}_\alpha^{Y|WD}$ *is IO contractible for all α and for all α*

then $Y \perp\!\!\!\perp^e_{\mathbb{P}_C} (W, id_C)|(H, D)$, $H \perp\!\!\!\perp^e_{\mathbb{P}_C} D|(W, id_C)$ *and*

Proof. Y^D is a function of Y and D (see Definition 4.3.10) and H is a function of Y^D. Say $f : Y \times D \to H$ is such that $H = f(Y, D)$ (see Definition 4.3.12). Because $H = f(Y, D)$, we have $H \perp\!\!\!\perp^e_{\mathbb{P}_C} (W, id_C)|(Y, D)$. Thus

For a sequence $d \in D^{\mathbb{N}}$ where each $j \in D$ occurs infinitely often, take $[d = j]_i$ to be the ith 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

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

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

$$f_d \circ \mathsf{lu}(y^D, d) = \underset{j \in D}{\times} A_j \mapsto \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \prod_{j \in D} \mathbb{1}_{A_j}(y^D_{\#_{[d=j]_i}j})$$

Let $g := (y^D, d) \mapsto f_d \circ \mathsf{lu}(y^D, d)$ for some $d \in D^{\mathbb{N}}$ where each $j \in D$ occurs infinitely often.

We aim to show that $g(\mathsf{Y}^D, d) \overset{\mathbb{P}_\alpha}{\cong} g(\mathsf{Y}^D, d')$ for all $d, d' \in D^{\mathbb{N}}$ such that each $j \in D$ occurs infinitely often.

Consider, for arbitrary $A \in \mathcal{Y}^D$

$$\mathbb{P}_\alpha(g(\mathsf{Y}^D, d)(A) \bowtie g(\mathsf{Y}^D, d')(A)) = \int_H \mathbb{P}_\alpha^{\mathsf{Id}_\Omega | \mathsf{H}}(g(\mathsf{Y}^D, d)(A) \bowtie g(\mathsf{Y}^D, d')(A) | \nu) \mathbb{P}_\alpha^{\mathsf{H}}(d\nu)$$

Note that

$$\mathbb{P}_\alpha^{\mathsf{Id}_\Omega | \mathsf{H}}(g(\mathsf{Y}^D, d)(A) \bowtie \nu(A) | \nu) = \mathbb{P}_\alpha^{\mathsf{Y}^D | \mathsf{H}}(\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \prod_{j \in D} \mathbb{1}_{A_j}(y^D_{\#_{[d=j]_i}, j}) \bowtie \nu(A) | \nu) \mathbb{P}_\alpha^{\mathsf{H}}(d\nu)$$

by independent permutability of the rows of Y^D (Lemma 4.3.17), for each row we can send $\#_{[d=j]_i}$ to i and obtain

$$\mathbb{P}_\alpha^{\mathsf{Y}^D | \mathsf{H}}(\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \prod_{j \in D} \mathbb{1}_{A_j}(y^D_{\#_{[d=j]_i}, j}) \bowtie \nu(A) | \nu) \mathbb{P}_\alpha^{\mathsf{H}}(d\nu) = \mathbb{P}_\alpha^{\mathsf{Y}^D | \mathsf{H}}(\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \prod_{j \in D} \mathbb{1}_{A_j}(y^D_{i,j}) \bowtie \nu(A) | \nu)$$

$$= \mathbb{P}_\alpha^{\mathsf{Y}^D_{iD} | \mathsf{H}}(\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{A}(y^D_{i,D}) \bowtie \nu(A) | \nu)$$

but by Lemma 4.3.18, the sequence $(\mathsf{Y}^D_{iD})_{i \in \mathbb{N}}$ are mutually independent conditional on H and for all α, $\mathbb{P}_\alpha^{\mathsf{Y}_{iD} | \mathsf{H}}(A | \nu) \overset{\mathbb{P}_C}{\cong} \nu(A)$. Thus, by the law of large numbers

$$\mathbb{P}_\alpha^{\mathsf{Y}^D | \mathsf{H}}(\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\prod_{j \in D} A_j}(y^D_{i,D}) \bowtie \nu(A) | \nu) = 1$$

which implies

$$\int_H \mathbb{P}_\alpha^{\mathsf{Id}_\Omega | \mathsf{H}}(g(\mathsf{Y}^D, d)(A) \bowtie g(\mathsf{Y}^D, d')(A) | \nu) \mathbb{P}_\alpha^{\mathsf{H}}(d\nu)$$
$$= \int_H \mathbb{P}_\alpha^{\mathsf{Id}_\Omega | \mathsf{H}}(g(\mathsf{Y}^D, d)(A) \bowtie \nu(A) \cap g(\mathsf{Y}^D, d')(A) \bowtie \nu(A) | \nu) \mathbb{P}_\alpha^{\mathsf{H}}(d\nu)$$
$$= 1$$

Because this holds for all A,

$$g(\mathsf{Y}^D, d) \overset{\mathbb{P}_\alpha}{\cong} g(\mathsf{Y}^D, d') \qquad \text{as this holds for all } A$$

And, as a consequence, defining

$$i : (y^d, d, d') \mapsto (\mathsf{lu}(\mathsf{Y}^D, d), g(\mathsf{Y}^D, d'))$$

we have

$$i(y^d, d, d) \stackrel{\mathbb{P}_\alpha}{\cong} i(y^d, d, d')$$

which in turn implies the almost sure equality of the associated Markov kernels:
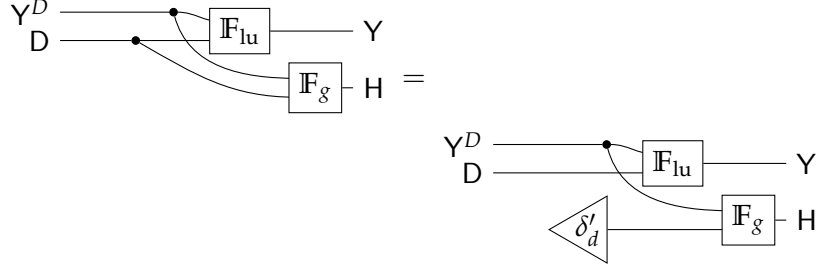
but we also have, by the definitions of f and g,
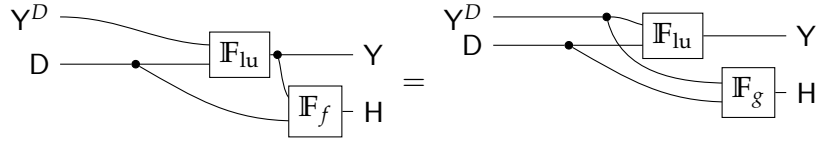
finally
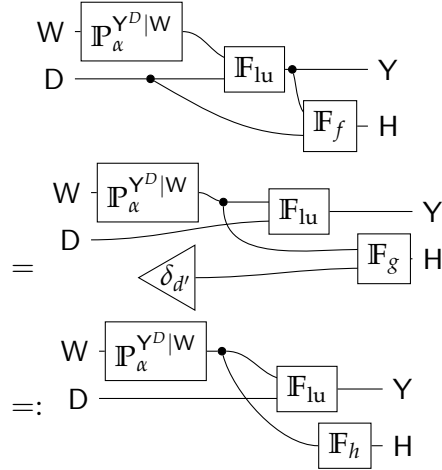
Noting that $\mathbb{F}_h \otimes \mathrm{Del}_W = \mathbb{P}_\alpha^{\mathsf{H}|\mathsf{Y}^D\mathsf{W}}$
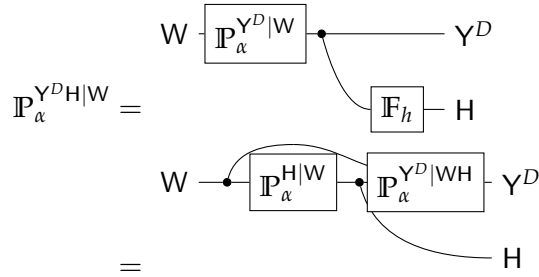
and so

$$\mathbb{P}_\alpha^{\mathsf{YH}|\mathsf{WD}} \;=\;$$

From Lemma 4.3.17 we also have $\mathsf{Y}^D \perp\!\!\!\perp_{\mathbb{P}_C}^e (\mathsf{W}, \mathrm{id}_C)|\mathsf{H}$, so

$$\mathbb{P}_\alpha^{\mathsf{YH}|\mathsf{WD}} \;=\;$$

and so by higher order conditionals $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathsf{W}|(\mathsf{H},\mathsf{D},\mathrm{id}_C)$ and

$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{HD}} \;=\;$$

Because the right hand side does not depend on α, we finally have $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_C}^e (\mathsf{W}, \mathrm{id}_C)|(\mathsf{H}, \mathsf{D})$ and the result

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{HD}} \;=\;$$

Furthermore, by marginalising the right hand side of Equation B.3 we have

$$\mathbb{P}_\alpha^{\mathsf{H}|\mathsf{WD}} \;=\;$$

Hence $\mathsf{H} \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathsf{D}|(\mathsf{W}, \mathrm{id}_C)$.   □

Theorem 4.3.21. *Suppose a sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *with sample space* $(\Omega, \mathcal{F})$ *is given with* D *countable and* D *infinitely supported. Then the following are equivalent:*

   1. There is some W *such that* $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ *is IO contractible for all* α

   2. For all i, $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (\mathsf{Y}_{\neq i}, \mathsf{D}_{\neq i}, id_C)|(\mathsf{H}, \mathsf{D}_i)$ *and for all* i, j

$$\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{HD}_i} = \mathbb{P}_C^{\mathsf{Y}_j|\mathsf{HD}_j}$$

   3. There is some $\mathbb{L} : H \times X \rightarrowtail Y$ *such that*

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{HD}} \;=\;$$

Proof. As a preliminary, we will show

$$
\mathbb{F}_{\mathrm{lu}} = \boxed{\begin{array}{c} Y^D \\ D \end{array} \quad \fbox{$\mathbb{F}_{\mathrm{lus}}$} - Y \\ i \in \mathbb{N}}
\tag{B.10}
$$

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

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

Recall that lu is the function

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

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

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

which is what we wanted to show.

$(1) \implies (3)$: From Lemma 4.3.17, we have some Y^D such that

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

and by Lemma 4.3.18

$$
\mathbb{P}_C^{\mathsf{Y}^D|\mathsf{H}} = \boxed{\begin{array}{c} \mathsf{H} - \bullet \qquad \fbox{$\mathbb{M}$} - \mathsf{Y}_i^D \\ i \in \mathbb{N} \end{array}}
\tag{B.11}
$$

By Lemma 4.3.17, for each $w \in W$

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

and so by Lemma 4.3.20

$$
\mathbb{P}_C^{\mathsf{Y}|\mathsf{HD}} = \begin{array}{c} \mathsf{H} - \fbox{$\mathbb{P}_C^{\mathsf{Y}^D|\mathsf{H}}$} \\ \mathsf{D} \end{array} \fbox{$\mathbb{F}_{\mathrm{lu}}$} - \mathsf{Y}
\tag{B.12}
$$

We can substitute Equations (B.11) and (B.10) into (B.12) for

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{HD}} =$$

$(3) \implies (2)$: If

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{HD}} =$$

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

$$\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{HD}_i\mathsf{Y}_{\neq i}\mathsf{D}_{\neq i}} \overset{\mathbb{P}_C}{\cong} \mathbb{L} \otimes \mathrm{Del}_{Y^{\mathbb{N}} \times X^{\mathbb{N}}}$$

hence $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (\mathsf{Y}_{\neq i}, \mathsf{D}_{\neq i}, \mathrm{id}_C)|(\mathsf{H}, \mathsf{D}_i)$

$(2) \implies (1)$: Take $\mathsf{W} := \mathsf{H}$. Because we assume $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (\mathsf{Y}_{[1,i)}, \mathsf{D}_{[1,i),\mathrm{id}_C})|(\mathsf{H}, \mathsf{D}_i)$ we can take $\mathbb{L} := \mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{HX}_i}$ for arbitrary i (by assumption, they are all the same) and

$$\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{HD}_i\mathsf{Y}_{[1,i)}\mathsf{D}_{[1,i)}} \overset{\mathbb{P}_C}{\cong} \mathbb{L} \otimes \mathrm{Del}_{Y^{i-1} \times X^{i-1}}$$

by taking the semidirect product of the conditionals

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{HD}} =$$

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

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

B.4 Proofs for symmetries of sequences with conditionally independent and identical responses

In this section Theorems 4.3.23, 4.3.24 and 4.3.26 are proved. Theorem 4.3.24 requires Lemma B.4.1, which is stated and proved in this section.

Theorem 4.3.23. *Suppose a sequential input-output model* $(\mathbb{P}_C, D, Y)$ *with sample space* $(\Omega, \mathcal{F})$ *is given with D countable and and D infinitely supported over $*$. Then the following are equivalent:*

1. $\mathbb{P}_\alpha^{Y|D}$ *is IO contractible for all α*

2. *For all i, $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{\neq i}, D_{\neq i}, id_C)|(H, D_i)$, for all i, j, α*

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

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

3. *There is some $\mathbb{L} : H \times X \rightharpoonup Y$ such that for all α,*
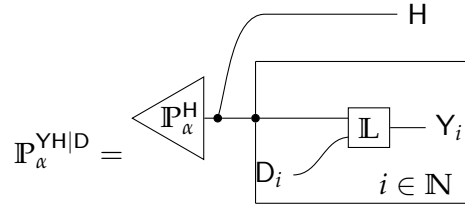
$$\mathbb{P}_\alpha^{YH|D} = \quad$$

Proof. $(1) \implies (3)$ From Lemmas 4.3.17 and 4.3.20, $(\mathbb{P}_C, D, Y)$ IO contractible over $*$ implies $H \perp\!\!\!\perp_{\mathbb{P}_C}^e D|(W, id_C)$.

From Theorem 4.3.21 we have $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{[1,i)}, D_{[1,i),id_C})|(H, D_i)$ and $\mathbb{P}_C^{Y_i|HD_i} = \mathbb{P}_C^{Y_j|HD_j}$ and

$$\mathbb{P}_C^{Y|HD} = \quad$$

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

$$\mathbb{P}_C^{YH|D} = \quad$$

$(3) \implies (2)$ From

$$\mathbb{P}_\alpha^{\mathsf{YH}|\mathsf{D}} = \quad (\text{B.13})$$

we have

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

and $\mathsf{H} \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathsf{D}|\mathrm{id}_C$, so we get all elements of (2) immediately except $\mathsf{D}_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (\mathsf{Y}_{[1,i)})|(\mathsf{D}_{[1,i)}, \mathrm{id}_C)$. But marginalising Equation (B.13) over H yields

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

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

$$\mathbb{P}_\alpha^{\mathsf{Y}_A|\mathsf{D}_{(A,\mathbb{N}\setminus A)}} = \mathbb{P}_\alpha^{\mathsf{Y}_A|\mathsf{D}_A} \otimes \mathrm{Del}_{\mathsf{D}_{|\mathbb{N}\setminus A|}}$$

hence, taking $A = \{i\}$, $\mathsf{D}_{\mathbb{N}\setminus\{i\}} \perp\!\!\!\perp_{\mathbb{P}_C}^e (\mathsf{Y}_i)|(\mathsf{D}_i, \mathrm{id}_C)$ which implies $\mathsf{D}_{(i,\infty]} \perp\!\!\!\perp_{\mathbb{P}_C}^e (\mathsf{Y}_i)|(\mathsf{D}_{[1,i]}, \mathrm{id}_C)$. $(2) \implies (1)$ From Theorem 4.3.21 and $\mathsf{H} \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathsf{D}|\mathrm{id}_C$,

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

and the argument from here is analogous to the section "$(2) \implies (1)$" from Theorem 4.3.21. In particular

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

hence $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ is exchange commutative. Furthermore, take $A \subset \mathbb{N}$. Then

$\square$
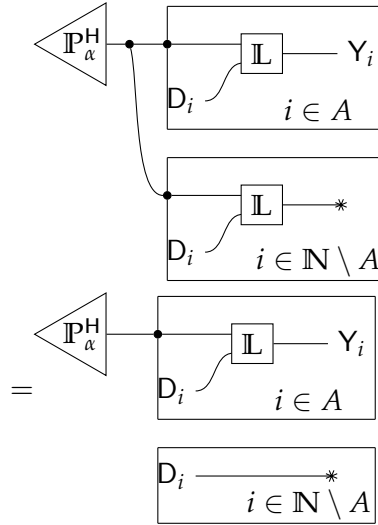
Lemma B.4.1 (Exchangeably dominated conditionals). *Given* $(\mathbb{P}_C, \Omega, \mathcal{F})$ *and variables* D, Y, *if for any* α *there is some* Q_α *such that* $\mathsf{Q}_\alpha^{\mathsf{DY}}$ *is exchangeable with directing random measure* G, D *is infinitely supported over* G *with respect to* Q_α *and for any* i, $\mathsf{Q}_\alpha^{\mathsf{Y}_i|\mathsf{DY}_{\{i\}^\complement}} \stackrel{\mathbb{P}}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{DY}_{\{i\}^\complement}}$ *then* $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{HD}}$ *is IO contractible (where* H *is the directing random conditional for* $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}}$*).*

Proof. By Kallenberg (2005b, Prop. 1.4), there is a G such that $(\mathsf{D}_i, \mathsf{Y}_i) \perp\!\!\!\perp_{\mathsf{Q}_C}^e (\mathsf{D}_{\{i\}^\complement} \mathsf{Y}_{\{i\}^\complement}) | (\mathsf{G}, \mathrm{id}_C)$ and for all i, j

$$\mathsf{Q}_\alpha^{\mathsf{Y}_i \mathsf{D}_i | \mathsf{G}} = \mathsf{Q}_\alpha^{\mathsf{Y}_j \mathsf{D}_j | \mathsf{G}} \tag{B.14}$$

There is some function $f : D^\mathbb{N} \times Y^\mathbb{N}$ such that $\mathsf{G} = f(\mathsf{D}, \mathsf{Y})$, i.e.

$$\stackrel{P}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}_i \mathsf{G} | \mathsf{DY}_{\{i\}^\complement}}$$

$$\implies \mathsf{Q}_\alpha^{\mathsf{Y}_i | \mathsf{GDY}_{\{i\}^\complement}} \stackrel{P}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}_i | \mathsf{GDY}_{\{i\}^\complement}} \tag{B.15}$$

It follows from weak union that

$$\mathsf{Y}_i \perp\!\!\!\perp_{\mathsf{Q}_C}^e (\mathsf{D}_{\{i\}^\complement} \mathsf{Y}_{\{i\}^\complement}) | (\mathsf{D}_i, \mathsf{G}, \mathrm{id}_C)$$

$$\iff \mathbb{P}_\alpha^{\mathsf{Y}_i | \mathsf{D}_i \mathsf{GY}_{\{i\}^\complement} \mathsf{D}_{\{i\}^\complement}} (A | d_i, g, d, y) \stackrel{P}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}_i | \mathsf{D}_i \mathsf{G}} (A | d_i, g) \qquad \forall A, d_i, g, d, y, \alpha \tag{B.16}$$

$$\implies \mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (\mathsf{D}_{\{i\}^\complement} \mathsf{Y}_{\{i\}^\complement}) | (\mathsf{D}_i, \mathsf{G}, \mathrm{id}_C)$$

where Eq. (B.16) follows from Eq. (B.15).

Finally, from Eq. (B.14) and Eq. (B.16)

$$\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{G}} \overset{\mathbb{P}}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}_j\mathsf{D}_j|\mathsf{G}}$$

Thus $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ features independent and identical responses conditioned on G, and by Lemma 4.3.22 it also has independent and identical responses conditioned on H. Finally, the infinite support of D over G with respect to Q_α implies D is also infinitely supported over G with respect to $\mathbb{P}_\alpha$, so by Theorem 4.3.21 $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{HD}}$ is IO contractible. $\square$

Theorem 4.3.24. *Given $(\mathbb{P}_C, \mathsf{Y}, \mathsf{D})$, if $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}}$ is exchange commutative for each α, and for each α $\mathbb{P}_\alpha^{\mathsf{D}}$ is absolutely continuous with respect to some exchangeable distribution $\mathsf{Q}_\alpha^{\mathsf{D}}$ in $\Delta(D^{\mathbb{N}})$ with directing random measure F and D infinitely supported over F with respect to Q_α, then $\mathbb{P}_C^{\mathsf{Y}|\mathsf{HD}}$ is IO contractible, where H is the directing random conditional of $(\mathbb{P}_C, \mathsf{Y}, \mathsf{D})$.*

Proof. For each α, extend $\mathsf{Q}_\alpha^{\mathsf{D}}$ to a distribution on (D, Y) by asserting that $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}} \overset{\mathsf{Q}_\alpha}{\cong} \mathsf{Q}_\alpha^{\mathsf{Y}|\mathsf{D}}$. Because $\mathsf{Q}_\alpha^{\mathsf{D}}$ dominates $\mathbb{P}_\alpha^{\mathsf{D}}$, we have in fact $\mathsf{Q}_\alpha^{\mathsf{Y}|\mathsf{D}} \overset{\mathbb{P}}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}}$

We will show $\mathsf{Q}_\alpha^{\mathsf{DY}}$ is unchanged by finite permutations of $(\mathsf{D}_i, \mathsf{Y}_i)$ pairs. For some finite permutation $\rho : \mathbb{N} \to \mathbb{N}$:

$$\mathsf{Q}_\alpha^{\mathsf{D}_\rho \mathsf{Y}_\rho} = \mathsf{Q}_\alpha^{\mathsf{D}_\rho \mathsf{Y}_\rho}(\mathrm{Swap}_{\rho,D^{\mathbb{N}}} \otimes \mathrm{Swap}_{\rho,Y^{\mathbb{N}}})$$

$$= \mathsf{Q}_\alpha^{\mathsf{D}} \odot \mathsf{Q}_\alpha^{\mathsf{Y}|\mathsf{D}}(\mathrm{Swap}_{\rho,D^{\mathbb{N}}} \otimes \mathrm{Swap}_{\rho,Y^{\mathbb{N}}})$$

(B.17)

(B.18)

(B.19)

$$= \mathsf{Q}_\alpha^{\mathsf{DY}}$$

Where line (B.17) follows from exchange commutativity, (B.18) follows from Theorem 2.2.6 and the fact that the swap map is deterministic and line (B.19) comes from the exchangeability of $\mathsf{Q}_\alpha^{\mathsf{D}}$.

Because $\mathbb{P}_\alpha^{\mathsf{D}}$ is dominated by $\mathsf{Q}_\alpha^{\mathsf{D}}$ by assumption, we have $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}} \overset{\mathbb{P}}{\cong} \mathsf{Q}_\alpha^{\mathsf{Y}|\mathsf{D}}$, which implies $\mathsf{Q}_\alpha^{\mathsf{Y}_i|\mathsf{DY}_{\{i\}^\complement}} \overset{\mathbb{P}}{\cong} \mathsf{Q}_\alpha^{\mathsf{Y}_i|\mathsf{DY}_{\{i\}^\complement}}$ and from Lemma B.4.1 we therefore have $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{HD}}$ IO contractible over

H, and from Theorem 4.3.21 we have $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathrm{id}_C | (\mathsf{D}, \mathsf{H})$ and so $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{HD}}$ IO contractible over H also. □

Theorem 4.3.26. *A data-independent sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *features conditionally independent and identical response functions* $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{G}}$ *with* D *infinitely supported over* G *only if for any sets* $A, B \subset \mathbb{N}$ *such that* D_A *and* D_B *are also infinitely supported over* G *and any* $i, j \in \mathbb{N}$ *such that* $i \notin A$, $j \notin B$,

$$\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{Y}_A,\mathsf{D}_A} = \mathbb{P}_\alpha^{\mathsf{Y}_j|\mathsf{D}_j|RV\mathsf{Y}_B\mathsf{D}_B}$$

. If in addition each $\mathbb{P}_\alpha^{\mathsf{YD}}$ *is dominated by some* $\mathbb{Q}_\alpha$ *such that* $\mathbb{Q}_\alpha^{\mathsf{YD}}$ *is exchangeable, then the reverse implication also holds.*

Proof. Only if: By Theorem 4.3.21 and Lemma 4.3.22, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{HD}}$ is IO contractible. By Theorem 4.3.19, H is almost surely a function of both $(\mathsf{D}_A, \mathsf{Y}_A)$ and $(\mathsf{D}_B, \mathsf{Y}_B)$ and, furthermore, $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{D}_A, \mathsf{Y}_A) | (\mathsf{D}_i, \mathsf{H}, \mathrm{id}_C)$, $\mathsf{Y}_j \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{D}_B, \mathsf{Y}_B) | (\mathsf{D}_j, \mathsf{H}, \mathrm{id}_C)$. Hence there is some $f : D^{\mathbb{N}} \times Y^{\mathbb{N}} \to H$ such that for all $E \in \mathcal{Y}, d_i \in D, d \in D^{\mathbb{N}}, y \in Y^{\mathbb{N}}$

$$\begin{aligned}
\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{Y}_A,\mathsf{D}_A}(E|d_i,y,d) &= \mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}}(E|d_i,f(y,d)) \\
&= \mathbb{P}_\alpha^{\mathsf{Y}_j|\mathsf{D}_j\mathsf{H}}(E|d_i,f(y,d)) \qquad\text{(B.20)} \\
&= \mathbb{P}_\alpha^{\mathsf{Y}_j|\mathsf{D}_j\mathsf{Y}_B,\mathsf{D}_B}(E|d_i,y,d)
\end{aligned}$$

Where Eq. (B.20) follows from Theorem 4.3.7.

If: By construction

$$\mathbb{Q}_\alpha^{\mathsf{Y}_i\mathsf{D}_i\mathsf{Y}_{\{i^{\complement}\}}\mathsf{D}_{\{i^{\complement}\}}} := \mathbb{Q}_\alpha^{\mathsf{D}_i\mathsf{Y}_{\{i^{\complement}\}}\mathsf{D}_{\{i^{\complement}\}}} \odot \mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{Y}_{\{i^{\complement}\}},\mathsf{D}_{\{i^{\complement}\}}}$$

is exchangeable, and by domination $\mathbb{Q}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{Y}_{\{i^{\complement}\}},\mathsf{D}_{\{i^{\complement}\}}} \overset{\mathbb{P}}{\cong} \mathbb{Q}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{Y}_{\{i^{\complement}\}},\mathsf{D}_{\{i^{\complement}\}}}$. The result follows from Lemma B.4.1. □

B.5   Proofs for data-dependent models

This section presents the proofs of all theorems and lemmas in Section 4.5.2

Theorem 4.5.11. *A Markov kernel* $\mathbb{K} : X^{\mathbb{N}} \nrightarrow Y^{\mathbb{N}}$ *is* IO contractible *if and only if for every* $n \in \mathbb{N}$ *and every* $A \subset \mathbb{N}$ *there exists some* $\mathbb{L} : X^n \nrightarrow Y^n$ *such that*

$$\mathbb{K}marg_A = swap_{A \to [n]}\mathbb{L} \otimes del_{X^{\mathbb{N}}}$$

Proof. Only if: By exchange commutativity

$$\mathrm{Swap}_{A \to [n]}\mathbb{K} = \mathbb{K}swap_{A \to [n]}$$

multiply both sides by $swap_{[n] \to A}$ on the right and, because $swap_{[n] \to A}$ is the inverse of $swap_{A \to [n]}$,

$$swap_{A \to [n]}\mathbb{K}swap_{[n] \to A} = \mathbb{K}$$

so

$$\mathbb{K}\mathrm{marg}_A = \mathrm{swap}_{A\to[n]}\mathbb{K}\mathrm{swap}_{[n]\to A}\mathrm{marg}_A$$
$$= \mathrm{swap}_{A\to[n]}\mathbb{K}\mathrm{marg}_{[n]}$$

By locality, there exists some $\mathbb{L} : X^n \nrightarrow Y^n$ such that

$$\mathbb{K}\mathrm{marg}_{[n]} = \mathbb{K}(\mathrm{id}_{[n]} \otimes \mathrm{del}_{X^{\mathbb{N}}})$$
$$= \mathbb{L} \otimes \mathbf{del}_{X^{\mathbb{N}}}$$

thus

$$\mathbb{K}\mathrm{marg}_A = \mathrm{swap}_{A\to[n]}\mathbb{L} \otimes \mathrm{del}_{X^{\mathbb{N}}}$$

as desired. 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

$$\mathrm{swap}_{A\to[n]}\mathbb{K}\mathrm{marg}_A = \mathrm{swap}_{A\to[n]}\mathbb{K}\mathrm{swap}_{[n]\to A}(\mathrm{id}_{[n]} \otimes \mathrm{del}_{X^{\mathbb{N}}})$$
$$= \mathbb{K}\mathrm{marg}_{[n]}$$
$$= \mathbb{K}(\mathrm{id}_{[n]} \otimes \mathrm{del}_{X^{\mathbb{N}}})$$

Then by Lemma B.1.2

$$\mathrm{swap}_{A\to[n]}\mathbb{K}\mathrm{swap}_{[n]\to[n]} = \mathbb{K}$$

as desired.

Consider an arbitrary finite permutation $\rho : \mathbb{N} \to \mathbb{N}$. 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 $\rho : \mathbb{N} \to \mathbb{N}$

$$\mathrm{swap}_{\rho}\mathbb{K}\mathrm{swap}_{\rho^{-1}} = \mathbb{K}$$

$\square$

Lemma 4.5.12. *A Markov kernel $\mathbb{K} : X^{\mathbb{N}} \nrightarrow Y^{\mathbb{N}}$ is IO contractible if and only if there exists a column exchangeable probability distribution $\mu \in \Delta(Y^{|X|\times\mathbb{N}})$ such that*

$$\Longleftrightarrow$$

$$\mathbb{K}(\underset{i\in\mathbb{N}}{\bigtimes} A_i | (x_i)_{i\in\mathbb{N}}) = \mu(\underset{i\in\mathbb{N}}{\bigtimes} Y^{x_i-1} \times A \times Y^{|X|-(x_i+1)})\forall A_i \in \mathcal{Y}$$

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

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

Proof. Only if: Choose $e := (e_i)_{i\in\mathbb{N}}$ such that $e_{i+|X|j}$ is the ith element of X for all $i, j \in \mathbb{N}$.

Define

$$\mu\left(\underset{(i,j)\in X\times\mathbb{N}}{\times} A_{ij}\right) := \mathbb{K}\left(\underset{(i,j)\in X\times\mathbb{N}}{\times} A_{ij}|e\right)\forall A_{ij}\in \mathcal{Y}$$

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

$$\mathbb{Q} : X^{\mathbb{N}} \twoheadrightarrow Y^{\mathbb{N}}$$

and consider some $A \subset \mathbb{N}$, $|A| = n$ and let $B := (x_i, i))_{i\in A}$. For arbitrary $\{C_i \in \mathcal{Y}|i \in A\}$, define $C_A := \mathrm{swap}_{[n]\to A}(\times_{i\in[n]}C_i \times Y^{\mathbb{N}})$; i.e. the Cartesian product with the C_is occupying the positions indexed by A and Y occupying the rest of the positions. Then, for arbitrary $x \in X^{\mathbb{N}}$

$$\mathbb{Q}(C_A|x) = \mu(\mathrm{lu}_x^{-1}(C_A)) \tag{B.21}$$

where $\mathrm{lu}_x^{-1}(C_A) = \{y^X \in Y^{X\times\mathbb{N}}|\mathrm{lu}(x, y^X) \in C_A\}$.

The argument of μ is

$$\begin{aligned}
\mathrm{ev}_x^{-1}(C_A) &= \{(y_{ji})_{j,i\in X\times\mathbb{N}}|(y_{x_i i})_{i\in\mathbb{N}} \in C_A\} \\
&= \underset{i\in\mathbb{N}}{\times}\underset{j\in X}{\times} V_{ji}
\end{aligned}$$

where

$$V_{ji} = \begin{cases} C_i & (j,i) \in B \\ Y & \text{otherwise} \end{cases}$$

let $\xi : D \times \mathbb{N} \to \mathbb{N}$ be the map that sends (j, i) to its rank in the enumeration of B. Then

$$\mathrm{swap}_{\xi(B)\to A}(\mathrm{lu}_x^{-1}(C_A)) = C_A \tag{B.22}$$

Substituting Equation (B.22) into (B.21)

$$\begin{aligned}
\mathbb{Q}(C_A|x) &= \mu\mathrm{swap}_{\xi(B)\to A}(C_A) \\
&= \mathbb{K}\mathrm{swap}_{\xi(B)\to A}(C_A|e) \\
&= \mathbb{K}\mathrm{swap}_{\xi(B)\to A}(C_A|(f_i)_{i\in\mathbb{N}}) \qquad \text{by locality}
\end{aligned}$$

where

$$f_{ji} = \begin{cases} e_{x_i i} & i \in A \\ x_i & \text{otherwise} \end{cases}$$

but by construction of e, this means $(f_i)_{i \in \mathbb{N}} = \text{swap}_{A \to \zeta(B)}(x)$. Thus

$$= \mathbb{K}\text{swap}_{\zeta(B) \to A}(C_A|(x))$$
$$= \text{swap}_{A \to \zeta(B)}\mathbb{K}\text{swap}_{\zeta(B) \to A}(C_A|x)$$
$$= \mathbb{K}(C_A|x) \qquad \qquad \text{by commutativity of exchange}$$

Because this holds for all x, $A \subset \mathbb{N}$, by Lemma B.1.2

$$\mathbb{Q} = \mathbb{K}$$

Next we will show μ is column exchangeable. Consider any column swap $\text{swap}_c : X \times \mathbb{N} \to X \times \mathbb{N}$ 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 $A \in \mathcal{Y}^{\mathbb{N}}$

$$\mathbb{K}(A|e) = \text{swap}_{c^{-1}}\mathbb{K}\text{swap}_c(A|e)$$
$$= \mathbb{K}\text{swap}_c(A|\text{swap}_{c^{-1}}(e))$$
$$= \mathbb{K}\text{swap}_c(A|e)$$

If: Suppose

$$\mathbb{K} = \quad \begin{array}{c} \mathsf{W} - \boxed{\mathbb{P}_\alpha^{Y^D|W}} \\ \qquad \qquad \searrow \boxed{\mathbb{F}_{\text{lu}}} \quad \text{---} \; \mathsf{Y} \\ \mathsf{D} \; \text{------} \nearrow \end{array}$$

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 $A_S = \text{swap}_{[n] \to S} \times_{i \in [n]} A_i \times Y^{\mathbb{N}}$, $A_T = \text{swap}_{S \to T}(A_S)$, $B = (x_i i)_{i \in S}$ and $C = (x_i i)_{i \in T} = (x_{\text{swap}_{S \to T}}(i) i)_{i \in S}$. By Equations (B.21) and (B.22)
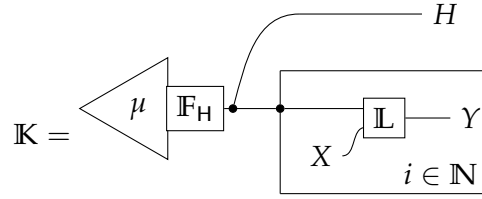
$$\mathbb{K}(A_S|x) = \mu\text{swap}_{S \to B}(A_S)$$
$$= \mu\text{swap}_{T \to C}(A_T) \qquad \qquad \text{by column exchangeability of } \mu$$
$$= \mathbb{K}(A_T|\text{swap}_{S \to T}(x))$$
$$= \text{swap}_{S \to T}\mathbb{K}(A_T|x)$$
$$= \text{swap}_{S \to T}\mathbb{K}\text{swap}_{S \to T}(A_S|x)$$

so $\mathbb{K}$ is IO contractible by Theorem 4.5.11. $\qquad \square$

Lemma 4.5.13. *Given* $\mathbb{K} : X^{\mathbb{N}} \to Y^{\mathbb{N}}$, *X and Y standard measurable, if*

$$\mathbb{K} = \quad \begin{array}{c} \triangleleft \mu \\ \qquad \searrow \\ \mathsf{D} \; \text{---} \boxed{\mathbb{F}_{\text{lu}}} \text{------} \; \mathsf{Y} \end{array}$$

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 $\mathsf{H} : Y^{X \times \mathbb{N}} \to H$ and $\mathbb{L} : H \times X \rightsquigarrow Y$ such that

Proof. As a preliminary, we will show

$$(B.23)$$

where $\mathbf{evs}_{Y^X \times X} : Y^X \times X \to Y$ is the single-shot evaluation function

$$(x, (y_i)_{i \in X}) \mapsto y_x$$

Recall that $\mathbf{ev}$ is the function

$$((x_i)_{\mathbb{N}}, (y_{ji})_{j,i \in X \times \mathbb{N}}) \mapsto (y_{x_i i})_{i \in \mathbb{N}}$$

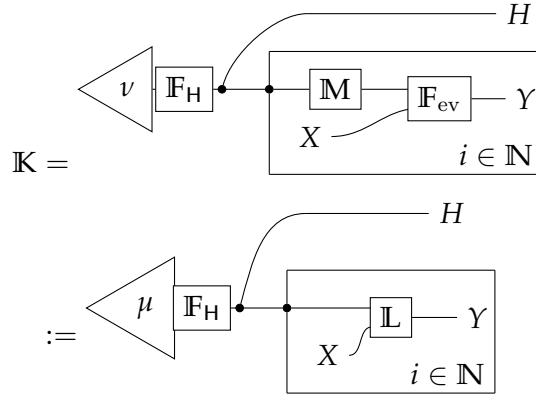By definition, for any $\{A_i \in \mathcal{Y} | i \in \mathbb{N}\}$

$$
\begin{aligned}
\mathbb{F}_{ev}(\underset{i \in \mathbb{N}}{\times} A_i | (x_i)_{\mathbb{N}}, (y_{ji})_{i \in X \times \mathbb{N}}) &= \delta_{(y_{x_i i})_{i \in \mathbb{N}}}(\underset{i \in \mathbb{N}}{\times} A_i) \\
&= \prod_{i \in \mathbb{N}} \delta_{y_{x_i i}}(A_i) \\
&= \prod_{i \in \mathbb{N}} \mathbb{F}_{evs}(A_i | x_i, (y_{ji})_{j \in X}) \\
&= \left(\bigotimes_{i \in \mathbb{N}} \mathbb{F}_{evs} \right) (\underset{i \in \mathbb{N}}{\times} A_i | (x_i)_{\mathbb{N}}, (y_{ji})_{j \in X \times \mathbb{N}})
\end{aligned}
$$

which is what we wanted to show.

Define $\mathbb{M} : H \rightsquigarrow Y^X$ 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 (2005b, Prop. 1.4) there is a directing random measure $\mathsf{H} : Y^{X \times \mathbb{N}} \to H$ such that

$$\Longleftrightarrow$$

$$\mu(\underset{i \in \mathbb{N}}{\times} A_i \times B) = \int_B \prod_{i \in \mathbb{N}} \mathbb{M}(A_i|h) \mu \mathbb{F}_{\mathsf{H}}(\mathrm{d}h) \qquad \forall A_i \in \mathcal{Y}^X$$

By Equations (B.24) and (B.23)

Where we can connect the copied outputs of $\mu \mathbb{F}_H$ to the inputs of each $\mathbb{M}$ "inside the plate" as the plates in Equations (B.10) and (B.11) are equal in number and each connected wire represents a single copy of Y^D. $\qquad\square$

Theorem 4.5.14. *Given a kernel* $\mathbb{K} : X^{\mathbb{N}} \to Y^{\mathbb{N}}$, *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* $\mathbb{L} : H \times X \to Y$ *such that*

$$\Longleftrightarrow$$

$$\mathbb{K}(\underset{i\in\mathbb{N}}{\bigtimes} A_i | (x_i)_{i\in\mathbb{N}}) = \int_H \prod_{i\in\mathbb{N}} \mathbb{L}(A_i | h, x_i) \mu(\mathrm{d}h)$$

Proof. Only if: By Lemma 4.5.12, we can represent the conditional probability $\mathbb{K}$ as

$$(B.24)$$

where $\nu \in \Delta(Y^{X \times \mathbb{N}})$ is column exchangeable.

Applying Lemma 4.5.13 yields the desired result.

If: By assumption, for any $\{A_i \in \mathcal{Y} | i \in \mathbb{N}\}$, $x := (x_i)_{i\in\mathbb{N}} \in X^{\mathbb{N}}$

$$\mathbb{K}(\underset{i\in\mathbb{N}}{\bigtimes} A_i | x) = \int_H \prod_{i\in\mathbb{N}} \mathbb{L}(A_i | h, x_i) \mu(\mathrm{d}h)$$

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 $A_T := \times_{i\in\mathbb{N}} B_{\mathrm{swap}_{S\to T}(i)}$.

$$\mathbb{K}(A_S|x) = \int_H \prod_{i \in S} \mathbb{L}(A_i|h, x_i)\mu(\mathrm{d}h)$$

$$= \int_H \prod_{i \in T} \mathbb{L}(A_i|h, x_{\mathrm{swap}_{S \to T}(i)})\mu(\mathrm{d}h)$$

$$= \mathrm{swap}_{S \to T}\mathbb{K}(A_T|x)$$

$$= \mathrm{swap}_{S \to T}\mathbb{K}\mathrm{Swap}_{T \to S}(A_S|x)$$

So by Theorem 4.3.7, $\mathbb{K}$ is IO contractible. $\square$

Theorem 4.5.15. *Given a sequential input-output model $(\mathbb{P}'_C, \mathsf{D}', \mathsf{Y}')$ on $(\Omega, \mathcal{F})$, then $\mathbb{P}'^{\mathsf{Y}' \wr \mathsf{D}'}_C$ 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 $\mathsf{G} : \Omega \times G \to G$ such that $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}'_C} (\mathsf{Y}_{<i}, \mathsf{X}_{<i}, \mathsf{C})|(\mathsf{X}_i, \mathsf{G})$ and $\mathbb{P}^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{G}}_C = \mathbb{P}^{\mathsf{Y}_j|\mathsf{X}_j\mathsf{G}}_C$ for all $i, j \in \mathbb{N}$ and $\mathsf{G} \perp\!\!\!\perp_{\mathbb{P}_C} (\mathsf{X}, id_C)$.*

Proof. If: By assumption, there is some $\mathbb{L} : G \times D \to Y$ such that

$$\mathbb{P}^{\mathsf{Y}_i|\mathsf{G}\mathsf{D}_i}_C = \mathbb{L}$$

and $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{<i}, \mathsf{D}_{<i})|(\mathsf{D}_i, \mathsf{G})$. Thus

$$\mathbb{P}^{\mathsf{Y}_i|\mathsf{G}\mathsf{D}_i\mathsf{Y}_{<i}\mathsf{D}_{<i}}_C = \mathbb{L} \otimes \mathrm{erase}_{Y^{i-1} \times D^{i-1}}$$

and so

$$\mathbb{P}^{\mathsf{Y} \wr \mathsf{D}}_C = \text{} \tag{B.25}$$

and so by Theorem 4.5.14, $\mathbb{P}^{\mathsf{Y} \wr \mathsf{D}}_C$ is IO contractible.

Only if: First, define the extension $\mathbb{P}_C$. From Theorem 4.5.14 and IO contractibility of $\mathbb{P}'^{\mathsf{Y}' \wr \mathsf{D}'}_C$ there is some set G, $\mu \in \Delta(G)$ and $\mathbb{L} : G \times D \to Y$ such that

$$\mathbb{P}'^{\mathsf{Y}' \wr \mathsf{D}'}_C = \text{}$$

thus, by the definition of the comb insert operation

$$\mathbb{P}'^{\mathsf{D}'_{[n]}\mathsf{Y}'_{[n]}}_\alpha = \mathbb{P}'^{\mathsf{D}_1}_\alpha \odot \mathrm{insert}\big(\mathbb{P}'^{\mathsf{D}'_{[2,n]} \wr \mathsf{Y}'_{[n-1]}}_\alpha, \mathbb{P}'^{\mathsf{Y}'_{[n]} \wr \mathsf{D}'_{[n]}}_C\big)$$

Let

$$\mathbb{P}^{\mathsf{Y}_i|\mathsf{G}\mathsf{D}_i}_C = \mathbb{L} \tag{B.26}$$

and let $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{<i}, \mathsf{D}_{<i})|(\mathsf{D}_i, \mathsf{G})$, and for all α set $\mathbb{P}^{\mathsf{W}|\mathsf{D}\mathsf{Y}}_\alpha = \mathbb{P}'^{\mathsf{W}'|\mathsf{D}'\mathsf{Y}'}_\alpha$ for all $\mathsf{W}' : \Omega \to W$ and $\mathbb{P}^{\mathsf{D}_i|\mathsf{Y}_{<i}\mathsf{D}_{<i}}_\alpha = \mathbb{P}'^{\mathsf{D}'_i|\mathsf{Y}'_{<i}\mathsf{D}''_{<i}}_\alpha$.

It remains to be shown that $\mathbb{P}^{\mathsf{D}\mathsf{Y}}_\alpha = \mathbb{P}'^{\mathsf{D}\mathsf{Y}}_\alpha$.

By Equation (B.26) and $Y_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (Y_{<i}, D_{<i})|(D_i, G)$, it follows (for identical reasons as Equation (B.25)) that

$$\mathbb{P}_C^{Y \wr D} = $$

$$=$$

$$= \mathbb{P}_C'^{Y' \wr D'}$$

And so for all $n \in \mathbb{N}$

$$\mathbb{P}_\alpha^{D_{[n]} Y_{[n]}} = \mathbb{P}_\alpha^{D_1} \odot \mathrm{insert}\big(\mathbb{P}_\alpha^{D_{[2,n]} \wr Y_{[n-1]}}, \mathbb{P}_C'^{Y_{[n]} \wr D_{[n]}}\big)$$

$$= \mathbb{P}_\alpha^{D_1} \odot \mathrm{insert}\big(\mathbb{P}_\alpha'^{D'_{[2,n]} \wr Y'_{[n-1]}}, \mathbb{P}_C'^{Y'_{[n]} \wr D'_{[n]}}\big)$$

$$= \mathbb{P}_\alpha'^{D'_{[n]} Y'_{[n]}}$$

$\square$

Appendix C

Proofs of key results in Chapter 5

C.1 Proofs related to causal Bayesian networks

Theorem 5.1.15. *Given an uncertain unrolled causal Bayesian network* $(\mathbb{P}., C, \Omega, (V_{ij})_{i \in A, j \in [n]}, H, G)$, *take* $C' \subset C$ *to be sequences of interventions that, for some* $i \in A$, *do not target a particular* V_{ij} *for any* $j \in [n]$ *and ensure every sequence* $V_{j[n]}$ *has infinite support. Then* $V_{i[n]} \perp\!\!\!\perp^e_{\mathbb{P}_{C'}} id_C | (H, \text{Pa}(V_{i[n]}))$ *and* $\mathbb{P}_C^{V_{i[n]}|\text{HPa}(V_{i[n]})}$ *is IO contractible over* H.

Proof. First we will prove $V_{i[n]} \perp\!\!\!\perp^e_{\mathbb{P}_{C'}} id_C | (H, \text{Pa}(V_{i[n]}))$. 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 5.1.13, for each $h \in H$

$$V_{Aj} \perp\!\!\!\perp^e_{\mathbb{P}_{\alpha,h}} V_{A[n]\setminus\{j\}} | id_C$$

which implies

$$
\begin{aligned}
& V_{Aj} \perp\!\!\!\perp^e_{\mathbb{P}_\alpha} V_{A[n]\setminus\{j\}} | (id_C, H) \\
&\implies V_{ij} \perp\!\!\!\perp^e_{\mathbb{P}_\alpha} V_{A[n]\setminus\{j\}} | (H, \text{Pa}(V_{i[n]}), id_C)
\end{aligned}
\tag{C.1}
$$

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 js with ks 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 $\mathbb{P}_C^{V_{i[n]}|\text{HPa}(V_{i[n]})}$ over H. By Eq. (C.1)

$$V_{ij} \perp\!\!\!\perp^e_{\mathbb{P}_\alpha} (V_{i[1,j)}, \text{Pa}(V_{i[1,j)})) | (H, \text{Pa}(V_{i[n]}), id_C)$$

furthermore, by 3* and the assumption that no intervention $\alpha \in C'$ targets V_{ij} for any j, for any $\alpha \in C'$

$$\mathbb{P}_\alpha^{\mathsf{V}_{ij}|\mathsf{HPa}(\mathsf{V}_{ij})}(A|h,y) = \mathbb{P}_\alpha^{\mathsf{V}_{ik}|\mathsf{HPa}(\mathsf{V}_{ik})}(A|h,y)$$

thus $\mathbb{P}_C$ has independent and identical response functions conditional on H, by assumption the inputs have infinite support and therefore by Theorem 4.3.21, $\mathbb{P}_C^{\mathsf{V}_{i[n]}|\mathsf{HPa}(\mathsf{V}_{i[n]})}$ is IO contractible over H. $\square$

C.2 Proofs related to precedent

Theorem 5.1.21. *Suppose we have a probability set $\mathbb{P}_C$ on $(\Omega, \mathcal{F})$ with variables $\mathsf{Y} := (\mathsf{Y}_i)_{i\in\mathbb{N}}$, $\mathsf{X} := (\mathsf{X}_i)_{i\in\mathbb{N}}$ and $\mathsf{Z}_i := (\mathsf{Z}_i)_{i\in\mathbb{N}}$, with X, Y, Z discrete and n-precedented with indexing variable T. Let G be the directing random conditional of $(\mathsf{P}_C, \mathsf{T}, \mathsf{Z})$ and H the directing random conditional of $(\mathbb{P}_C, (\mathsf{T}, \mathsf{Z}), (\mathsf{X}, \mathsf{Y}))$. Suppose there is some $N \in \mathcal{H}$ such that $\mathbb{P}_\alpha^{\mathsf{H}}(N) > 0$ for all α, and define the probability conditioned on $\mathsf{H} \in N$:*

$$probP_{\alpha,\mathsf{H}\in N}(A) := \frac{\mathbb{P}_\alpha^{id_\Omega\mathsf{H}}(A \times N)}{\mathbb{P}_\alpha^{\mathsf{H}}(N)}$$

Suppose for all i, $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_{C,\mathsf{H}\in N}} \mathsf{Z}_i|(\mathsf{H},\mathsf{X}_i)$, and for all $t \in T$ and $z, z' \in Z$,

$$\mathbb{P}_{\alpha,\mathsf{H}\in N}^{\mathsf{G}_z^t|\mathsf{HG}_{z'}^t}(\cdot|h, g_{z'}^t) \ll U_{[0,1]}$$

where $U_{[0,1]}$ is the uniform probability measure on $([0,1], \mathcal{B}([0,1]))$.

Then $\mathbb{P}_{\alpha,\mathsf{H}\in N}^{\mathsf{Y}|\mathsf{XH}}$ is IO contractible over H.

Proof. For arbitrary $h \in N$, define matrices k and l by

$$\mathbb{P}_{\alpha,\mathsf{H}\in N}^{\mathsf{Y}_i|\mathsf{HT}_i\mathsf{Z}_i\mathsf{X}_i}(y|h,t,z,x) \overset{\mathbb{P}_{\alpha,\mathsf{H}\in N}}{\cong} : k_{tzx}^y$$

$$\mathbb{P}_{\alpha,\mathsf{H}\in N}^{\mathsf{X}_i|\mathsf{HT}_i\mathsf{Z}_i}(x|h,t,z) \overset{\mathbb{P}_{\alpha,\mathsf{H}\in N}}{\cong} : l_{tz}^x$$

noting that both *RVK* and l are almost surely deterministic functions of H.

The assumption $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Z}_i|(\mathsf{H},\mathsf{X}_i, id_C)$ implies, for $\mathbb{P}_\alpha$-almost all α, z, z', x, y

$$\sum_{t\in T} k_{tzx}^y \frac{l_{tz}^x g_z^t}{\sum_{t'\in T} l_{t'z}^x g_z^{t'}} \overset{\mathbb{P}_C}{\cong} \sum_{t\in T} k_{tz'x}^y \frac{l_{tz'}^x g_{z'}^t}{\sum_{t'\in T} l_{t'z'}^x g_{z'}^{t'}} \tag{C.2}$$

For fixed k, l and g_z^t, Eq. (C.2) defines a polynomial constraint on $g_{z'}^t$. We will show that, unless $k_{tzx}^y = k_{t'zx}^y$ for all t, t' and z, that this constraint is nontrivial for some z'. Consequently, the set of solutions subject to the restriction $k_{tzx}^y \neq k_{t'zx}^y$ has Lebesgue measure 0.

For arbitrary t, assume $k_{tz'x}^y > k_{t<z'x}^y$ for some $t^<$.

Then either $l_{tz'}^x = l_{t<z'}^x$, $l_{tz'}^x < l_{t<z'}^x$ or $l_{tz'}^x = l_{t>z'}^x$. Consider the first case, and take g' such that $g_{z'}'^t = g_{z'}^t - \epsilon$ and $g_{z'}'^{t<} = g_{z'}^{t<} + \epsilon$ and equal to g otherwise. There is almost surely some ϵ such

that g' is a Markov kernel, as $g_{z'}^t > 0$ almost surely. Then

$$\frac{l_{tz}^x g_z^t}{\sum_{t' \in T} l_{t'z}^x g_z^{t'}} > \frac{l_{tz}^x g_z'^t}{\sum_{t' \in T} l_{t'z}^x g_z'^{t'}}$$

$$\frac{l_{t<z}^x g_z^{t<}}{\sum_{t' \in T} l_{t'z}^x g_z^{t'}} < \frac{l_{t<z}^x g_z'^{t<}}{\sum_{t' \in T} l_{t'z}^x g_z'^{t'}}$$

because by assumption the denominator remains the same. But then

$$\sum_{t \in T} k_{tzx}^y \frac{l_{tz}^x g_z^t}{\sum_{t' \in T} l_{t'z}^x g_z^{t'}} > \sum_{t \in T} k_{tz'x}^y \frac{l_{tz'}^x g_{z'}'^t}{\sum_{t' \in T} l_{t'z'}^x g_{z'}'^{t'}} \tag{C.3}$$

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 $l_{tz'}^x > l_{t<z'}^x$. Then we still have

$$\frac{l_{tz}^x g_z^t}{\sum_{t' \in T} l_{t'z}^x g_z^{t'}} > \frac{l_{tz}^x g_z'^t}{\sum_{t' \in T} l_{t'z}^x g_z'^{t'}}$$

$$\frac{l_{t<z}^x g_z^{t<}}{\sum_{t' \in T} l_{t'z}^x g_z^{t'}} < \frac{l_{t<z}^x g_z'^{t<}}{\sum_{t' \in T} l_{t'z}^x g_z'^{t'}}$$

For the first inequality, both the numerator and the denominator shrink. For the second, note that

$$\frac{l_{t<z}^x g_z^{t<}}{\sum_{t' \in T} l_{t'z}^x g_z^{t'}} = \frac{l_{t<z}^x (g_z^{t<} + \epsilon)}{(1 + \frac{\epsilon}{g_z^{t'}}) \sum_{t' \in T} l_{t'z}^x g_z^{t'}}$$

$$< \frac{l_{t<z}^x (g_z^{t<} + \epsilon)}{\sum_{t' \in T} l_{t'z}^x g_z^{t'}}$$

and so the conclusion in Eq. (C.3) follows for the same reasons. Finally considering $l_{tz'}^x < l_{t<z'}^x$, analogous reasoning implies Eq. (C.3) once again.

Thus, unless $k_{tzx}^y = k_{t'zx}^y$ for all t, t' and z, Eq. (C.2) implies a nontrivial constraint on $g_{z'}^t$ for some z'. Thus, for some z', x, t, t', y the set of solutions $S := \{g_{z'}^t | \text{Eq. (C.2) is satisfied} \wedge k_{tzx}^y \neq k_{t'zx}^y\}$ has Lebesgue measure 0 in the set $[0,1]^D$ (Okamoto, 1973), and so

$$\mathbb{P}_{\alpha, H \in N}^{G_z^t | HG_{z'}^t}(S | h, g_z^t) = 0$$

by the assumption that this probability is dominated by the Lebesgue measure. On the other hand, by assumption, the set $B := \{g_{z'}^t | \text{Eq. (C.2) is satisfied}\}$ has measure 1. Thus we conclude that $k_{tzx}^y = k_{t'zx}^y$ with probability 1. That is, $Y_i \perp\!\!\!\perp_{\mathbb{P}_{C,H \in \mathbb{N}}}^e T_i | (H, Z_i, X_i, \text{id}_C)$. By contraction with $Y_i \perp\!\!\!\perp_{\mathbb{P}_{C,H \in \mathbb{N}}}^e Z_i | (H, X_i, \text{id}_C)$, we have $Y_i \perp\!\!\!\perp_{\mathbb{P}_{C,H \in \mathbb{N}}}^e (Z_i, T_i) | (H, X_i, \text{id}_C)$. Because we already know $\mathbb{P}_{\alpha}^{Y_i | HX_i Z_i}$ are identical for all i, this implies $\mathbb{P}_{\alpha, H \in N}^{Y_i | HX_i}$ are independent and identical response functions. $\qquad\square$