# Causal Inference Without Interventions

A Preprint

David O. Johnston[*]
Australian National University
davidoj@fastmail.com.au

Cheng Soon Ong
Data61

Robert C. Williamson
Universität Tübingen

May 22, 2023

## Abstract

Valid causal inferences from diverse data are a tantalizing but elusive prospect. Assumptions are necessary to make any inferences, and these assumptions cannot all be washed out by sufficiently large datasets. A decision maker must supply these assumptions, and believe in them. We argue that structural interventions can embody assumptions not entailed by typical prior knowledge or by the given data. Abandoning structural interventions as a foundation for models of decision making requires an alternative. We argue that a critical role of structural interventional models is to identify relationships between variables that are shared by data arising from observation and by data arising from the consequences of a decision. We analyse such assumptions of shared relationships directly. Our starting point is is to ground causal models in the problem of making good decisions, and our first result is an equivalence between decision models with ”sequences of identical conditionals” and decision models that feature a symmetry we call ”IO contractibility”. This can be seen as a generalisation of De Finetti’s work on exchangeability, itself an attempt to justify the conventional but seemingly mysterious assumption of sequences featuring a ”shared but unknown probability distribution”. We argue that IO contractibility is often an unreasonable assumption. Our second result is that IO contractibility may be implied by a combination of an observed conditional independence and a weaker assumption based on the principle of “precedent”, which requires that everything that might be done has already been done before in some way and has been seen to work. We discuss a connection between this latter result and constraint based causal discovery.

Keywords causal inference · decision theory

## 1 Introduction

Judea Pearl’s causal hierarchy distinguishes between three types of problems: prediction problems, intervention problems, and counterfactual problems [Pearl and Mackenzie, 2018]. Modelling an intervention problem requires a different kind of knowledge than that required for modelling predictions, and modelling a counterfactual problem requires a different kind of knowledge than that needed for modelling intervention problems.

While we think that Pearl’s hierarchy offers an important insight into the differences between causal inference and classical statistics, we feel the terminology of interventions is confusing. In Pearl’s theory, structural interventions are used to model what he describes as intervention problems. Structural interventions are operations that transform a probability distribution according to a graphical causal model. However, “intervention problems” refer to a broad class of problems that ask questions like “what will happen if I do this? what will happen if I do something else instead?” – problems that we frequently encounter whether or not we make use of graphical causal models. This kind of problem often comes up we want to make a decision;

[*]

given various options and some idea of the outcomes we would like to achieve, we want to know what the likely consequences of each option are.

As the terminology suggests, structural interventions and graphical causal models are often used as a tool for modelling the consequences of different options (that is, "structural interventional models" are used to model "intervention problems" broadly understood). If a decision maker wants to use structural interventions to model the consequences of different options, they must identify each of their options with a structural intervention. Even if they are able to learn many of the relevant causal relationships from a dataset – using, for example, a method of causal discovery – they usually will not learn the correspondence between each of their options $\alpha$ and a particular structural intervention. This correspondence is usually taken to be prior knowledge. However, we contend that such a correspondence often goes beyond the prior knowledge a decision maker is likely to have.

The most common kind of structural intervention is known as a perfect or hard intervention. A perfect intervention is often referred to with the symbol $\text{do}(\mathsf{X} = x)$. Given a probability distribution $\mathbb{P}$, a variable $\mathsf{X}$ and a causal graphical model $\mathcal{G}$ which, among other things, specifies a set of causal parents $\text{Pa}(\mathsf{X})$ of $\mathsf{X}$, the intervention $\text{do}(\mathsf{X} = x)$ yields a new probability distribution $\mathbb{P}'$ such that the conditional probability $\mathbb{P}'^{\mathsf{X}|\text{Pa}(\mathsf{X})}$ is everywhere $\delta_x$, while all other conditional distributions of a "child" conditional on its "parents" according to $\mathcal{G}$ match their counterparts in $\mathbb{P}$ [Pearl, 2009, Sec. 1.3.1]

Thus identifying a perfect intervention $\text{do}(\mathsf{X} = x)$ with an option $\alpha$ embodies a collection of assumptions – first, it embodies the assumption that selecting the option $\alpha$ will force future observations of the variable $\mathsf{X}$ to take the value $x$. This assumption may be justified where a decision maker knows at the outset how to deterministically control $\mathsf{X}$. However, the identification of options with perfect interventions also embodies the assumption that selecting the option $\alpha$ will leave all "parental conditionals" with respect to $\mathcal{G}$ unchanged with the exception of $\mathbb{P}'^{\mathsf{X}|\text{Pa}(\mathsf{X})}$. What kind of knowledge could a decision maker have that justifies these further assumptions?

One possibility is that the decision maker also knows that all of these conditionals will remain unchanged. For example, if the decision maker is a programmer working on a program represented by the graph $\mathcal{G}$ and is reasoning about changing a particular function to output a constant. However, this is not a typical situation where someone is interested in causal inference or causal discovery.

Alternatively, the decision maker might simply assume that their options correspond to perfect interventions (or some other default transformation) on on some unknown causal graph, and attempt to determine the correct graph via a combination of prior knowledge and causal discovery. The notion of a default correspondence between options and interventions is compatible with the common idea that the causal graph is all you need to answer key interventional questions. Success in causal discovery is typically measured by the structural intervention distance [Peters and Bühlmann, 2015] or the structural hamming distance, neither of which have any dependence on the options under consideration [Scherrer et al., 2022, Toth et al., 2022, Brouillard et al., 2020, Ng et al., 2019, Forré and Mooij, 2018, Zheng et al., 2018, Chickering, 2003, Spirtes et al., 1993]. The idea that the relationships captured by a causal graph are independent of the options under consideration is also defended in Pearl [2018].

The justification for a default choice of intervention is hard to see. One could imagine an extensive survey of decision problems concluding that a particular kind of intervention is nearly always suitable, but we are not aware of any such undertaking. Furthermore, it does not appear to be particularly difficult to come up with common decision problems where perfect interventions care inappropriate. Hernán and Taubman consider the example of different options that are known a priori to affect a person's body mass index, including diet plans, gastric bypass surgery and limb removal [Hernán and Taubman, 2008, Hernán, 2016]. These will all plausibly affect an individual's risk of death differently, and so they cannot all be modeled by the same kind of intervention on body mass index. Further, it seems to us (as well as other authors in this exchange [Pearl, 2018, Hernán and Cole, 2009, Shahar, 2009]) that none of these options stand out as a strong candidate to be identified with a perfect intervention on body mass index.

One possible response is to model consequences via perfect interventions on diet, gastric surgery or limb count. This approach raises similar concerns about interventions on the new variables (there's more than one way to raise the probability that someone adopts a diet). It also seems to undermine the project of learning useful causal structures from convenient datasets. In particular, it suggests that causal structures are of little use unless some special "intervenable" variables can be included in these structures.

So far our remarks have concerned perfect interventions, but there is a diverse array of structural interventions that can be found in the literature. A non-exhaustive review turns up perfect interventions or "hard" interventions [Pearl, 2009, Hauser and Bühlmann, 2012, ch. 1], soft interventions [Correa and Bareinboim, 2020, Eberhardt and Scheines, 2007], general or fat-hand interventions [Eberhardt and Scheines, 2007, Yang et al., 2018, Glymour and Spiegelman, 2017] and general interventions with unknown targets [Brouillard et al., 2020]. We could consider using any of these families of interventions to address our decision maker's dilemma: how to relate their actions and their causal graph to distributions over consequences. The richer family of interventions can address the problem of nonuniqueness above – some actions that affect the same target may be modeled as different types of intervention. While generalised interventions address the problem that it is impossible to represent all actions as perfect interventions, they still embody assumptions that go beyond the causal graph and , we note that the references cited for generalised interventions do not use them for the purposes of evaluating prospective actions However, the problem still seems to be a difficult one: our decision maker who may, in general, be quite ignorant about many of the details of the relevant causal graph, must say something about which action corresponds to which kind of intervention. While this problem may be solvable, we investigate the prospects for avoiding it altogether.

To clarify our claim: we doubt that models for data-driven decision making should rely on structural interventions as a primitive notion. We do not doubt that structural interventional models are practically useful in a wide range of cases. In some cases, interventions describe very precisely the outcomes that should be expected – for example, consider the consequences of altering a single function to output a constant value in a piece of computer software. In many more cases the idea of structural interventions are useful enough even where our understanding of why they provide a good model of actions is less precise. For example, causal graphical models have been widely used for understanding interactions between genes [Badsha and Fu, 2019]. Any approch to building models for data-driven decision making that avoids structural interventions as primitives probably needs to include structural interventions as a common and important special case.

We must make some assumptions to licence inferences about the consequences of our actions. It is not enough to assume that observations exhibit some regularity (for example, being independent and identically distributed) and that we have prior knowledge about how different actions control some variables. We want to leverage the observations to draw conclusions about how actions will affect variables which we do not know how to control at the outset. The assumption of invariant parental conditional distributions that are a part of structural interventional models allow us to make these inferences. If a decision maker wants to learn from some observed data so as to make a better decision, then these assumptions tell them what the observational data and the consequences of their decisions will have in common: the parental conditional distributions.

In this work, we ask what other kinds of assumptions might capture the idea that the consequences of our actions are somehow related to our observations. We draw inspiration from the classic statistical assumption of exchangeability, which expresses the idea that observations taken at different points in time are (in a particular way) just like one another. While it does capture a sense in which observations after taking an action are related to observations made before taking an action, it is not a useful assumption for deceision making. In decision problems, future events are affected by the choices made by the decision maker and therefore are not just like observations taken prior to making a decision.

We investigate two modified assumptions motivated by the idea of exchangeability. First, we introduce the idea of input-output contractibility (IO contractibility), which can be viewed as a generalisation of exchangeability to decision models. In particular, we show a theorem analogous to De Finetti's famous representation [de Finetti, [1937] 1992] theorem holds; IO contractibility is equivalent to the assumption that there is a shared but unknown input-output map for both the observations and the consequences of a decision maker's choices (Theorem 3.18). Unlike exchangeability, however, IO contractibility is not an appealing assumption in many data-driven decision problems.

We subsequently explore the related idea of precedent, which holds that every option a decision maker has has been exercised before, and the consequences observed, though they generally don't know exactly when each option has been exercised or its consequences observed. We show that under the assumption of diverse precedent, based on the principle of precedent, conditional independences imply the stronger conclusion of IO contractibility and with it the possibility of estimating an input-output conditional distribution from the given observational data (Theorem 4.7). We discuss, speculatively, how the assumption of diverse precedent may be related to causal structures, noting similar assumptions that appear in the causal discovery literature [Meek, 1995, Janzing, 2021].

Section 2 introduces the formalism of decision models. These differ from probability models in that they are a map from a set of options to distributions over consequences. We make use of the notion of extended conditional independence introduced by Constantinou and Dawid [2017], which is a notion of conditional independence relevant to decision models. Section 3 introduces the idea of shared conditionally independent and identical responses, and shows that this is equivalent to the assumption of input-output in Theorem 3.18. Section 4 explains the assumption of precedent and proves Theorem 4.7 and discusses the interpretation of this assumption and its connection to structural causal models.

### 1.1 Previous work on symmetries in causal inference

Our approach starts with the assumption that we are trying to model a decision problem, and not address any other kind of problem that comes under the umbrella of causal inference. This assumption motivates the formalism of "decision models" that we use to investigate the questions raised here. The broad idea of starting with the options available to a decision maker rather than starting with some foundational notion of causation is often called the decision theoretic approach to causal inference [Heckerman and Shachter, 1995, Dawid, 2012, 2020]. Lattimore and Rohde [2019a,b] also document an approach to causal modelling that demands explicit consideration of the set of interventions, and is arguably an example of the decision theoretic approach.

Lindley and Novick [1981] discussed sequences of exchangeable observations along with "one more observation". Lindley mentioned 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], Imbens and Rubin [2015] made use of the assumption of models with exchangeable potential outcomes to prove several identification results. Saarela et al. [2020], used graphical causal models to propose conditional exchangeability, defined as the exchangeability of the non-intervened causal parents of a target variable under intervention on its remaining 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. Hernán and Robins [2006], Hernán [2012], Greenland and Robins [1986], Banerjee et al. [2017], Dawid [2020] all discuss similar experimental symmetries. These symmetries are reminiscent of exchange commutativity discussed here. They're not identical, however – exchange commutativity can be justified by the equivalence of certain prediction problems that arise from a single experiment, instead of an equivalence of different experiments that arise from, for example, interchanging experimental subjects.

A different kind of regularity of causal models is given by the stable unit treatment distribution assumption (SUTDA) in Dawid [2020] and the stable unit treatment value assumption (SUTVA) in [Rubin, 2005]. This regularity is similar to the condition of locality introduced here.

Theorem 4.7 was influenced by the idea of causal inference by invariant prediction [Peters et al., 2016]. While both the assumptions and the conclusions drawn in that work differ from the assumptions and conclusion of Theorem 4.7, both proceed from an idea that can be roughly described as "things I can do have been done before" and both look for variable pairs $\mathsf{X}$ and $\mathsf{Y}$ such that the distribution of $\mathsf{Y}$ given $\mathsf{X}$ doesn't change after actions are taken. Finally, the variable described in that work as "the environment" is similar to the variable $\mathsf{Z}$ in Theorem 4.7. A key difference is that their result is proved for certain classes of structural interventions as the "things that can be done", while Theorem 4.7 does not depend on classes of structural interventions.

Finally, Guo et al. [2022] have recently generalised De Finetti's theorem to causal graphs in a different manner to the present work and analysed how causal structure may be inferred from independences in exchangeable models.

### 1.2 Outline

Section 2 outlines our mathematical framework and defines key terms. Section 2.1 introduces probability theory that may be familiar to many readers, while Section 2.2 introduces decision models, and extension of standard probability models and the basic kind of model we consider for the rest of this work. We also describe extended conditional independence, an extension of standard conditional independence suitable for decision models. Extended conditional independence is not original to this work, and was introduced by Constantinou and Dawid [2017].

Section 3.2 first explains decision models with conditionally independent and identical responses, an analogue of the common assumption of conditionally independent and identicallly distributed variables. We then introduce and explains Theorem 3.18, a "decision model analogue" for De Finetti's represention theorem, and finally argue, using Theorem 3.18, that the assumption of conditionally independent and identical responses is often unreasonable.

Section 4, in response to the need for a weaker assumption than conditionally independent and identical responses, introduces the idea of precedent via an example, and then introduces Theorem 4.7 which establishes that under some additional conditions the assumption of precedent can imply conditionally independent and identical responses. One of these additional requirements is a conditional independence, which is a common requirement to draw causal conclusions from observational data. The other condition is a stronger version of precedent known as diverse precedent. The interpretation of this assumption is not obvious. We postulate that it may be derivable from assumptions of causal direction along with the principle of independent causal mechanisms, but leave this as an open question.

## 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. Because decision models are stochastic functions rather than probability measures (Section 2.2), we make use a generalisation of conditional independence called extended conditional independence, explained in Section 2.3.

Section 2.4 defines some standard terms relating to directed acyclic graphs, which are likely familiar to anyone acquainted with structural causal models.

### 2.1 Probability Theory

#### 2.1.1 Measurable spaces

Definition 2.1 (Sigma algebra). Given a set $A$, a $\sigma$-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 2.2 (Measurable space). A measurable space $(A, \mathcal{A})$ is a set $A$ along with a $\sigma$-algebra $\mathcal{A}$.

Definition 2.3 (Sigma algebra generated by a set of events). Given a set $A$ and an arbitrary collection of subsets $U \supset \mathcal{P}(A)$, the $\sigma$-algebra generated by $U$, $\sigma(U)$, is the smallest $\sigma$-algebra containing $U$.

Common $\sigma$ algebras For any $A$, $\{\emptyset, A\}$ is a $\sigma$-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 $\sigma$-algebra.

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

#### 2.1.2 Probability measures and Markov kernels

Definition 2.4 (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 $\{A_i\} \subset \mathcal{E}$, $\mu(\cup_i A_i) = \sum_i \mu(A_i)$

Definition 2.5 (Set of all probability measures). The set of all probability measures on $(E, \mathcal{E})$ is written $\Delta(E)$. We equip $\Delta(E)$ with the coarsest $\sigma$-algebra such that the evaluation maps $\eta_B : \nu \mapsto \nu(B)$ are measurable for all $B \in \mathcal{F}$.

Definition 2.6 (Probability space). A probability space is a triple $(\mu, E, \mathcal{E})$ consisting of a probability measure and a measurable space.

Definition 2.7 (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 $x \in E$

Notation 2.8 (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 $\mathbb{M} : E \rightarrowtriangle F$, read "$\mathbb{M}$ maps from $E$ to probability measures on $F$".

Definition 2.9 (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 2.10 (Dirac measure). The Dirac measure $\delta_x \in \Delta(X)$ is a probability measure such that $\delta_x(A) = [\![x \in A]\!]$

Definition 2.11 (Markov kernel associated with a function). Given measurable $f : (X, \mathcal{X}) \to (Y, \mathcal{Y})$, $\mathbb{F}_f : X \rightarrowtriangle Y$ is the Markov kernel given by $x \mapsto \delta_{f(x)}$

Definition 2.12 (Markov kernel associated with a probability measure). Given $(X, \mathcal{X})$, a one-element measurable space $(\{*\}, \{\{*\}, \emptyset\})$ and a probability measure $\mu \in \Delta(X)$, the associated Markov kernel $\mathbb{Q}_\mu : \{*\} \rightarrowtriangle X$ is the unique Markov kernel $* \mapsto \mu$

### 2.1.3 Variables, conditionals and marginals

Definition 2.13 (Random variable). Given a measurable space $(\Omega, \mathcal{F})$, which we refer to as a sample space, and a measurable space of values $(X, \mathcal{X})$, an $X$-valued random variable on $\Omega$ is a measurable function $\mathsf{X} : (\Omega, \mathcal{F}) \to (X, \mathcal{X})$.

A sequence of random variables is also a random variable.

Definition 2.14 (Sequence of variables). Given a sample 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 $\mathsf{Y}$ is higher than $\mathsf{X}$ if $\mathsf{X}$ is given by application of a function to $\mathsf{Y}$. For example, $\mathsf{Y} \preccurlyeq (\mathsf{W}, \mathsf{Y})$ as $\mathsf{Y}$ can be obtained by composing a projection with $(\mathsf{W}, \mathsf{Y})$.

Definition 2.15 (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 $\mathsf{X} = f \circ \mathsf{Y}$.

We use superscripts to specify marginal and conditional distributions, as subscrips (which are a somewhat more common notation) are reserved for specifying options in decision models (Section 2.2).

Definition 2.16 (Marginal distribution). Given a probability space $(\mu, \Omega, \mathcal{F})$ and a variable $\mathsf{X} : \Omega \to (X, \mathcal{X})$, the marginal distribution of $\mathsf{X}$ with respect to $\mu$, $\mu^{\mathsf{X}} : \mathcal{X} \to [0, 1]$ by $\mu^{\mathsf{X}}(A) := \mu(\mathsf{X}^{-1}(A))$ for any $A \in \mathcal{X}$.

Definition 2.17 (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 $\mathsf{Y}$ given $\mathsf{X}$ is any Markov kernel $\mu^{\mathsf{Y}|\mathsf{X}} : X \rightarrowtriangle Y$ such that

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

Definition 2.18 (Trivial variable). We let $*$ stand for a single-valued variable $* : \Omega \to \{*\}$.

## 2.2 Decision models

A decision model is a Markov kernel $\mathbb{P}.$ from an option set $(C, \mathcal{C})$ to a sample space $(\Omega, \mathcal{F})$.

Definition 2.19 (Decision model). A decision model is a triple $(\mathbb{P}., (\Omega, \mathcal{F}), (C, \mathcal{C}))$ where $\mathbb{P}. : C \rightarrowtriangle \Omega$ is a Markov kernel, $(\Omega, \mathcal{F})$ is the sample space and $(C, \mathcal{C})$ is the option set.

For an option $\alpha \in C$, we say $\mathbb{P}_\alpha$ is the model $\mathbb{P}.$ evaluated at $\alpha$.

Definition 2.20 (Almost sure equality). Given a decision model $(\mathbb{P}., (\Omega, \mathcal{F}), (C, \mathcal{C}))$ and random variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, two Markov kernels $\mathbb{K} : X \rightarrowtriangle Y$ and $\mathbb{L} : X \rightarrowtriangle Y$ are $\mathbb{P}., \mathsf{X}, \mathsf{Y}$-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 $\mathbb{K} \overset{\mathbb{P}^{\mathsf{X}}_{\cdot}}{\cong} \mathbb{L}$.

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

### 2.3 Extended conditional independence

Because decision models aren't standard probability spaces, we need some version of conditional independence for decision models. Such a notion has already been worked out in some detail: it is the idea of extended conditional independence defined in Constantinou and Dawid [2017]. Extended conditional independence is substantially more general than we need for our purposes, and in fact we only consider two special cases of it. However, we still make use of the notational convention introduced in that paper.

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 exist, is equivalent to other common definitions Fritz [2020].

Definition 2.21 (Conditional independence). Given a decision model $(\mathbb{P}_{\cdot}, (\Omega, \mathcal{F}), (C, \mathcal{C}))$, variables $\mathsf{X}, \mathsf{Y}, \mathsf{Z}$ and fixing some $\alpha \in C$, we say $\mathsf{Y}$ is conditionally independent of $\mathsf{X}$ given $\mathsf{Z}$, written $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_\alpha} \mathsf{X}|\mathsf{Z}$, if there exists some $\mathbb{K} : Z \rightarrowtriangle Y$ such that

$$\mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}(A|x,z) \overset{\mathbb{P}^{\mathsf{XZ}}_{\alpha}}{\cong} \mathbb{K}(A|z) \qquad\qquad \forall A \in \mathcal{Y}$$

Extended conditional independence as introduced by Constantinou and Dawid [2017] is defined using pairs of "complementary nonstochastic variables" on the option set $C$.

Definition 2.22 (Nonstochastic variable). Given a decision model $(\mathbb{P}_{\cdot}, (\Omega, \mathcal{F}), (C, \mathcal{C}))$ and a measurable set $(X, \mathcal{X})$, a nonstochastic variable is a measurable function $\phi : C \to X$.

Definition 2.23 (Complementary nonstochastic variables). A pair of nonstochastic variables $\phi$ and $\xi$ are complementary if the pair $(\phi, \xi)$ is invertible.

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 2.24 (Extended conditional independence). Given a probability set $\mathbb{P}_C$, variables $\mathsf{X}$, $\mathsf{Y}$ and $\mathsf{Z}$ and complementary nonstochastic variables $\phi$ and $\xi$, the extended conditional independence $\mathsf{Y} \perp\!\!\!\perp^{e}_{\mathbb{P}_{\cdot}} \mathsf{X}\phi|\mathsf{Z}\xi$ holds if for each $a \in \xi(C)$, $\alpha, \alpha' \in \xi^{-1}(a)$,

$$\mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}_{\alpha} \overset{\mathbb{P}_{\cdot}}{\cong} \mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}_{\alpha'}$$

and for all $\alpha \in C$

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

In this work we only ever consider the complimentary pair $(\mathrm{id}_C, *)$ where $*$ is the trivial variable $\cdot \mapsto *.$, in which case extended conditional independence breaks down into two special cases: global conditional independence and uniform conditional independence. The former can be understood as "conditional independence for every $\alpha \in C$", while the latter means "conditional independence for every $\alpha \in C$, and moreover conditionally independent of $\mathrm{id}_C$".

Definition 2.25 (Global conditional independence). Given a decision model $(\mathbb{P}_{\cdot}, (\Omega, \mathcal{F}), (C, \mathcal{C}))$ and variables $\mathsf{X}$, $\mathsf{Y}$ and $\mathsf{Z}$, $\mathsf{Y}$ is globally independent of $\mathsf{X}$ given $\mathsf{Z}$, written $\mathsf{Y} \perp\!\!\!\perp^{e}_{\mathbb{P}_{\cdot}} \mathsf{X}|(\mathsf{Z}, \mathrm{id}_C)$ if for each $\alpha \in C$

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

Definition 2.26 (Uniform conditional independence). Given a decision model $(\mathbb{P}_{\cdot}, (\Omega, \mathcal{F}), (C, \mathcal{C}))$ and variables $\mathsf{X}$, $\mathsf{Y}$ and $\mathsf{Z}$, the uniform conditional independence $\mathsf{Y} \perp\!\!\!\perp^{e}_{\mathbb{P}_{\cdot}} (\mathsf{X}, \mathrm{id}_C)|\mathsf{Z}$ holds if $\mathsf{Y} \perp\!\!\!\perp^{e}_{\mathbb{P}_{\cdot}} \mathsf{X}|(\mathsf{Z}, \mathrm{id}_C)$ and furthermore for all $\alpha, \alpha' \in C$

$$\mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}_{\alpha} \overset{\mathbb{P}^{\mathsf{XZ}}_{\alpha}}{\cong} \mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}_{\alpha'}$$

For countable sets $C$ we can reason with collections of extended conditional independence statements as if they were regular conditional independence statements. In the following rules, $\phi$ and $\xi$ refer to complementary variables on the set $C$ (see Constantinou and Dawid [2017] for details), but for our purposes we only consider the cases where either $\phi = \text{id}_C$ and $\xi = *$ or $\phi = *$ and $\xi = \text{id}_C$. In the rest of this text, we will omit the trivial variable from extended conditional independence statements.

1. Symmetry: $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}.} (\mathsf{Y}, \phi)|(\mathsf{Z}, \xi)$ iff $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}.} (\mathsf{X}, \phi)|(\mathsf{Z}, \xi)$
2. $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}.} (\mathsf{Y}, \text{id}_C)|(\mathsf{Y}, \text{id}_C)$
3. Decomposition: $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}.} (\mathsf{Y}, \phi)|\mathsf{W}\xi$ and $\mathsf{Z} \preccurlyeq \mathsf{Y}$ implies $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}.} (\mathsf{Z}, \phi)|(\mathsf{W}, \xi)$
4. Weak union:
   (a) $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}.} (\mathsf{Y}, \phi)|(\mathsf{W}, \xi)$ and $\mathsf{Z} \preccurlyeq \mathsf{Y}$ implies $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}.} (\mathsf{Y}, \phi)|(\mathsf{Z}, \mathsf{W}, \xi)$
   (b) $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}.} \mathsf{Y}\text{id}_C|\mathsf{W}$ implies $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}.} \mathsf{Y}|(\mathsf{W}, \text{id}_C)$
5. Contraction: $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}.} (\mathsf{Z}, phi)|(\mathsf{W}, \xi)$ and $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}.} (\mathsf{Y}, \phi)|(\mathsf{Z}, \mathsf{W})\xi$ implies $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}.} (\mathsf{Y}, \mathsf{Z}, \phi)|(\mathsf{W}, \xi)$

If we have the extended conditional independence $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}.} \text{id}_C|\mathsf{X}$, then by definition for all $\alpha, \alpha' \in C$ we have $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_\alpha = \mathbb{P}^{\mathsf{Y}|\mathsf{X}}_{\alpha'}$. In this case, we use the notation $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_C$ to indicate that the conditional distribution does not depend on the choice of $\alpha$

Definition 2.27 (Uniform conditional distribution). Given a decision model $(\mathbb{P}., (\Omega, \mathcal{F}), (C, \mathcal{C}))$ and variables $\mathsf{X}$, $\mathsf{Y}$, if $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}.} \text{id}_C|\mathsf{X}$ then

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

for any $\alpha \in C$. If $\mathsf{Y} \not\perp\!\!\!\perp^e_{\mathbb{P}.} \text{id}_C|\mathsf{X}$ then $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_C$ is not defined.

### 2.4 Directed graphs

Definition 2.28 (Directed graph). A directed acyclic graph $\mathcal{G}$ is a set of nodes $\mathcal{V}$ and a set of edges $\mathcal{E}$. Each edge is an ordered pair of nodes $(V_i, V_j) \in \mathcal{V}^2$, with $V_i$ the source and $V_j$ the destination. An acyclic graph must have no directed path that begins and ends at $V_i$ for any $V_i \in \mathcal{V}$.

Definition 2.29 (Directed path). Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a directed path is a sequence of edges $((V^1_k, V^2_k))_{k \in [n]}$ from $\mathcal{E}$ such that for any $k \in [n]$, $V^2_k = V^1_{k+1}$. A directed path begins as $V^1_1$ and ends at $V^2_k$.

Definition 2.30 (Directed acyclic graph). A directed graph $\mathcal{G}$ is a directed acyclic graph if it contains no directed paths beginning and ending at the same node.

Definition 2.31 (Parents). Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the parents of a node $V_i$ are all the nodes $V_j$ such that there is an edge $(V_j, V_i) \in \mathcal{E}$: $\text{Pa}(V_i) = \{V_j | (V_j, V_i) \in \mathcal{E}\}$.

Definition 2.32 (Model graph association). Given a set of variables $(\mathsf{V}_i)_{i \in [k]}$, an associated directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a graph with a node $V_i$ for each variable $\mathsf{V}_i$. We define the parents of a variable via this association: $\text{Pa}_\mathcal{G}(\mathsf{V}_i) = \{\mathsf{V}_j | (V_j, V_i) \in \mathcal{E}\}$.

## 3 Inferring consequences when observations and consequences share responses

Returning to the example of body mass index and mortality raised in the introduction, we will present an implausible but relatively simple account of how a decision maker might leverage observations to draw conclusions about how their options that are known to affect body mass index can affect mortality. In particular, they may assume that whether or not an individual dies during the study period is related by a "fixed but unknown stochastic function" to their body mass index at the start of the study period – and this is true for individuals who were part of the observation set and individuals who may be influenced by the decision maker alike.

In more detail, suppose our decision maker has a decision model $(\mathbb{P}., (C, \mathcal{C}), (\Omega, \mathcal{F}))$ and a sequences of random variable pairs $(\mathsf{X}_i, \mathsf{Y}_i)_{i \in [m+n]}$ where $[m+n]$ is the set $\{1, 2, ..., m+n\}$ where $\mathsf{X}_i$ is an individual's body mass index and $\mathsf{Y}_i$ is a variable taking values in $\{0, 1\}$ indicating whether or not they died during the follow-up period. The first $m$ pairs in the sequence are observations unaffected by the decision maker and the next $n$ pairs are affected by their choice. The decision maker wants to learn something from the uncontrolled pairs of observations $(\mathsf{X}_{[m]},\mathsf{Y}_{[m]})$ to help make a decision that will promote good outcomes among the controlled pairs $(\mathsf{X}_{[m+1,n]},\mathsf{Y}_{[m+1,n]})$. In order to do this, the decision maker might assume:

- They already know how their choices determine the marginal distribution $\mathbb{P}_\alpha^{\mathsf{X}_i}$ for $i > m$
- There is an unknown response $\mathsf{H}$ taking values in $\Delta(Y)^X$ shared by all pairs $(\mathsf{X}_i,\mathsf{Y}_i)$, $i \in [m+n]$ that maps an individual's body mass index to their risk of death in the followup period; that is, $\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}} = \mathbb{P}_C^{\mathsf{Y}_j|\mathsf{D}_j\mathsf{H}} = \mathsf{H}$ for all $i,j \in [m+n]$
- For all $i$, whether or not an individual dies $\mathsf{Y}_i$ is independent of $(\mathsf{X}_j,\mathsf{Y}_j)_{j\neq i}$ conditional on $\mathsf{X}_i$ and $\mathsf{H}$; for all $i$, $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_\cdot} (\mathrm{id}_C, \mathsf{X}_{[m+n]\setminus\{i\}}, \mathsf{Y}_{[m+n]\setminus\{i\}})|(\mathsf{X}_i,\mathsf{H})$

In this case, the decision maker can use the relationship observed in the first $m$ pairs of observations $(x_{[m]},y_{[m]})$ to compute a posterior distribution $\mathbb{P}_C^{\mathsf{H}|\mathsf{D}_{[m]}\mathsf{Y}_{[m]}}$ and thereby estimate the effects of their options on $\mathsf{Y}_i$, $i>m$

$$\begin{aligned}\mathbb{P}_\alpha^{\mathsf{Y}_i}(A) &= \int_{\Delta(Y)^X}\int_X \mathbb{P}_C^{\mathsf{Y}|\mathsf{X}\mathsf{H}}(A|x,h)\mathbb{P}_\alpha^{\mathsf{X}_i}(\mathrm{d}x)\mathbb{P}_C^{\mathsf{H}|\mathsf{X}_{[m]}\mathsf{Y}_{[m]}}(\mathrm{d}h|x_{[m]},y_{[m]})\\ &= \int_{\Delta(Y)^X}\int_X h(A|x)\mathbb{P}_\alpha^{\mathsf{X}_i}(\mathrm{d}x)\mathbb{P}_C^{\mathsf{H}|\mathsf{X}_{[m]}\mathsf{Y}_{[m]}}(\mathrm{d}h|x_{[m]},y_{[m]})\end{aligned}$$

Once again, the key assumption enabling this deduction is that the same response $\mathsf{H}$ is shared by both the observations $(\mathsf{X}_i,\mathsf{Y}_i)$, $i\in[m]$ and the consequences $(\mathsf{X}_j,\mathsf{Y}_j)$, $j>m$. Why might we buy this assumption? One reason is that the system we're modelling is like an engineered system. In such systems, significant effort is often invested to ensure that components respond to the system's state in reliable and predictable ways, and as a result we might consider it likely that components remain predictable if we act on the system's state. The relationship between body mass index and mortality does not arise in such a system, and so this justification is unavailable – but this does not imply that we need to reject the assumption either.

While engineered systems might be designed just so that their components exhibit repeatable responses, we might suppost that most systems exhibit regular response relationships if we knew exactly where to look for these relationships. We frequently observe predictable behaviour in non-engineered systems from the weather to people's online shopping. This predictable behaviour is a result, perhaps, of observing systems that respond to inputs in regular ways, primed with regular inputs. This might be so, but it tells us little about whether particular pairs of variables share identical probabilistic responses. We will discuss this intuition further in Section 4.

In this section we explore an alternative way to view this assumption. We draw inspiration from De Finetti's work that furnished us with an alternative view of the assumption that a sequence of variables is independent and identically distributed with an unknown distribution de Finetti [[1937] 1992]. De Finetti showed how this assumption is equivalent to the assumption that someone judges that their predictive model for this sequence should not be changed if the sequence is rearranged, an assumption known as exchangeability. Exchangeability is inappropriate for decision models, because the fundamental premise of using formal models to assist decision making is that the decision maker's choices lead to differences in the distribuitons of some variables, so swapping these controlled variables with other variables should lead to a change in the model we used to assess consequences.

We explore a generalisation of De Finetti's equivalence more appropriate for decision making models. Instead of examining sequences of independent and identically distributed (IID) variables, we examine sequences of variable pairs that share independent and identical responses of the kind we have already discussed. Where IID sequences are associated models that are unchanged by arbitrary variable swaps, we characterise sequences of pairs with independent and identical responses as being unchanged by swaps of infinite subsequences of pairs. An informal statment of this result is that an analyst accepting an assumption of independent and identical responses is tantamount to announcing that they are sure their results would be unchanged whether their data was derived from an experiment or passive observation.

In the example presented here – concerning the relationship between body mass index and mortality – we consider such an assumption to be unreasonable. In fact, we argue that this result suggests that an assumption of independent and identical responses is untenable in many situations. If, for example, an analyst believes that there would be any benefit in checking their results against experimental data, then our result indicates that this attitude amounts to a rejection of the assumption of independent and idential responses. While we argue on this basis that the assumption of independent and identical responses is often untenable, we argue in Section 4 that a special case of this assumption we refer to as precedent is more plausible.

### 3.1 Conditionally independent and identical responses

We formalise decision models with "shared but unknown responses" as sequential models of variable pairs with conditionally independent and identical responses (CIIRs). A sequence of pairs $(\mathsf{D}_i, \mathsf{Y}_i)_{i\in\mathbb{N}}$ share conditionally independent and identical responses if there is an unknown stochastic function $\mathsf{H}$ taking values in $\mathit{\Delta}(Y)^D$ – in the set of maps from $D$ to probability distributions over $Y$ – such that every output $\mathsf{Y}_i$ "responds to" $\mathsf{D}_i$ according to the same $\mathsf{H}$. While our example featured a decision maker who has prior knowledge about how to control some of the inputs $\mathsf{D}_i$, this is a separate assumption and is not required by the assumption of CIIR pairs.

We define sequential input-output models as a shorthand for a decision model along with a sequence of variable pairs.

Definition 3.1 (Sequential input-output model). A decision model $(\mathbb{P}_\cdot, (C, \mathcal{C}), (\mathit{\Omega}, \mathcal{F}))$ and two sequences of variables $\mathsf{Y} := (\mathsf{Y}_i)_{i\in\mathbb{N}}$ and $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ is a sequential input-output model, which we specify with the shorthand $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$.

The formal definition of CIIRs follows. In this work we consider the response $\mathsf{H}$ to be a random variable.

Definition 3.2 (Conditionally independent and identical responses). Given a sequential input-output model $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$, the $(\mathsf{D}_i, \mathsf{Y}_i)$ pairs are related by independent and identical responses conditional on $\mathsf{H}$ if for all $i$, $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{D}_{[1,i)}, \mathsf{Y}_{[1,i)})|(\mathsf{D}_i, \mathsf{H}, \mathrm{id}_C)$ and $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}} \overset{\mathbb{P}_\alpha^{\mathsf{D}_i|\mathsf{H}}}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}_j|\mathsf{D}_j\mathsf{H}}$ for all $i, j$.

Definition 3.2 asserts that there are versions of all the conditional distributions $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}}$ that are pairwise congruent, and Theorem 3.3 shows that this is sufficient for the existence of a single conditional distribution that is a version of $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}}$ for all $i$.

Theorem 3.3 (Existence of representative conditional distribution). Given a sequential input-output model $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$, if the $(\mathsf{D}_i, \mathsf{Y}_i)$ pairs are related by independent and identical responses conditional on $\mathsf{H}$, then for every $\alpha$, $\mathbb{P}_\alpha$-almost all $h \in H$ there is a representative conditional distribution $h_X^Y$ such that $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{H}}(\cdot|h) \overset{\mathbb{P}_\alpha^{\mathsf{D}_i|\mathsf{H}}(\cdot|h)}{\cong} h_X^Y$ for all $i$.

We refer to the function $\mathsf{H}_X^Y : h \mapsto h_X^Y$ as a representative conditional distribution.

Proof. Fix $h$ and take $h_{X,i}^Y := \mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{H}}(\cdot|\cdot, h)$ to be an arbitrary version of the conditional distribution for all $i$.

For $i, j \in \mathbb{N}$, take $S_{ij} := \{x|h_{x,i}^Y$ is not a version of $\mathbb{P}_\alpha^{\mathsf{Y}_j|\mathsf{X}_j\mathsf{H}}(\cdot|\cdot, h)\}$. Note that $S_i := \cup_{j\in\mathbb{N}} S_{ij}$ is a countable union of sets of $\mathbb{P}_\alpha^{\mathsf{X}_i|\mathsf{H}}(\cdot|h)$-measure 0, hence is also a set of $\mathbb{P}_\alpha^{\mathsf{X}_i|\mathsf{H}}(\cdot|h)$-measure 0.

Define

$$h_X^Y(A|x) := \sum_{i\in N} \mathbb{1}_{S_i^\complement \setminus \cup_{j\in[i]} S_j^\complement}(x) h_{X,i}^Y(A|x)$$

By construction, $h_X^Y$ differs from each $h_{X,i}^Y$ by a measure 0 set with respect to $\mathbb{P}_\alpha^{\mathsf{X}_i|\mathsf{H}}(\cdot|h)$. Hence it is a version of $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{H}}(\cdot|h)$ for every $i$. □

In general, the definition for conditionally independent and identical responses only requires that the outputs $\mathsf{Y}_i$ are independent of previous inputs and outputs conditional on $\mathsf{H}$ and $\mathsf{D}_i$. If $\mathsf{D}_i$ is selected based on previous data, then in general there may be relationships between $\mathsf{D}_j$ and $\mathsf{Y}_i$ for $j > i$ even after conditioning on $\mathsf{D}_i$ and $\mathsf{H}$. However, for present purposes we make the additional simplifying assumption that inputs are weakly data-independent, which means that conditional on $\mathsf{H}$ and past inputs $\mathsf{D}_{[1,i]}$, $\mathsf{Y}_i$ is also independent of all future inputs. Generalising our theory to data-dependent inputs is an open question.

Definition 3.4 (Weakly data-independent). A sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ with independent and identical responses conditional on $\mathsf{H}$ is weakly data-independent if $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{D}_{\mathbb{N}\setminus\{i\}}|(\mathsf{D}_i, \mathsf{H}, \mathrm{id}_C)$.

### 3.2 Symmetries of sequential conditional probabilities

Given the previously mentioned sequences $\mathsf{D}$ and $\mathsf{Y}$, the decision maker has for each option $\alpha \in C$ a conditional probability $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}}$. An obvious symmetry of this conditional probability we could consider is symmetry to paired permutations of $\mathsf{D}$ and $\mathsf{Y}$. That is, given any permutation $\rho : \mathbb{N} \to \mathbb{N}$, define $\mathsf{Y}_\rho := (\mathsf{Y}_{\rho(i)})_{i\in\mathbb{N}}$ and $\mathsf{D}_\rho$ similarly. Then symmetry to paired permutations means for all $\alpha$, $\rho$

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

This symmetry is reminiscent of exchangeability, and in Theorem B.9 we show that it implies that the $(\mathsf{D}_i, \mathsf{Y}_i)$ share conditionally independent and identical responses. However, the converse is not true, as Example 3.5 shows.

Example 3.5. Suppose there is a machine with two arms $D = \{0, 1\}$, one of which always pays out \$100 50% of the time and nothing otherwise, and the other that pays out nothing. A decision maker (DM) doesn't know which is which, but DM watches a sequence of people operate the machine. The first person in the sequence was told yesterday exactly which arm is good, and most likely remembers. The second one has no idea which arm is good, and does not observe the first person's choice. The DM is sure that they all want the money, so the first person will pull the good arm $1 - \epsilon$ of the time, while the second person will pull the good arm 50% of the time. The response $\mathsf{H}$ takes values that can be summarised as "0 is good" and "1 is good" (which we'll just refer to as $\{0, 1\}$), and the DM assigns 50% probability to each initially. Then

$$\begin{aligned}\mathbb{P}_C^{\mathsf{Y}_2|\mathsf{D}_2}(100|1) &= \mathbb{P}_C^{\mathsf{Y}_2|\mathsf{D}_2\mathsf{H}}(100|1,0)\mathbb{P}_C^{\mathsf{H}|\mathsf{D}_2}(0|1) + \mathbb{P}_C^{\mathsf{Y}_2|\mathsf{D}_2\mathsf{H}}(100|1,1)\mathbb{P}_C^{\mathsf{H}|\mathsf{D}_2}(1|1)\\ &= 0 \cdot 0.5 + 0.5 \cdot 0.5\\ &= 0.25\end{aligned}$$

while

$$\begin{aligned}\mathbb{P}_C^{\mathsf{Y}_1|\mathsf{D}_1}(100|1) &= \mathbb{P}_C^{\mathsf{Y}_1|\mathsf{D}_1\mathsf{H}}(100|1,0)\mathbb{P}_C^{\mathsf{H}|\mathsf{D}_1}(0|1) + \mathbb{P}_C^{\mathsf{Y}_1|\mathsf{D}_1\mathsf{H}}(100|1,1)\mathbb{P}_C^{\mathsf{H}|\mathsf{D}_1}(1|1)\\ &= 0 \cdot \epsilon + 0.5(1-\epsilon)\\ &= 0.5(1-\epsilon)\\ &\neq \mathbb{P}_C^{\mathsf{Y}_2|\mathsf{D}_2}(100|1)\end{aligned}$$

Example 3.5 motivates the weaker symmetry we call exchange commutativity. The key difference is that exchange commutativity allows for the permutability of pairs after conditioning on some arbitrary variable $\mathsf{W}$. A sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ is exchange commutative if there is some variable $\mathsf{W}$ such that the conditional $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ is symmetric to paired swaps of $\mathsf{Y}$ and $\mathsf{D}$.

Definition 3.6 (Exchange commutativity). Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ along with some $\mathsf{W} : \Omega \to W$, we say $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ commutes with exchange over $\mathsf{W}$ if for all finite permutations $\rho : \mathbb{N} \to \mathbb{N}$ and all $\alpha \in C$

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

We say $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ commutes with exchange if there is some $\mathsf{W}$ such that $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ commutes with exchange over $\mathsf{W}$.

A second regularity condition we will consider can be roughly understood as the idea that $\mathsf{Y}_i$ doesn't "depend on" $\mathsf{D}_j$ for $j \neq i$. As Example 3.5 suggests, this cannot be an assumption that $\mathsf{Y}_i$ doesn't depend on $\mathsf{D}_j$ unconditionally; $\mathsf{D}_j$ could, after all, offer some evidence about the state of the shared response $\mathsf{H}$. Instead, we assume that $\mathsf{Y}_i$ doesn't depend on non-corresponding $\mathsf{X}_j$ after conditioning on some auxiliary $\mathsf{W}$.

Definition 3.7 (Locality). Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ along with some $\mathsf{W} : \Omega \to W$, the model is local over $\mathsf{W}$ if for all $\alpha \in C$, $n \in \mathbb{N}$, $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{X}_{\{i,\infty)}|(\mathsf{W}, \mathsf{X}_i, \mathrm{id}_C)$. If there is some $\mathsf{W}$ such that $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ is local over $\mathsf{W}$ then we say $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ is local.

If an input-output model is both exchange commutative and local, then we say it is input-output contractible. This term is chosen because such a model is unchanged by contractions of the input and output indices - see Theorem 3.9.

Definition 3.8 (Input-output contractibility). A sequential input-output model $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ along with some $\mathsf{W} : \Omega \to W$ is input-output contractible (IO contractible) over $\mathsf{W}$ if it is local and commutes with exchange.

Theorem 3.9 (Equality of equally sized subsequence conditionals). Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ and some $\mathsf{W}$, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ is IO contractible over $\mathsf{W}$ if and only if for all subsequences $A, B \subset \mathbb{N}$ with $|A| = |B|$ and for every $\alpha$

$$\begin{aligned}\mathbb{P}_\alpha^{\mathsf{Y}_A|\mathsf{WD}_{A,\mathbb{N}\setminus A}} &= \mathbb{P}_\alpha^{\mathsf{Y}_B|\mathsf{WD}_{B,\mathbb{N}\setminus B}} \\ &= \mathbb{P}_\alpha^{\mathsf{Y}_A|\mathsf{WD}_A} \otimes \mathrm{del}_{D^{|\mathbb{N}\setminus A|}}\end{aligned}$$

Proof. Appendix B.1 □

Appendix B.2 sets out two additional properties of these symmetries. Example B.4 shows that neither locality nor exchange commutativity is implied by the other, and Example B.5 shows that locality by itself does not rule out everything that we might intuitively describe as "interference" between pairs.

### 3.3 Representation of IO contractible models

In this section, we state Theorem 3.18, which shows that a sequential input output model $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ features pairs $(\mathsf{D}_i, \mathsf{Y}_i)$ related by conditionally independent and identical responses if and only if it is IO contractible over some variable $\mathsf{W}$.

The proof of the theorem is involved, and can be found in its entirety Appendix B.3. Note that we employ a string diagram notation in some steps of the proof, explained in Appendix A. Here we just introduce enough to explain the key terms in the theorem statement.

### 3.4 Preliminaries

Definition 3.10 (Input count variable). Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ with countable $D$, $\#_j^k$ is the variable

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

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

If we have an infinite sequence of pairs $(\mathsf{D}_i, \mathsf{Y}_i)$, we can wrap the sequence $\mathsf{Y}$ into a table $\mathsf{Y}^D$ such that $\mathsf{Y}_{11}^D$ is equal to the value of the first $\mathsf{Y}_i$ such that $\mathsf{D}_i = 1$, $\mathsf{Y}_{21}^D$ is equal to the value of the second such $\mathsf{Y}_i$ and so forth. We call it a "tabulated conditional" because, under the assumption of CIIRs, we can evaluate a conditional $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}}(\cdot|d_1, d_2, ...)$ by "looking up" the marginal distribution $\mathbb{P}_\alpha^{\mathsf{Y}_{1d_1}^D \mathsf{Y}_{2d_2}^D \cdots}$ over the appropriate elements of $\mathsf{Y}^D$.

Definition 3.11 (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 $\mathsf{Y}^D : \Omega \to Y^{\mathbb{N}\times D}$ by

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

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

The directing random measure of a sequence of exchangeable variables is defined as the map from the set of events of each variable in the sequence the limit of normalised partial sums of indicator functions over that set [Kallenberg, 2005]. The directing random measure is a probability measure. For completeness, we also define a directing random measure in the case that the relevant limit does not exist, although we are only practically interested in using the definition where the limit does exist.

Definition 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 $\mathsf{X}$ written $\mathsf{H} : \Omega \to \Delta(X)$ 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 input and output sequences $\mathsf{D}$ and $\mathsf{Y}$ we define the directing random conditional as the directing random measure of the tabulated conditional $\mathsf{Y}^D$ interpreted as a sequence of column vectors $((\mathsf{Y}^D_{1j})_{j\in D}, (\mathsf{Y}^D_{2j})_{j\in D}, ...)$.

Definition 3.13 (Directing random conditional). Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$, we will say the directing random conditional $\mathsf{H} : \Omega \to \Delta(Y^D)$ is the function

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

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 3.14 (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 3.16 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 we do not assume that the tabulated conditional $\mathsf{Y}^D$ is interpretable as potential outcomes. By representing a conditional probability as an exchangeable regular probability distribution, we can apply De Finetti's theorem, which is a key step in proving the main result of Theorem 3.18.

To prove Lemma 3.16, 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, we would have to accept positive probability that we run out of $\mathsf{D}_i$s taking some value $j \in D$ preventing us from filling out the "tabulated conditional" $\mathsf{Y}^D$ correctly. We call this side condition infinite support.

Definition 3.15 (Infinite support). Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ with $D$ countable if, letting $E \subset D^{\mathbb{N}}$ be the set of all sequences such that for all $j \in D$

$$x \in E \implies \sum_{i=0} [\![x_i = j]\!] = \infty$$

we have $\mathbb{P}_\alpha^{\mathsf{D}|\mathsf{W}}(E|w) = 1$ for all $\alpha, w$, then we say $\mathsf{D}$ is infinitely supported over $\mathsf{W}$.

The key property of the tabulated conditional is that we can evaluate the regular conditional $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ by "looking up" the appropriate marginal of $\mathbb{P}_\alpha^{\mathsf{Y}^D}$.

Lemma 3.16. Suppose a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ is given with $D$ countable and $\mathsf{D}$ infinitely supported over $\mathsf{W}$. Then for some $\mathsf{W}$, $\alpha$, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ is IO contractible if and only if

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

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

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

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 $\mathsf{Y}^D$ selected according to $d \in D^{\mathbb{N}}$.

If: We construct a conditional probability according to Definition 3.11 and verify that it satisfies IO contractibility.

The full proof can be found in Appendix B.3. Note that the proof uses string diagram notation explained in Appendix A. □

Because the distribution $\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}$ from Lemma 3.16 is row-exchangeable, the limit in the definition of the directing random conditional $\mathsf{H}$ exists almost surely (see Lemma B.6). In fact, we do not need the full sequence of pairs $(\mathsf{D}, \mathsf{Y})$ to calculate $\mathsf{H}$; any subsequence $A \subset \mathbb{N}$ that satisfies the condition that $\mathsf{D}_A$ is infinitely supported over $\mathsf{W}$ is sufficient.

Theorem 3.17. Suppose a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ is given with $D$ countable, $\mathsf{D}$ infinitely supported over $\mathsf{W}$ and for some $\mathsf{W}$, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ is IO contractible for all $\alpha$. Consider an infinite set $A \subset \mathbb{N}$, and let $\mathsf{D}_A := (\mathsf{D}_i)_{i\in A}$ and $\mathsf{Y}_A := (\mathsf{Y}_i)_{i\in A}$ such that $\mathsf{D}_A$ is also infinitely supported over $\mathsf{W}$. Then $\mathsf{H}_A$, the directing random conditional of $(\mathbb{P}_C, \mathsf{D}_A, \mathsf{Y}_A)$ is almost surely equal to $\mathsf{H}$, the directing random conditional of $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$.

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

The proof is in Appendix B.3. □

We are now ready to state the main result, Theorem 3.18. Assuming a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ (Definition 3.1) with inputs $\mathsf{D}$ infinitely supported (Definition 3.15) over some random variable $\mathsf{W}$, $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ is IO contractible over the same $\mathsf{W}$ if and only if the pairs $(\mathsf{D}_i, \mathsf{Y}_i)$ share conditionally independent and identical responses (Definition 3.2), given by the directing random conditional $\mathsf{H}$ (Definition 3.13) and $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ is weakly data-independent.

### 3.5 Statement of the representation theorem

Theorem 3.18 (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 $\mathsf{D}$ infinitely supported over $\mathsf{W}$. Then the following are equivalent:

1. There is some $\mathsf{W}$ such that $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ is IO contractible for all $\alpha$
2. For all $i$, $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{\neq i}, \mathsf{D}_{\neq i}, \mathrm{id}_C)|(\mathsf{H}, \mathsf{D}_i)$ and for all $i, j, \alpha$
$$\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{H}\mathsf{D}_i} = \mathbb{P}_\alpha^{\mathsf{Y}_j|\mathsf{H}\mathsf{D}_j}$$
3. There is some $\mathbb{L} : H \times X \rightarrowtriangle Y$ such that for all $\alpha$,
$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{DH}}(\bigtimes_{i\in\mathbb{N}} A_i|d, h) = \prod_{i\in\mathbb{N}} \mathbb{P}_C^{\mathsf{Y}_1|\mathsf{D}_1\mathsf{H}}(A_i|d_i, h)$$

Proof. (1) $\implies$ (3): We apply Lemma 3.16 followed by Lemma B.6 followed by Lemma B.7.

(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 $\mathsf{W}$ for all $\alpha$.

See Appendix B.4 for the full proof. Note that the proof uses string diagram notation explained in Appendix A. □

Whenever we have an input-output model with conditionally independent and identical responses given some arbitrary $\mathsf{W}$, then we also have conditionally independent and identical responses given the directing random conditional $\mathsf{H}$.

Corollary 3.19. If a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ has independent and identical responses conditional on some variable $\mathsf{W}$ and $\mathsf{D}$ has infinite support over the same $\mathsf{W}$, then letting $\mathsf{H}$ be the directing random conditional with respect to inputs $\mathsf{D}$ and outputs $\mathsf{Y}$, it follows that for for all $i$, $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{W}|(\mathsf{D}_i, \mathsf{H}, \mathrm{id}_C)$ and for all $\alpha, i, j$, $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}} = \mathbb{P}_\alpha^{\mathsf{Y}_j|\mathsf{D}_j\mathsf{H}}$.

Proof. We have by Theorem 3.18 that $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ is IO contractible over $\mathsf{W}$. The conclusion follows by applying Theorem 3.18 a second time. □

Building on Corollary 3.19, Theorem 3.20 shows the assumption that the pairs $(\mathsf{D}_i, \mathsf{Y}_i)$ are related by conditionally independent and identical responses implies that, for the purposes of learning the response function $\mathsf{H}$, all infinite subsequences of $(\mathsf{D}_i, \mathsf{Y}_i)$ pairs with appropriate support are interchangeable. That is, suppose we have some infinite $A \subset \mathbb{N}$ for such that $(\mathbb{P}_., \mathsf{D}_A, \mathsf{Y}_A)$ is unimpeachably IO contractible over $*$ −

perhaps all pairs indexed by $A$ are derived from a carefully conducted experiment in precisely the conditions of interest to the decision maker and are therefore considered interchangeable in this strong sense. If we have some other infinite set $B \subset \mathbb{N} \setminus A$ of pairs derived from passive observation, then the assumption of conditionally independent and identical responses for the whole collection of pairs $(\mathsf{D}_i, \mathsf{Y}_i)_{i\in\mathbb{N}}$ implies that while we may not be able to swap individual pairs in $A$ with individual pairs in $B$, we must be able to swap the whole set $A$ for the whole set $B$ for the purposes of learning the response function $\mathsf{H}$.

Theorem 3.20. A data-independent sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ with directing random conditional $\mathsf{H}$ and $\mathsf{D}$ infinitely supported over $\mathsf{H}$ features conditionally independent and identical response functions $\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}}$ only if for any sets $A, B \subset \mathbb{N}$ such that $\mathsf{D}_A$ and $\mathsf{D}_B$ are also infinitely supported over $\mathsf{H}$ and any $i, j \in \mathbb{N}$ such that $i \notin A$, $j \notin B$,

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

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

Proof. See Appendix B.5. □

### 3.6 Does IO contractibility help us understand identification?

One of the key contributions of De Finetti's representation theorem was to provide an alternative justification for the common modelling assumption that a sequence of variables were all distributed according to a shared but unknown "true distribution". De Finetti regarded the notion of an "unknown true distribution" as nonsensical, and through his representation theorem suggested that we could instead justify this structure by arguing that the experiment that produced the sequence of variables was, from the point of view of the analyst seeking to make predictions, invariant to reindexing the variables in the sequence.

Can IO contractibility help to justify common causal assumptions in a similar way? This question is less straightforward because IO contractibility is not such a straightforward symmetry. However, we think it does offer some insight into a common kind of causal assumption. Rather than lending justification to this assumption, the we think that it strengthens the case that this assumption is usually unreasonable.

The particular assumption we have in mind is, in the world of causal graphical models, the assumption that backdoor adjustment is possible and in the world of potential outcomes it is the assumption of conditional ignorability [Rubin, 2005]. Both assumptions hold that, given a treatment $\mathsf{D}_i$, covariates $\mathsf{X}_i$ and an outcome $\mathsf{Y}_i$, there is an unknown but common conditional distribution of $\mathsf{Y}_i$ given $\mathsf{D}_i$ and $\mathsf{X}_i$ for all $i$, where $i$ ranges over passive observations as well as the consequences of actions. That is, we assume that the pairs $((\mathsf{D}_i, \mathsf{X}_i), \mathsf{Y}_i)$ share conditionally independent and identical responses. The key implication is Theorem 3.20, which holds that, if the sequences of observations and consequences are both infinite, then for the purpose of learning the response function the problem is unchanged by swapping any subset of the indices corresponding to observations with any subset of those corresponding to consequences. That is, there is no difference between predicting the response function of the passive observations from an infinite sequence of passive observational data and predicting the response function of the consequences of the decision makers actions from the same sequence of passive observational data.

In practice, we propose that it would be very rare to have both of these datasets and treat them as interchangeable in this manner. Example 3.21 makes a similar point.

Example 3.21. Suppose an experiment is done which assigns some medical treatment $\mathsf{D}_i$ uniformly according to some random signal to patients for even $i$, and allows assignment by patient and doctor discretion for odd $i$. $\mathsf{Y}_i$ is a binary variable recording some health outcome of interest and $\mathsf{X}_i$ is some vector of covariates. The sequence $(\mathsf{D}_c, \mathsf{X}_c, \mathsf{Y}_c)$ is associated with the consequences of a decision maker's choices, where $c$ is some special character not in $\mathbb{N}$.

According to Theorem 3.20, the assumption of conditionally independent and identical responses applied to $((\mathsf{D}, \mathsf{X}), \mathsf{Y})$ implies

$$\begin{aligned}
\mathbb{P}_\alpha^{\mathsf{Y}_c|\mathsf{D}_c\mathsf{X}_c\mathsf{D}_{\text{odds}}\mathsf{X}_{\text{odds}}\mathsf{Y}_{\text{odds}}} &= \mathbb{P}_\alpha^{\mathsf{Y}_c|\mathsf{D}_c\mathsf{X}_c\mathsf{D}_{\text{evens}\setminus\{0\}}\mathsf{X}_{\text{evens}\setminus\{0\}}\mathsf{Y}_{\text{evens}\setminus\{0\}}} \\
&= \mathbb{P}_\alpha^{\mathsf{Y}_2|\mathsf{D}_2\mathsf{X}_2\mathsf{D}_{\text{evens}}\mathsf{X}_{\text{evens}\setminus\{2\}}\mathsf{Y}_{\text{evens}\setminus\{2\}}} \\
&= \mathbb{P}_\alpha^{\mathsf{Y}_2|\mathsf{D}_2\mathsf{X}_2\mathsf{D}_{\text{odds}}\mathsf{X}_{\text{odds}}\mathsf{Y}_{\text{odds}}}
\end{aligned}$$

That is, under this assumption, the following four problems are deemed identical:

- 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
- Predicting a held-out experimental outcome from the experimental data
- Predicting a held-out experimental outcome from the observational data

Any answer to one problem is, under this assumption, an answer for all of them. This is an assumption; we do not conclude this by comparing answers to these different problems and finding them to be the same, we simply assume it is so. The proposition that these problems are identical is hard to swallow: it seems very unlikely, for example, if an analyst aiming to predict experimental results with access to the experimental data would be satisfied with their previous answer derived from the observational data.

In practice, when both experimental and observational data are available, they are not assumed to be interchangeable in this sense – in fact, the question of how well the observational data predicts experimental outputs is one of substantial interest Eckles and Bakshy [2021], Gordon et al. [2018, 2022].

## 4 Inferring consequences when options have precedent

We have suggested that conditionally independent and identical responses is usually an unreasonably strong assumption for a decision maker to make, on the grounds that it implies overly strong interchangeability properties between different datasets. One way to get around this objection is to suppose that conditionally independent and identical responses are shared by pairs $(\mathsf{E}_i, \mathsf{X}_i)$ where the $\mathsf{E}_i$ are in fact latent variables. In this case, the assumption would still assert that infinite $(\mathsf{E}_i, \mathsf{X}_i)$ sequences arising from observation would be interchangeable with infinite $(\mathsf{E}_j, \mathsf{X}_j)$ sequences arising as consequences of actions, but because the $\mathsf{E}_i$ are never observed these interchanges do not imply that we would use the same model for different experiments.

To simplify the presentation, we will consider a specific kind of decision model featuring long sequence of exchangeable observations indexed by natural numbers that are unresponsive to the decision maker's choice and "one more" variable representing the "consequences of action" indexed by the special character $c$ that may be responsible ot the decision maker's choice. That is, we have $(\mathsf{X}_i)_{i\in\mathbb{N}}$ unresponsive to the decision maker and $(\mathsf{X}_c)$ responsive to the decision maker. Call this setup a "see-do model".

Definition 4.1 (See-do model). A see-do model is an decision model $(\mathbb{P}_\cdot, \Omega, C)$ along with a sequence of variables $\mathsf{X}_{\mathbb{N}\cup\{c\}}$ where $\mathsf{X}_{\mathbb{N}} \perp\!\!\!\perp^{e}_{\mathbb{P}_\cdot} \mathrm{id}_C$. Variables indexed with $i \in \mathbb{N}$ are referred to as observations and variables indexed with the special index $c$ are referred to as consequences. We specify a see-do model with the shorthand $(\mathbb{P}_\cdot, \mathsf{X}_{\mathbb{N}\cup\{c\}})$.

In this section, we will consider the following kind of "standard" see-do model: we have some observed variables $(\mathsf{X}, \mathsf{Y}, \mathsf{Z})$ and an unobserved variable $\mathsf{E}$ such that the observation pairs $(\mathsf{Z}_i, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i))_{i\in\mathbb{N}}$ share conditionally independent and identical responses. Typically, this might be because we assume observations are exchangeable, but we also allow for cases where $\mathsf{Z}_i$ is not exchangeable – for example, perhaps it is a time variable which monotonically increases. We also assume that the pairs $(\mathsf{E}_i, (\mathsf{X}_i, \mathsf{Y}_i))_{i\in\mathbb{N}\cup\{c\}}$ share conditionally independent and identical responses for all indices.

Recall that in Section 3 we suggested that many systems might exhibit (probabilistically) regular input-output behaviours, but where we might not know or observe the right "inputs". The assumption that the pairs $(\mathsf{E}_i, (\mathsf{X}_i, \mathsf{Y}_i))_{i\in\mathbb{N}\cup\{c\}}$ share conditionally indpendent and identical responses can be viewed as a formalisation of this intuition; there is some unknown and unobserved state $\mathsf{E}_i$ which $\mathsf{X}_i$ and $\mathsf{Y}_i$ respond to in a regular manner no matter what else is happening.

Note that we make no assumptions about the distribution of $\mathsf{Z}_c$.

Definition 4.2 (Latent CIIR see-do model). A latent CIIR see-do model is a see-do model $(\mathbb{P}_\cdot, (\mathsf{E}_i, \mathsf{Z}_i, \mathsf{X}_i, \mathsf{Y}_i)_{i\in\mathbb{N}\cup\{c\}})$ such that the observation pairs $(\mathsf{Z}_i, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i))_{i\in\mathbb{N}}$ share conditionally independent and identical responses and the pairs $(\mathsf{E}_i, (\mathsf{X}_i, \mathsf{Y}_i))_{i\in\mathbb{N}\cup\{c\}}$ also share conditionally independent and identical responses. We say the $\mathsf{E}_i$s are "latent" variables, which informally means that we typically do not get to observe them. We adopt the convention that the directing random conditional of $(\mathbb{P}_\cdot, \mathsf{Z}_{\mathbb{N}}, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i)_{i\in\mathbb{N}})$.

We can take any see-do model $(\mathbb{P}_\cdot, \mathsf{X}_{\mathbb{N}\cup\{c\}})$ with exchangeable observations and turn it into a latent CIIR see-do model by setting $\mathsf{Z}_i = *$ and $\mathsf{E}_i = (\mathsf{X}_i, \mathsf{Y}_i)$. This trivial construction typically isn't very helpful, though. One particular feature we might want is for a latent CIIR model to express the fact that "things we can do have been done before"; that is, any setting of the unobserved state $\mathsf{E}_c$ that our actions might yield has positive probability in the observed data. Example 4.3 illustrates model constructions with and without this property.

Example 4.3. Suppose we have a see-do model $(\mathbb{P}_\cdot, \mathsf{X}_{\mathbb{N}\cup\{c\}})$ where each $\mathsf{X}_i$ takes values in a binary set, and the control we can exert is to choose either $\mathbb{P}_0^{\mathsf{X}_c} = \frac{1}{4}\delta_0 + \frac{3}{4}\delta_1$ or $\mathbb{P}_1^{\mathsf{X}-c} = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$, independent of all other observations. Suppose further that for $i \in \mathbb{N}$, $\mathbb{P}_C^{\mathsf{X}_i} = \frac{3}{4}\delta_0 + \frac{1}{4}\delta_1$ independent of all other observations. Then we can consider this model to be IO contractible with latent binary inputs $\mathsf{E}_i$ such that

$$\mathbb{P}_\alpha^{\mathsf{X}_i|\mathsf{E}_i}(\cdot|e) = \delta_e$$

This is not the only way to construct such a model. We could instead choose latent binary inputs $\mathsf{E}'_i$ such that

$$\mathbb{P}_\alpha^{\mathsf{X}_i|\mathsf{E}'_i}(\cdot|e) = \begin{cases} \frac{3}{4}\delta_0 + \frac{1}{4}\delta_1 & e = 0 \\ \frac{1}{4}\delta_0 + \frac{3}{4}\delta_1 & e = 1 \end{cases}$$

On the other hand, the choice $\mathsf{E}''_i$ with

$$\mathbb{P}_\alpha^{\mathsf{X}_i|\mathsf{E}''_i}(\cdot|e) = \begin{cases} \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1 & e = 0 \\ \frac{1}{4}\delta_0 + \frac{3}{4}\delta_1 & e = 1 \end{cases}$$

cannot be latent binary inputs for a conditionally independent and identical response model, as the observational distribution cannot be written as any convex combination of $\mathbb{P}_\alpha^{\mathsf{X}_i|\mathsf{E}''_i}(\cdot|0)$ and $\mathbb{P}_\alpha^{\mathsf{X}_i|\mathsf{E}''_i}(\cdot|1)$.

In the first construction in Example 4.3, but not the following two, we have $\mathbb{P}_C^{\mathsf{E}_i} \gg \mathbb{P}_\alpha^{\mathsf{E}_c}$ for all $\alpha$. We say under this construction the options have precedent; they have, in a sense, "been done before". The assumption of precedent by itself has some implications – for example, if a decision maker considers precedent a reasonable assumption and they have access to a lot of data, they they should not expect any of their actions to lead to consequences that have never appeared before in the observational data. In Theorem 4.7, we will make use a stronger version of this assumption where the conditional distribution over $\mathsf{E}_i$ is different for each value of $\mathsf{Z}_i$, which leads to stronger conclusions. We will discuss the plausibility of the stronger assumption afterwards.

Theorem 4.7 is motivated by the following example:

Example 4.4. Suppose a decision maker collects data about a group of peope who have variously engaged the services of dietiticians, sporting coaches, general practitioners, bariatric surgeons and none of the above, with practitioner choice recorded under the variable $\mathsf{Z}_i$. The decision maker has also collected data on each person's body mass index $\mathsf{X}_i$ at the beginning of the study and followed mortality outcomes $\mathsf{Y}_i$ for a considerable period of time. A decision maker is reviewing this data, and in particular is wondering if steps they take to manage their weight $\mathsf{X}_c$ are likely to improve their own mortality prospects $\mathsf{Y}_c$.

Our decision maker presumes that each group of people $\mathsf{Z}_i$ has, in aggregate, different strategies for pursuing weight management and different contextual reasons for doing so (though, for the sake of this example, we suppose that the decision maker doesn't collect data on any of these facts). Because of this variation, the decison maker reasons, people in these different groups with different levels of body mass index should see different mortality results if, conditional on body mass index, the different circumstances and management strategies actually lead to different results. Conversely, if there is no variation in results for these different groups of people, then it would appear that, at least with regard to mortality, the eventual body mass index achieved is apparently the only important feature of any management plan.

This inference might fail if, for any reason, the variation in treatment plans and contexts between the different groups of people surveyed masks the variation in their effects. For example, if all groups of people overwhelmingly choose to pursue diet changes in the end and other dimensions of variation are simply not very important to the outcome, then their results will not reveal any variation in mortality outcomes due to different treatment strategies. Alternatively, it might be the case that everybody is making choices that achieve nearly optimal mortality prospects given their unobserved context and that the best achievable mortality outcomes are approximately the same for each person's achievable level of body mass index. In this case there may still be substantial variation in outcomes from different weight management strategies, but it is masked by the fact that everyone is making near-optimal choices.

If the decision maker finds that $\mathsf{Y}_i$ is not independent of $\mathsf{Z}_i$ given $\mathsf{X}_i$, they may also consider whether $\mathsf{Y}_i$ is independent of $\mathsf{Z}_i$ given $(\mathsf{V}_i, \mathsf{X}_i)$ for some set of covariates $\mathsf{V}_i$.

Theorem 4.7 establishes formal conditions for the informal deduction described in Example 4.4. We assume that all variables of interest are discrete, and make use of an alternative notation for discrete conditional probabilities.

Definition 4.5 (Index notation for discrete conditionals). Given a joint probability distribution $\mu^{\mathsf{XY}}$ with $\mathsf{X}$ and $\mathsf{Y}$ discrete, let $\mu_x^y := \mu^{\mathsf{Y}|\mathsf{X}}(\{y\}|x)$ and $\mu_X^Y := (x, y) \mapsto \mu_x^y$

The key assumption for Theorem 4.7 is an assumption we call diverse precedent. It's a rather complicated assumption. It imposes a domination condition that requires (roughly speaking) that the distribution of the latent input $\mathsf{E}_i$ given event $\mathsf{Z}_i = z$ almost surely dominates the distribution induced by any option we can choose (as in the discussion of precedent above) and is almost surely "diverse" for different values of $\mathsf{Z}_i$.

Definition 4.6 (Diverse precedent). Given a latent CIIR see-do model $(\mathbb{P}_\cdot, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i, \mathsf{Z}_i)_{i\in\mathbb{N}\cup\{c\}})$ with $E, X, Y$ and $Z$ all discrete, recall $\mathsf{G}$ is the directing random conditional of $(\mathbb{P}_\cdot, \mathsf{Z}_\mathbb{N}, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i)_{i\in\mathbb{N}})$.

We say that the options $C$ have diverse precedent with respect to $(\mathbb{P}_\cdot, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i, \mathsf{Z}_i)_{i\in\mathbb{N}\cup\{c\}})$ if $\mathbb{P}_\cdot$ satisfies the diversity condition:

$$\mathbb{P}_\alpha^{\mathsf{G}_Z^{EX}|\mathsf{G}_{EXZ}^Y}(\cdot|g_{EXZ}^Y) \ll U_{\Delta(E)} \qquad\qquad \forall \alpha, z, \mathbb{P}_\alpha - \text{almost all } g_{EXZ}^Y$$

as well as the precedent condition:

$$\mathbb{P}_\alpha^{\mathsf{E}_c|\mathsf{G}} \ll \sum_{z\in Z} \mathbb{P}_\alpha^{\mathsf{E}_i|\mathsf{G}}(\cdot|g) \qquad\qquad \mathbb{P}_\alpha - \text{almost all } g$$

Where $U_{\Delta(E)}$ is the uniform measure on the $|E - 1|$ simplex of discrete probability distributions with $|E|$ outcomes.

For Theorem 4.7, we assume that on the basis of observations we condition the probability on some event $I$ (in particular, we are interested in the case where $I$ is the event that a certain conditional independence holds).

Theorem 4.7 (Latent to observable IO contractibility). Given a latent CIIR see-do model $(\mathbb{P}_\cdot, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i, \mathsf{Z}_i)_{i\in\mathbb{N}\cup\{c\}})$ with $E, X, Y$ and $Z$ all discrete, recall $\mathsf{G}$ is the directing random conditional of $(\mathbb{P}_\cdot, \mathsf{Z}_\mathbb{N}, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i)_{i\in\mathbb{N}})$.

Let $I \subset \Delta(Y)^{XZ}$ be the event $\mathsf{G}_{Xz}^Y = \mathsf{G}_{Xz'}^Y$ for all $z, z' \in Z$; i.e. the event that $\mathsf{Y}_i$ is independent of $\mathsf{Z}_i$ conditional on $\mathsf{X}_i$ and $\mathsf{G}_{XZ}^Y$. Define $\mathbb{Q}_\alpha \in \Delta(\Omega)$ to be the probability measure such that, for all $A \in \mathcal{F}$

$$\mathbb{Q}_\alpha(A) := \mathbb{P}_\alpha^{\mathrm{id}_\Omega|\mathbb{1}_I\circ\mathsf{G}}(A|1)$$

i.e. $\mathbb{Q}_\alpha$ is $\mathbb{P}_\alpha$ conditioned on $\mathsf{G}_{XZ}^Y \in I$, so $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{Q}_\cdot}^e \mathsf{Z}_i|(\mathsf{X}_i, \mathrm{id}_C)$.

If the options $C$ have diverse precedent with respect to $(\mathbb{Q}_\cdot, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i, \mathsf{Z}_i)_{i\in\mathbb{N}\cup\{c\}})$, then $(\mathbb{Q}_\cdot, \mathsf{X}, \mathsf{Y})$ is also IO contractible.

Proof. We show that the assumption of conditional independence imposes a polynomial constraint on $\mathsf{G}_z^d$ which is nontrivial unless $\mathsf{Y}_i \perp\!\!\!\perp^e (\mathsf{Z}_i, \mathsf{E}_i, \mathrm{id}_C)|(\mathsf{X}_i, \mathsf{H})$, and hence the solution set $S$ for this constraint has measure 0 when this conditional independence does not hold.

Full proof in Appendix C. □

## 5 Diversity, causal discovery and the principle of independent causal mechanisms

In this section we will present a somewhat informal argument that th

We've already discussed the "precedent" part of the diverse precedent assumption: it is the assumption that whatever the decision maker can do has been done before. The interpretation of the diversity part of the assumption is less obvious. Informally, Theorem 4.7 argues as follows:

- If $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{Q}}^e \mathsf{Z}_i|(\mathsf{X}_i, \mathsf{G}, \mathrm{id}_C)$ then we must either have "alignment" between $\mathsf{G}_Z^{XE}$ and $\mathsf{G}_{EXZ}^Y$ or we also have $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{Q}}^e \mathsf{E}_i|(\mathsf{X}_i, \mathsf{G}, \mathrm{id}_C)$
- If we assume diversity of $\mathsf{G}_Z^{XE}$ conditioned on $\mathsf{G}_{EXZ}^Y$ – i.e. we rule out "alignment" – then we must conclude $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{Q}}^e \mathsf{E}_i|(\mathsf{X}_i, \mathsf{G}, \mathrm{id}_C)$

What we mean by "alignment" here is that $\mathsf{G}_Z^{XE}$ is restricted to a set of Lebesgue measure 0 conditioned on $\mathsf{G}_{EXZ}^{Y}$. The principal of independent causal mechanisms offers us one possible way of deciding whether such alignment is or is not likely [Lemeire and Janzing, 2013, Peters et al., 2017]. This is an informal principle that suggests conditionals like $\mathsf{G}_Z^{XE}$ and $\mathsf{G}_{EXZ}^{Y}$ will be "independent" or "unaligned" under causal structure assumptions that identify both of these conditionals as "causal mechanisms". It is further argued (for example in Peters et al. [2017, Ch. 2]) that conditionals that do not correspond to "causal mechanisms" may exhibit "alignment".

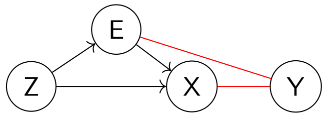
Given a directed graph $\mathcal{G}$, a "causal mechanism" is a conditional distribution $\mathsf{G}^X_{\mathrm{Pa}_{\mathcal{G}}(X)}$.

Causal discovery refers to a variety of strategies that aim to infer a causal structure from a set of observations. A common approach to causal discovery is to learn the causal structures that are faithful to the conditional independence structure of the observed data. We can use this approach to causal discovery to motivate the assumption of diversity as follows:

- Begin with a set of possible causal structures $\mathcal{G}$
- "Observe" $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{Z}_i|(\mathsf{X}_i, \mathsf{G}, \mathrm{id}_C)$ and retain the subset $\mathcal{G}'$ faithful to this conditional indepdenence
- Check whether $\mathsf{G}_Z^{XE}$ and $\mathsf{G}_{EXZ}^{Y}$ are independent causal mechanisms in the resulting set of causal structures

Example 5.1 (Using causal discovery to justify Theorem 4.7). Suppose we have a latent CIIR see-do model $(\mathbb{P}_\cdot, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i, \mathsf{Z}_i)_{i\in\mathbb{N}\cup\{c\}})$, and we hypothesize that it may be associated with any of the following graphs:

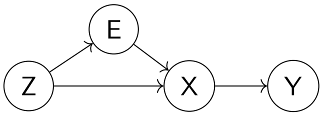
$\mathcal{G} :=$



(1)

Here red edges correspond to edges that may or may not be present, and may be oriented in any direction. Of these graphs, only one is faithful to the conditional indepdendence $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{Z}_i|(\mathsf{X}_i, \mathsf{G}, \mathrm{id}_C)$, yielding the graph

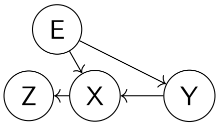
$\mathcal{G} =$



We note in this graph that $\mathrm{Pa}_{\mathcal{G}}(\mathsf{E}, \mathsf{X}) = \mathsf{Z}$ and $\mathrm{Pa}_{\mathcal{G}}(\mathsf{Y}) = \mathsf{X}$

Then Diagram (1) in combination with the principle of probabilistically independent causal mechanisms implies $\mathsf{G}_Z^{E} \perp\!\!\!\perp^e_{\mathbb{P}} \mathsf{G}_{EZ}^{XY}|\mathrm{Id}_C$. By Theorem ??, if we also assume that we have unconditional diversity of $\mathsf{G}_Z^E$ then we have diverse precedent. If, in addition, the observed data $\mathsf{T}$ enables us to conclude that $\mathsf{G}_{xz}^Y = \mathsf{G}_{xz'}^Y$ for all $z, z'$, then the conditions for Theorem 4.7 are satisfied. We can therefore determine the effect of any action $\alpha$ on $\mathsf{Y}$ if we already know its effect on $\mathsf{X}$.

On the other hand, if we observe the same conditional independence but associate the model with the following graph:



We note that $\mathsf{G}_{EXZ}^{Y}$ is not identified by this graph as a "causal mechanism", and therefore the principle of probabilistically independent causal mechanisms implies neither $\mathsf{G}_Z^{E} \perp\!\!\!\perp^e_{\mathbb{P}} \mathsf{G}_{EZ}^{XY}|\mathrm{Id}_C$ nor $\mathsf{G}_{EZ}^{X} \perp\!\!\!\perp^e_{\mathbb{P}} \mathsf{G}_{\mathsf{EXZ}}^{\mathsf{Y}}|\mathrm{Id}_C$. Thus in this case we cannot in general conclude from $\mathsf{G}_{xz}^Y = \mathsf{G}_{xz'}^Y$ that the effect of any action $\alpha$ on $\mathsf{Y}$ can be determined by its effect on $\mathsf{X}$.

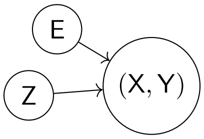
### 5.0.1 Connection to causal discovery

We will

The principle of independent causal mechanisms is usually used to derive rules to test for causal directions Peters et al. [2017, Ch. 4]. That is, rather than assuming causal directions and inferring independence of conditionals, this principle has typically been used to develop tests for independent conditionals in order to infer causal directions. For Theorem 4.7, we cannot test for all of the relevant independences because they involve the unobserved variable $\mathsf{E}$.

We might speculatively justify $\mathsf{E}_i$ as a causal parent of $\mathsf{X}_i$ and $\mathsf{Y}_i$ by referring to our original position that $\mathsf{E}_i$ represents some latent state which gives rise to probabilistically consistent response behaviour from $\mathsf{X}_i$ and $\mathsf{Y}_i$.

We also need to justify the assumption that $\mathsf{G}^{\mathsf{EZ}}$ is Lebesgue dominated. One way this might be violated is if a decision maker expects that $\mathsf{E}_i$ might be independent of $\mathsf{Z}_i$. In our original example this might occur if every doctor employs the same treatment plan (perhaps because there is a widely-agreed upon optimal plan) and has a similar mixture of patients. In that case, however, we would also observe the treatment $\mathsf{X}_i$ independent of $\mathsf{Z}_i$.



In this graph, $\mathsf{G}^{\mathsf{XY}}_{\mathsf{E}}$ and $\mathsf{G}^{\mathsf{EXY}}_{\mathsf{Z}}$ are identified as independent causal mechanisms.

However we arrive at them, once we have the assumptions that $\mathsf{E}_i$ is a causal parent of $\mathsf{X}_i$ and $\mathsf{Y}_i$ and that $\mathsf{G}^{\mathsf{EZ}}$ is Lebesgue dominated, a sufficient additional condition for Theorem ?? is that $\mathsf{Z}_i$ is a causal parent of $(\mathsf{X}_i, \mathsf{Y}_i)$. As both $\mathsf{Z}_i$ and $(\mathsf{X}_i, \mathsf{Y}_i)$ are observed

In light of the problems associated with conditioning on a set of measure 0, it would be very useful to extend Theorem 4.7 to an approximate result. Specifically, in the event $\mathsf{Y}_i$ is "approximately independent" of $\mathsf{Z}_i$ given $\mathsf{X}_i$ and $\mathsf{G}$, under what conditions is $\mathsf{Y}_i$ also approximately independent of $\mathsf{E}_i$ given $\mathsf{X}_i$ and $\mathsf{G}$? We speculate that a stronger version of the diverse precedent assumption will be necessary for such a theorem.

The diverse precedent assumption has a connection to the theory of causal graphical models. Meek [1995] justified the faithfulness condition for causal graphs associated with discrete probability models on the assumption that the distribution of parameters of a distribution consistent with a particular causal graph are dominated by the Lebesgue measure. In this theory, we have a discrete set of hypotheses over causal structures that imply some conditional independences, and Lebesge-dominated priors over the directing measure after conditioning on any of the causal structure hypotheses and their associated independences. Applying similar reasoning to the present case, we posit an argument along these lines: if we have the independence $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}} (\mathsf{E}_i, \mathsf{Z}_i)|(\mathsf{X}_i, \mathsf{G}, \mathrm{id}_C)$ but not the independence $\mathsf{E}_i \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{Z}_i|(\mathsf{G}, \mathrm{id}_C)$ and furthermore $\mathsf{Z}_i$ is an ancestor of $\mathsf{E}_i$ and $(\mathsf{E}_i, \mathsf{Z}_i)$ is an ancestor of $(\mathsf{X}_i, \mathsf{Y}_i)$ (so that $\mathsf{G}^E_Z$ and $\mathsf{G}^{XY}_{EZ}$ are associated with forward edges in the causal model) then the diverse precedent assumption may be supported. Note that it may be possible to rule out the independence $\mathsf{E}_i \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{Z}_i|(\mathsf{G}, \mathrm{id}_C)$ on the basis of the non-independence of $\mathsf{Z}_i$ and $\mathsf{X}_i$.

Another relation between theory of causal graphical models and the present work may be found in the causal version of the principle of maximum entropy [Sun et al., 2006, Janzing, 2021]. The causal version of the principle of maximum entropy, in contrast to the standard version of the principle, suggests that priors be specified by sequentially maximising the entropy of a cause, then maximising the conditional entropy of the first effect given the cause and so forth. While the cited articles discuss using the principle of entropy maximisation to specify prior distributions over observed variables rather than distributions over directing conditionals, the same principle may perhaps be applied to the specification of priors over directing conditionals. We posit that the causal version of the prinicple of maximum entropy might support a similar line of argument: if $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}} (\mathsf{E}_i, \mathsf{Z}_i)|(\mathsf{X}_i, \mathsf{G}, \mathrm{id}_C)$ but not $\mathsf{E}_i \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{Z}_i|(\mathsf{G}, \mathrm{id}_C)$ and $\mathsf{Z}_i$ is an ancestor of $\mathsf{E}_i$ and $(\mathsf{E}_i, \mathsf{Z}_i)$ is an ancestor of $(\mathsf{X}_i, \mathsf{Y}_i)$, then perhaps the causal version of the principle of maximum entropy offers some support for the diverse precedent assumption. Note that this (as well as the implication suggested in the previous paragraph) are highly speculative.

The causal version of the principle of maximum entropy is itself motivated by the general principle of independent causal mechanisms.

## 6 Conclusion

We employ a decision theoretic approach to causal inference to investigate two different approaches to answering the question “how do my observations relate to the consequences of my choices?”. Firstly, we examined the assumption of conditionally independent and identical responses, and its equivalent form in IO contractibility, which we argued was often an unreasonable assumption and secondly, we examined an approach based on the principle of precedent, or the idea that the decision maker’s options have been taken before, and some of their consequences observed. Our approach allows us to consider the question of what observations and consequences have in common independently from any prior knowledge the decision maker might have about how their choices influence outcomes – neither Theorem 3.18 nor Theorem 4.7 depend on any assumptions about a decision maker’s prior knowledge of the effects of their different options (though the plausibility of the assumptions in both theorems may well depend on such prior knowledge).

The grand aim of this work is to facilitate causal inference in situations where a decision maker has relatively little causal knowledge at the outset. We think avoiding structured interventions in this setting is advantageous because we regard the question of whether an action is known in advance to influence a particular variable as substantially more transparent than the question of whether it is well modeled by a structured intervention (of any type) on that variable.

Nevertheless, this work leaves many open questions for causal inference in the low prior knowledge setting. We have argued that the assumptions required for Theorem 3.18 are unlikely to be compelling in many situations. While the diverse precedent assumption may be more broadly plausible, it is at this stage difficult to evaluate. Speculatively, it may be possible to make progress on this question by better understanding when structural assumptions support this conclusion, via for example the causal version of the principle of maximum entropy.

For practical purposes, a generalisation of Theorem 4.7 to approximate independence is in order, and such a generalisation may also bring additional clarity to the diverse precedent assumption.

Despite these challenges, we are encouraged by a number of features of this work. Using decision making as a starting point for constructing models means that, at the outset, we are only making commitments a decision maker is likely to already be making if they want to apply a formal theory of decision making. The informal idea of precedent that underpins Theorem 4.7 seems like a general principle that may be applicable in a broad range of data-driven decision making problems. Finally, the apparent connection between Theorem 4.7 suggests that much of the work already done in the world of causal graphical models may be applicable to our alternative perspective. Causal inference under circumstances of limited prior knowledge presents many hard conceptual as well as practical problems, and our approach is a promising new avenue of investigation.

## References

Md. Bahadur Badsha and Audrey Qiuyan Fu. Learning Causal Biological Networks With the Principle of Mendelian Randomization. Frontiers in Genetics, 10, 2019. ISSN 1664-8021. URL https://www.frontiersin.org/articles/10.3389/fgene.2019.00460.

A. V. Banerjee, S. Chassang, and E. Snowberg. Chapter 4 - Decision Theoretic Approaches to Experiment Design and External Validity. In Abhijit Vinayak Banerjee and Esther Duflo, editors, Handbook of Economic Field Experiments, volume 1 of Handbook of Field Experiments, pages 141–174. North-Holland, January 2017. doi: 10.1016/bs.hefe.2016.08.005. URL https://www.sciencedirect.com/science/article/pii/S2214658X16300071.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable Causal Discovery from Interventional Data. In Advances in Neural Information Processing Systems, volume 33, pages 21865–21877. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f8b7aa3a0d349d9562b424160ad18612-Abstract.html.

Erhan Çinlar. Probability and Stochastics. Springer, 2011.

David Maxwell Chickering. Optimal Structure Identification with Greedy Search. J. Mach. Learn. Res., 3: 507–554, March 2003. ISSN 1532-4435. doi: 10.1162/153244303321897717. URL https://doi.org/10.1162/153244303321897717.

Kenta Cho and Bart Jacobs. Disintegration and Bayesian inversion via string diagrams. Mathematical Structures in Computer Science, 29(7):938–971, August 2019. ISSN 0960-1295, 1469-8072. doi: 10.1017/S0960129518000488.

Panayiota Constantinou and A. Philip Dawid. Extended conditional independence and applications in causal inference. The Annals of Statistics, 45(6):2618–2653, 2017. ISSN 0090-5364. URL http://www.jstor.org/stable/26362953.

Juan Correa and Elias Bareinboim. A Calculus for Stochastic Interventions:Causal Effect Identification and Surrogate Experiments. Proceedings of the AAAI Conference on Artificial Intelligence, 34(06):10093–10100, April 2020. ISSN 2374-3468. doi: 10.1609/aaai.v34i06.6567. URL https://ojs.aaai.org/index.php/AAAI/article/view/6567. Number: 06.

A. Philip Dawid. Decision-theoretic foundations for statistical causality. arXiv:2004.12493 [math, stat], April 2020. URL http://arxiv.org/abs/2004.12493. arXiv: 2004.12493.

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

Bruno de Finetti. Foresight: Its Logical Laws, Its Subjective Sources. In Samuel Kotz and Norman L. Johnson, editors, Breakthroughs in Statistics: Foundations and Basic Theory, Springer Series in Statistics, pages 134–174. Springer, New York, NY, [1937] 1992. ISBN 978-1-4612-0919-5. doi: 10.1007/978-1-4612-0919-5_10. URL https://doi.org/10.1007/978-1-4612-0919-5_10.

Frederick Eberhardt and Richard Scheines. Interventions and Causal Inference. Philos. Sci., 74, December 2007. doi: 10.1086/525638.

Dean Eckles and Eytan Bakshy. Bias and High-Dimensional Adjustment in Observational Studies of Peer Effects. Journal of the American Statistical Association, 116(534):507–517, April 2021. ISSN 0162-1459. doi: 10.1080/01621459.2020.1796393. URL https://doi.org/10.1080/01621459.2020.1796393.

Brendan Fong. Causal Theories: A Categorical Perspective on Bayesian Networks. arXiv: 1301.6201 [math], January 2013. URL http://arxiv.org/abs/1301.6201. arXiv: 1301.6201.

Patrick Forré and Joris M. Mooij. Constraint-based Causal Discovery for Non-Linear Structural Causal Models with Cycles and Latent Confounders. arXiv:1807.03024 [cs, stat], July 2018. URL http://arxiv.org/abs/1807.03024. arXiv: 1807.03024.

Tobias Fritz. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. Advances in Mathematics, 370:107239, August 2020. ISSN 0001-8708. doi: 10.1016/j.aim.2020.107239. URL https://www.sciencedirect.com/science/article/pii/S0001870820302656.

M. Maria Glymour and Donna Spiegelman. 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):81–85, January 2017. ISSN 0090-0036. doi: 10.2105/AJPH.2016.303539. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5308179/.

Brett R. Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook. SSRN Scholarly Paper ID 3033144, Social Science Research Network, Rochester, NY, September 2018. URL https://papers.ssrn.com/abstract=3033144.

Brett R. Gordon, Robert Moakler, and Florian Zettelmeyer. Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement. arXiv:2201.07055 [econ], January 2022. URL http://arxiv.org/abs/2201.07055. arXiv: 2201.07055.

Sander Greenland and James M Robins. Identifiability, Exchangeability, and Epidemiological Confounding. International Journal of Epidemiology, 15(3):413–419, September 1986. ISSN 0300-5771. doi: 10.1093/ije/15.3.413. URL https://doi.org/10.1093/ije/15.3.413.

Siyuan Guo, Viktor Tóth, Bernhard Schölkopf, and Ferenc Huszár. Causal de Finetti: On the Identification of Invariant Causal Structure in Exchangeable Data, March 2022. URL https://www.researchgate.net/publication/359574681_Causal_de_Finetti_On_the_Identification_of_Invariant_Causal_Structure_in_Exchangeable_Data.

Alain Hauser and Peter Bühlmann. Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs. Journal of Machine Learning Research, 13(79):2409–2464, 2012. ISSN 1533-7928. URL http://jmlr.org/papers/v13/hauser12a.html.

D. Heckerman and R. Shachter. Decision-Theoretic Foundations for Causal Reasoning. Journal of Artificial Intelligence Research, 3:405–430, December 1995. ISSN 1076-9757. doi: 10.1613/jair.202. URL https://www.jair.org/index.php/jair/article/view/10151.

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

Miguel A Hernán. Beyond exchangeability: The other conditions for causal inference in medical research. Statistical Methods in Medical Research, 21(1):3–5, February 2012. ISSN 0962-2802. doi: 10.1177/0962280211398037. URL https://doi.org/10.1177/0962280211398037.

Miguel A. Hernán. Does water kill? A call for less casual causal inferences. Annals of Epidemiology, 26 (10):674–680, October 2016. ISSN 1047-2797. doi: 10.1016/j.annepidem.2016.08.016. URL http://www.sciencedirect.com/science/article/pii/S1047279716302800.

Miguel A. Hernán and Stephen R. Cole. Invited Commentary: Causal Diagrams and Measurement Bias. American Journal of Epidemiology, 170(8):959–962, October 2009. ISSN 0002-9262. doi: 10.1093/aje/kwp293. URL https://academic.oup.com/aje/article/170/8/959/145135. Publisher: Oxford Academic.

Miguel A Hernán and James M Robins. Estimating causal effects from epidemiological data. Journal of Epidemiology and Community Health, 60(7):578–586, July 2006. ISSN 0143-005X. doi: 10.1136/jech.2004.029496. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2652882/.

Guido W. Imbens and Donald B. Rubin. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, Cambridge, 2015. ISBN 978-0-521-88588-1. doi: 10.1017/CBO9781139025751. URL https://www.cambridge.org/core/books/causal-inference-for-statistics-social-and-biomedical-sciences/71126BE90C58F1A431FE9B2DD07938AB.

Dominik Janzing. Causal versions of maximum entropy and principle of insufficient reason. Journal of Causal Inference, 9(1):285–301, January 2021. ISSN 2193-3685. doi: 10.1515/jci-2021-0022. URL https://www.degruyter.com/document/doi/10.1515/jci-2021-0022/html. Publisher: De Gruyter.

Olav Kallenberg. The Basic Symmetries. In Probabilistic Symmetries and Invariance Principles, Probability and Its Applications, pages 24–68. Springer, New York, NY, 2005. ISBN 978-0-387-28861-1. doi: 10.1007/0-387-28861-9_2. URL https://doi.org/10.1007/0-387-28861-9_2.

Finnian Lattimore and David Rohde. Causal inference with Bayes rule. arXiv:1910.01510 [cs, stat], October 2019a. URL http://arxiv.org/abs/1910.01510. arXiv: 1910.01510.

Finnian Lattimore and David Rohde. Replacing the do-calculus with Bayes rule. arXiv:1906.07125 [cs, stat], December 2019b. URL http://arxiv.org/abs/1906.07125. arXiv: 1906.07125.

Jan Lemeire and Dominik Janzing. Replacing Causal Faithfulness with Algorithmic Independence of Conditionals. Minds and Machines, 23(2):227–249, May 2013. ISSN 0924-6495, 1572-8641. doi: 10.1007/s11023-012-9283-1. URL https://link.springer.com/article/10.1007/s11023-012-9283-1.

D. V. Lindley and Melvin R. Novick. The Role of Exchangeability in Inference. The Annals of Statistics, 9 (1):45–58, 1981. ISSN 0090-5364. URL https://www.jstor.org/stable/2240868.

Christopher Meek. Strong Completeness and Faithfulness in Bayesian Networks. In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, UAI'95, pages 411–418, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-385-1. URL http://dl.acm.org/citation.cfm?id=2074158.2074205. event-place: Montréal, Qué, Canada.

Ignavier Ng, Shengyu Zhu, Zhitang Chen, and Zhuangyan Fang. A Graph Autoencoder Approach to Causal Structure Learning. November 2019. doi: 10.48550/arXiv.1911.07420. URL http://arxiv.org/abs/1911.07420. Number: arXiv:1911.07420 arXiv:1911.07420 [cs, stat].

Masashi Okamoto. Distinctness of the Eigenvalues of a Quadratic form in a Multivariate Sample. The Annals of Statistics, 1(4):763–765, 1973. ISSN 0090-5364. URL https://www.jstor.org/stable/2958321. Publisher: Institute of Mathematical Statistics.

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

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

Judea Pearl and Dana Mackenzie. The Book of Why: The New Science of Cause and Effect. Basic Books, New York, 1 edition edition, May 2018. ISBN 978-0-465-09760-9.

Jonas Peters and Peter Bühlmann. Structural Intervention Distance for Evaluating Causal Graphs. Neural Computation, 27(3):771–799, January 2015. ISSN 0899-7667. doi: 10.1162/NECO_a_00708. URL https://doi.org/10.1162/NECO_a_00708.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78(5):947–1012, 2016. ISSN 1467-9868. doi: 10.1111/rssb.12167. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12167.

Jonas Peters, Dominik Janzing, and Bernard Schölkopf. Elements of Causal Inference. MIT Press, 2017.

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

Olli Saarela, David A. Stephens, and Erica E. M. Moodie. The role of exchangeability in causal inference. June 2020. doi: 10.48550/arXiv.2006.01799. URL https://arxiv.org/abs/2006.01799v3.

Nino Scherrer, Olexa Bilaniuk, Yashas Annadani, Anirudh Goyal, Patrick Schwab, Bernhard Schölkopf, Michael C. Mozer, Yoshua Bengio, Stefan Bauer, and Nan Rosemary Ke. Learning Neural Causal Models with Active Interventions, March 2022. URL http://arxiv.org/abs/2109.02429. arXiv:2109.02429 [cs, stat].

P. Selinger. A Survey of Graphical Languages for Monoidal Categories. In Bob Coecke, editor, New Structures for Physics, Lecture Notes in Physics, pages 289–355. Springer, Berlin, Heidelberg, 2011. ISBN 978-3-642-12821-9. doi: 10.1007/978-3-642-12821-9_4. URL https://doi.org/10.1007/978-3-642-12821-9_4.

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

Peter Spirtes, Clark Glymour, and Richard Scheines. Causation, Prediction, and Search, volume 81. January 1993. doi: 10.1007/978-1-4612-2748-9.

Xiaohai Sun, Dominik Janzing, and Bernhard Schölkopf. Causal Inference by Choosing Graphs with Most Plausible Markov Kernels. January 2006.

Christian Toth, Lars Lorch, Christian Knoll, Andreas Krause, Franz Pernkopf, Robert Peharz, and Julius von Kügelgen. Active Bayesian Causal Inference, June 2022. URL http://arxiv.org/abs/2206.02063. arXiv:2206.02063 [cs, stat].

Karren Yang, Abigail Katoff, and Caroline Uhler. Characterizing and Learning Equivalence Classes of Causal DAGs under Interventions. In International Conference on Machine Learning, pages 5537–5546, July 2018. URL http://proceedings.mlr.press/v80/yang18a.html.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/e347c51419ffb23ca3fd5050202f9c3d-Abstract.html.

## A String Diagrams

We use a string diagram notation 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 A.

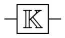
We make use of string diagram notation for probabilistic reasoning. Graphical models are often employed in causal reasoning, and string diagrams are a kind of graphical notation for representing Markov kernels. The notation comes from the study of Markov categories, which are abstract categories that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

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

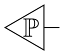
### A.1 Elements of string diagrams

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:

$\mathbb{K} :=$

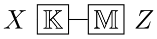


$\mu :=$



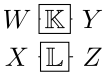
Given two Markov kernels $\mathbb{L} : X \rightarrowtriangle Y$ and $\mathbb{M} : Y \rightarrowtriangle Z$, the product $\mathbb{LM}$ is represented by drawing them side by side and joining their wires:

$\mathbb{LM} :=$



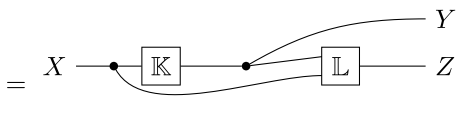
Given kernels $\mathbb{K} : W \rightarrowtriangle Y$ and $\mathbb{L} : X \rightarrowtriangle Z$, the tensor product $\mathbb{K} \otimes \mathbb{L} : W \times X \rightarrowtriangle Y \times Z$ is graphically represented by drawing kernels in parallel:

$\mathbb{K} \otimes \mathbb{L} :=$



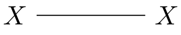
Given $\mathbb{K} : X \rightarrowtriangle Y$ and $\mathbb{L} : Y \times X \rightarrowtriangle Z$, the semidirect product is graphically represented by connecting $\mathbb{K}$ and $\mathbb{L}$ and keeping an extra copy

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



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

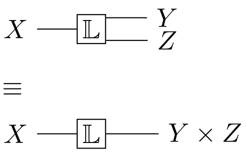
$\mathrm{Id}^X :=$



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 $X \times Y$:

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

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



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

## A.2 Special maps

Definition A.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.

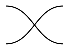
$$\mathrm{id}_X := \; X \text{-} X$$

Definition A.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:

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

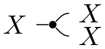
Definition A.3 (Swap map). The swap map $\mathrm{Swap}_{X,Y} : X \times Y \to Y \times X$ is defined by $(\mathrm{Swap}_{X,Y})(A \times B|x, y) = \delta_x(B)\delta_y(A)$ for $(x, y) \in X \times Y$, $A \in \mathcal{X}$ and $B \in \mathcal{Y}$. It swaps two inputs and is represented by crossing wires:

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



Definition A.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:

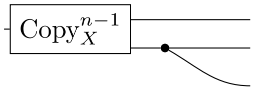
$\mathrm{Copy}_X :=$



Definition A.5 ($n$-fold copy map). The $n$-fold copy map $\mathrm{Copy}^n_X : X \to X^n$ is given by the recursive definition

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

$\mathrm{Copy}^n_X =$

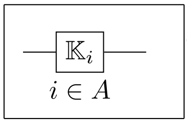


$n > 1$

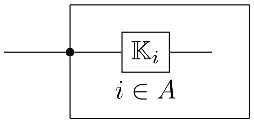
Plates In a string diagram, a plate that is annotated $i \in A$ means the tensor product of the $|A|$ elements that appear inside the plate. A wire crossing from outside a plate boundary to the inside of a plate indicates an $|A|$-fold copy map, which we indicate by placing a dot on the plate boundary. 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 $\mathbb{K}_i : X \rightarrowtriangle Y$ for $i \in A$,

$\bigotimes_{i\in A} \mathbb{K}_i :=$

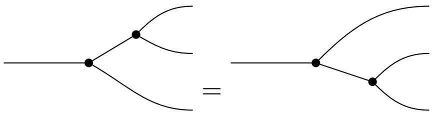


$\mathrm{Copy}_X^{|A|}(\bigotimes_{i\in A} \mathbb{K}_i) :=$

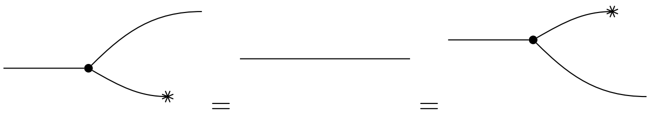


## A.3 Commutative comonoid axioms

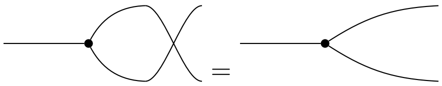
Diagrams in Markov categories satisfy the commutative comonoid axioms.



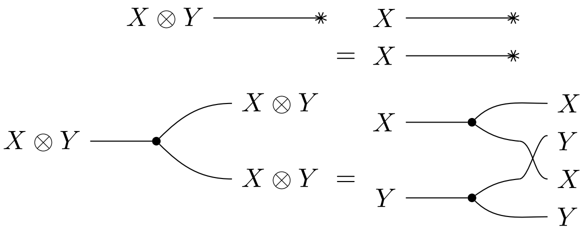
(2)



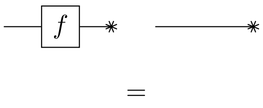
(3)



as well as compatibility with the monoidal structure



and the naturality of Del, which means that

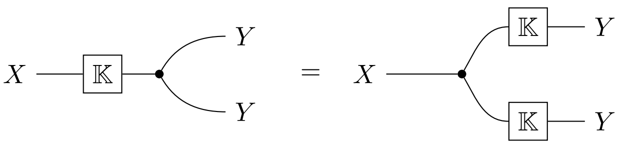


(4)

## A.4 Manipulating String Diagrams

Planar deformations along with the applications of Equations (2) through to Equation (4) are almost the only rules we have for transforming one string diagram into an equivalent one. One further rule is given by Theorem A.6.

Theorem A.6 (Copy map commutes for deterministic kernels [Fong, 2013]). For $\mathbb{K} : X \rightarrowtriangle Y$

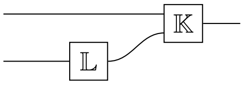


holds iff $\mathbb{K}$ is deterministic.

### A.4.1 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.

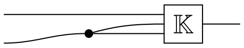
For arbitrary $\mathbb{K} : X \times Y \rightarrowtriangle Z$, $\mathbb{L} : W \rightarrowtriangle Y$



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

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

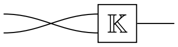
For arbitrary $\mathbb{K} : X \times Y \times Y \rightarrowtriangle Z$:



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

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

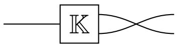
For arbitrary $\mathbb{K} : X \times Y \rightarrowtriangle Z$



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

The swap map before a kernel switches the input arguments.

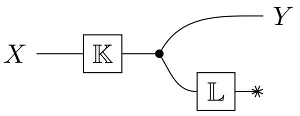
For arbitrary $\mathbb{K} : X \rightarrowtriangle Y \times Z$



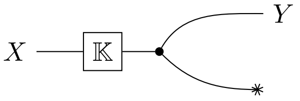
$$\begin{aligned}
&= \mathbb{K}\mathrm{Swap}_{YZ}\\
(\mathbb{K}\mathrm{Swap}_{YZ})(A \times B|x) &= \int_{Y\times Z} \delta_y(B)\delta_z(A)\mathbb{K}(\mathrm{d}y \times \mathrm{d}z|x)\\
&= \int_{B\times A} \mathbb{K}(\mathrm{d}y \times \mathrm{d}z|x)\\
&= \mathbb{K}(B \times A|x)
\end{aligned}$$

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

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

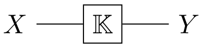


$=$



by Eq. (4)

$=$



by Eq. (3)

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

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

## B Symmetries of conditional probabilities

### B.1 Equality of equally sized contractions

This is the proof of Theorem 3.9.

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 (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 $i$th element of $A$ to the $i$th element of $B$ and vise versa. Note that $B \to A$ is the inverse of $A \to B$.

Lemma B.2 is used to extend conditional probabilities of finite sequences to infinite ones.

Lemma B.2 (Infinitely extended kernels). Given a collection of Markov kernels $\mathbb{K}_i : W \times X^{\mathbb{N}} \rightarrowtriangle Y^i$ for all $i \in \mathbb{N}$, if we have for every $j > i$

$$\mathbb{K}_j(\mathrm{id}_{Y^i} \otimes \mathrm{Del}_{Y^{j-i}}) = \mathbb{K}_i \otimes \mathrm{Del}_{X^{j-i}} \tag{5}$$

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

$$\mathbb{K}(\mathrm{id}_{Y^i} \otimes \mathrm{Del}_{Y^{\mathbb{N}}}) = \mathbb{K}_i \otimes \mathrm{Del}_{X^{j-i}}$$

Proof. Take any $x \in X^{\mathbb{N}}$ and let $x_{|m} \in X^n$ be the first $n$ elements of $x$. By Equation (5), 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 $\rho : [n] \to [n]$

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

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

$$\mathbb{Q}_x(\bigtimes_{i\in[n]} A_i \times Y^{\mathbb{N}}) = \mathbb{K}_n(\bigtimes_{i\in[n]} A_{\rho(i)}|x_{[n]}) \tag{6}$$

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

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

so by the Monotone convergence theorem, the sequence $\mathbb{Q}_x(\bigtimes_{i\in[n]} A_i)$ converges as $n \to \infty$ to $\mathbb{Q}_x(\bigtimes_{i\in\mathbb{N}} A_i)$. $x \mapsto \mathbb{Q}_x^{\mathsf{Z}_n}(\bigtimes_{i\in[n]} A_i)$ is measurable for all $n$, $\{A_i \in \mathcal{Y}|i \in \mathbb{N}\}$ by Equation (6), and so $x \mapsto Q_x$ is also measurable.

Thus $x \mapsto Q_x$ is the desired Markov kernel $\mathbb{K}$. □

Corollary B.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 $\mathsf{W}$ 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

$$\begin{aligned}\mathbb{P}^{\mathsf{X}_{[n]}|\mathsf{W}\mathsf{V}_{[n]}} \otimes \mathrm{Del}_{W^{\mathbb{N}}} &= \mathbb{P}^{\mathsf{X}|\mathsf{W}\mathsf{V}}(\mathrm{id}_{X^n} \otimes \mathrm{Del}_{X^{\mathbb{N}}})\\ \mathbb{P}^{\mathsf{Z}_{[n]}|\mathsf{W}\mathsf{Y}_{[n]}} \otimes \mathrm{Del}_{W^{\mathbb{N}}} &= \mathbb{P}^{\mathsf{Z}|\mathsf{W}\mathsf{Y}}(\mathrm{id}_{X^n} \otimes \mathrm{Del}_{X^{\mathbb{N}}})\end{aligned}$$

hence for all $n, m > n$

$$\begin{aligned}\mathbb{P}^{\mathsf{X}_{[m]}|\mathsf{W}\mathsf{V}_{[m]}}(\mathrm{id}_{X^n} \otimes \mathrm{Del}_{X^{m-n}}) &= \mathbb{P}^{\mathsf{Z}_{[m]}|\mathsf{V}\mathsf{Y}_{[m]}}(\mathrm{id}_{X^n} \otimes \mathrm{Del}_{X^{m-n}})\\ &= \mathbb{P}^{\mathsf{X}_{[n]}|\mathsf{W}\mathsf{V}_{[n]}} \otimes \mathrm{Del}_{W^{m-n}}\end{aligned}$$

and, in particular, by lemma B.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. □

Theorem 3.9. Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ and some $\mathsf{W}$, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{W}\mathsf{D}}$ is IO contractible over $\mathsf{W}$ if and only if for all subsequences $A, B \subset \mathbb{N}^{|A|}$ and for every $\alpha$

$$\begin{aligned}\mathbb{P}_\alpha^{\mathsf{Y}_A|\mathsf{W}\mathsf{D}_{A,\mathbb{N}\setminus A}} &= \mathbb{P}_\alpha^{\mathsf{Y}_B|\mathsf{W}\mathsf{D}_{B,\mathbb{N}\setminus B}}\\ &= \mathbb{P}_\alpha^{\mathsf{Y}_A|\mathsf{W}\mathsf{D}_A} \otimes \mathrm{del}_{D^{|\mathbb{N}\setminus A|}}\end{aligned}$$

Proof. Only if: For $Z \in \mathbb{N}^{|A|}$, let $\mathrm{del}_{Z^{\mathsf{c}}}$ be the Markov kernel associated with the map that sends $\mathsf{Y}$ to $\mathsf{Y}_Z := (\mathsf{Y}_i)_{i\in Z}$.

If $A$ is finite, then let $n := |A|$ and by exchange commutativity

$$\begin{aligned}\mathbb{P}_\alpha^{\mathsf{Y}_A|\mathsf{W}\mathsf{D}_{\mathsf{A},\mathbb{N}\setminus\mathsf{A}}} &= \mathbb{P}_\alpha^{\mathsf{Y}_A|\mathsf{W}\mathsf{D}_{\mathsf{A}\to[\mathsf{n}]}}\\ &= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{W}\mathsf{D}_{\mathsf{A}\to[\mathsf{n}]}}\mathrm{del}_{A^{\mathsf{c}}}\\ &= \mathbb{P}_\alpha^{\mathsf{Y}_{[n]\to A}|\mathsf{W}\mathsf{D}}\mathrm{del}_{A^{\mathsf{c}}}\end{aligned}$$

Use the fact that $[n] \to A \circ B \to [n] = B \to A$ and apply exchange commutativity to get

$$\begin{aligned}\mathbb{P}_\alpha^{\mathsf{Y}_{[n]\to A}|\mathsf{W}\mathsf{D}}\mathbb{F}_{\Pi_A} &= \mathbb{P}_\alpha^{\mathsf{Y}_{B\to A}|\mathsf{W}\mathsf{D}_{B\to[n]}}\mathrm{del}_{A^{\mathsf{c}}}\\ &= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{W}\mathsf{D}_{B\to[n]}}\mathrm{del}_{B^{\mathsf{c}}}\\ &= \mathbb{P}_\alpha^{\mathsf{Y}_B|\mathsf{W}\mathsf{D}_{\mathsf{B},\mathbb{N}\setminus\mathsf{B}}}\end{aligned}$$

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

$$\begin{aligned}\mathbb{P}_\alpha^{\mathsf{Y}_{A_m}|\mathsf{W}\mathsf{D}_{\mathsf{A_m}\to[\mathsf{m}]}} &= \mathbb{P}_\alpha^{\mathsf{Y}_{[m]}|\mathsf{W}\mathsf{D}}\\ &= \mathbb{P}_\alpha^{\mathsf{Y}_{B_m}|\mathsf{W}\mathsf{D}_{\mathsf{B_m}\to[\mathsf{m}]}}\end{aligned}$$

then by Corollary B.3

$$\mathbb{P}_\alpha^{\mathsf{Y}_{A}|\mathsf{W}\mathsf{D}_{\mathsf{A}\to[\mathsf{n}]}} = \mathbb{P}_\alpha^{\mathsf{Y}_{B_m}|\mathsf{W}\mathsf{D}_{\mathsf{B_m}\to[\mathsf{m}]}}$$

Finally, by locality

$$\mathbb{P}_\alpha^{\mathsf{Y}_{A}|\mathsf{W}\mathsf{D}_{\mathsf{A}\to[\mathsf{n}]}} = \mathbb{P}_\alpha^{\mathsf{Y}_{A}|\mathsf{W}\mathsf{D}_{A}} \otimes \mathrm{Del}_{D^{|\mathbb{N}\setminus A}}$$

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

### B.2 Examples of symmetries

These are the examples referenced in Section 3.2. Example B.4 shows that neither locality nor exchange commutativity is implied by the other.

Example B.4. 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 $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}} : D^{\mathbb{N}} \rightarrowtriangle Y^{\mathbb{N}}$ is given by

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

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

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

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

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

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 $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}} : D^{\mathbb{N}} \rightarrowtriangle Y^{\mathbb{N}}$ is given by

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

then

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

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

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

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

Although locality seems to an assumption that there is no interference between inputs and outputs of different indices, by itself it actually permits models with certain kinds of interference. This is shown in Example B.5.

Example B.5. Consider an experiment where I first flip a coin and record the results of this flip as the outcome $\mathsf{Y}_1$ 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 $\mathsf{D}_1 = 1$). $\mathsf{D}_2$ 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 $\mathsf{D}_1$ affects the joint distribution of the results of both steps, which is not ruled out by locality.

## B.3 Representation theorem preliminaries

This is the proof of Lemmas 3.16 and B.6 and Theorem 3.17 from Section 3.4. In addition, Lemmas B.6 and B.7 are presented and proved, which will be later used in the proof of Theorem 3.18.

The following definitions are reproduced for the reader's convenience. Note that these proofs use the string diagram notation explained in Appendix A.

Definition 3.10. Given a sequential input-output model $(\mathbb{P}_{\cdot}, \mathsf{D}, \mathsf{Y})$ 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 $i < k$.

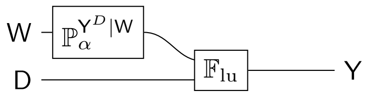
Definition 3.11. Given a sequential input-output model $(\mathbb{P}_{\cdot}, \mathsf{D}, \mathsf{Y})$ on $(\Omega, \mathcal{F})$, define the tabulated conditional distribution $\mathsf{Y}^D : \Omega \to Y^{\mathbb{N}\times D}$ by

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

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 $i$th 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.

Lemma 3.16. Suppose a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ is given with $D$ countable and $\mathsf{D}$ infinitely supported. Then for some $\mathsf{W}$, $\alpha$, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ is IO contractible if and only if

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



$$\iff$$

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

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

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

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

$$\mathbb{P}_\alpha^{(\mathsf{Y}_{ij}^D)_{\mathbb{N}\times D}|\mathsf{W}} = \mathbb{P}_\alpha^{(\mathsf{Y}_{\eta(i,j)}^D)_{\mathbb{N}\times D}|\mathsf{W}} \qquad (7)$$

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 $\mathsf{Y}^D$ 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 $\mathsf{D}_k$ is the $i$th 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. $\mathsf{R}_k$ returns the $(i,j)$ pair where $j$ is the value of $\mathsf{D}_k$ and $i$ is the count of $j$ occurrences up to $\mathsf{D}_k$). Let $\mathsf{R} : \mathbb{N} \to \mathbb{N}\times D$ by $k \mapsto \mathsf{R}_k$. $\mathsf{R}$ is almost surely bijective and

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

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

$$\begin{aligned}\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{WD}_{\mathsf{R}^{-1}}}(A|w,d) &= \mathbb{P}_\alpha^{\mathsf{Y}_{\mathsf{R}^{-1}}|\mathsf{WD}_{\mathsf{R}^{-1}}}(A|w,d)\\ &\overset{\mathbb{P}_\cdot}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}_{\mathsf{R}^{-1}}|\mathsf{WD}_{\mathsf{R}^{-1}}}(A|w,e)\\ &= \mathbb{P}_\alpha^{\mathsf{Y}^D}(A|w) \end{aligned} \tag{8}$$

for any $d \in D^{\mathbb{N}}$.

Now,

$$\mathbb{P}_\alpha^{\mathsf{Y}_{\mathsf{R}^{-1}}|\mathsf{WD}_{\mathsf{R}^{-1}}}(A|w,d) = \int_R \mathbb{P}_\alpha^{\mathsf{Y}_\rho|\mathsf{WD}_\rho}(A|d)\mathbb{P}_\alpha^{\mathsf{R}^{-1}|\mathsf{WD}_{\mathsf{R}^{-1}}}(\mathrm{d}\rho|w,d) \tag{9}$$

For each $\rho$, define $\rho^n : \mathbb{N} \to \mathbb{N}$ as the finite permutation that agrees with $\rho$ 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.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 (9)

$$\begin{aligned}\mathbb{P}_\alpha^{\mathsf{Y}_{\mathsf{R}^{-1}}|\mathsf{WD}_{\mathsf{R}^{-1}}}(A|w,d) &\overset{\mathbb{P}_\cdot}{\cong} \int_R \mathbb{P}_\alpha^{\mathsf{Y}_\rho|\mathsf{WD}_\rho}(A|d)\mathbb{P}_\alpha^{\mathsf{R}^{-1}|\mathsf{WD}_{\mathsf{R}^{-1}}}(\mathrm{d}\rho|w,d)\\ &\overset{\mathbb{P}_\cdot}{\cong} \int_R \mathbb{P}_\cdot^{\mathsf{Y}|\mathsf{WD}}(A|w,d)\mathbb{P}_\alpha^{\mathsf{R}^{-1}|\mathsf{WD}_{\mathsf{R}^{-1}}}(\mathrm{d}\rho|w,d)\\ &\overset{\mathbb{P}_\cdot}{\cong} \mathbb{P}_\cdot^{\mathsf{Y}|\mathsf{WD}}(A|w,d)\end{aligned} \tag{10}$$

for all $i, j \in \mathbb{N}$. Then by Equation (8) and Equation (10)

$$\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}(A|w) = \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}(A|w,e) \tag{11}$$

Take some $d \in D^{\mathbb{N}}$. From Equation (11) and IO contractibility of $\mathbb{P}_\cdot^{\mathsf{Y}|\mathsf{WD}}(A|e)$,

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

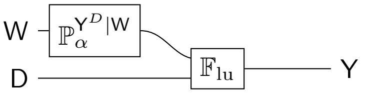
It remains to be shown that $\mathsf{Y}^D$ is invariant to finite permutations within rows. Consider some finite permutation within columns $\eta : \mathbb{N} \times D \to \mathbb{N} \times D$, note that $e_{\eta(i,j)} = j$ and hence $(e_{\eta(i,j)})_{i\in\mathbb{N},j\in D} = e$. Thus

$$\begin{aligned}
\mathbb{P}_\alpha^{(\mathsf{Y}^D_{\eta(i,j)})_{\mathbb{N}\times D}|\mathsf{W}}(A|w) &= \mathbb{P}_\alpha^{(\mathsf{Y}^D)_{\mathbb{N}\times D}|\mathsf{W}}\mathrm{Swap}_\eta(A|w) \\
&= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}\mathrm{Swap}_\eta(A|w,e) && \text{from Eq. (11)} \\
&= \mathbb{P}_\alpha^{\mathsf{Y}_\eta|\mathsf{WD}}(A|w,e) \\
&= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}_{\eta^{-1}}}(A|w,e) && \text{by exchange commutativity} \\
&= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}(A|w,(e_{\eta^{-1}(i,j)})_{i\in\mathbb{N},j\in D}) \\
&= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}(A|w,e) \\
&= \mathbb{P}_\alpha^{(\mathsf{Y}^D_{ij})_{\mathbb{N}\times D}|\mathsf{W}}(A|w) && \text{from Eq. (11)}
\end{aligned}$$

If: We construct a conditional probability according to Definition 3.11 and verify that it satisfies IO contractibility.

Suppose

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


W
$\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}$
D
$\mathbb{F}_{\mathrm{lu}}$
Y

where $\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}$ satisfies Equation (7).

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

$$\begin{aligned}
\mathbb{P}_\alpha^{\mathsf{Y}_S|\mathsf{WD}}(\bigtimes_{i\in[n]} A_i|w,d) &= \mathbb{P}_\alpha^{(\mathsf{Y}^D_{id_i})_{i\in S}|\mathsf{W}}(\bigtimes_{i\in[n]} A_i|w) \\
&= \mathbb{P}_\alpha^{(\mathsf{Y}^D_{S\leftrightarrow T(i)d_i})_{i\in S}|\mathsf{W}}(\bigtimes_{i\in[n]} A_i|w) \\
&= \mathbb{P}_\alpha^{(\mathsf{Y}^D_{id_{S\leftrightarrow T(i)}})_{i\in T}|\mathsf{W}}(\bigtimes_{i\in[n]} A_i|w) \\
&= \mathbb{P}_\alpha^{(\mathsf{Y}^D_{id'_i})_{i\in T}|\mathsf{W}}(\bigtimes_{i\in[n]} A_i|w) \\
&= \mathbb{P}_\alpha^{\mathsf{Y}_T|\mathsf{WD}}(\bigtimes_{i\in[n]} A_i|w,d')
\end{aligned}$$

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

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

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

$$\mathbb{P}_\alpha^{\mathsf{Y}_S|\mathsf{WD}}(\bigtimes_{i\in[n]} A_i|w,d) = (\mathbb{P}_\alpha^{\mathsf{Y}_{[n]}|\mathsf{WD}_{[n]}} \otimes \mathrm{Del}_{D^{\mathbb{N}}})(\bigtimes_{i\in[n]} A_i|w,d')$$

so $\mathbb{K}$ is IO contractible by Theorem 3.9. □

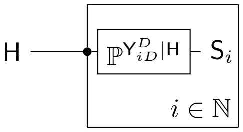
As a consequence of Lemma 3.16 along with De Finetti's representation theorem, we can say that given $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ IO contractible, conditioning on $\mathsf{H}$ renders the columns of $\mathsf{Y}^D$ independent and identically distributed.

Lemma B.6. Suppose a sequential input-output model $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ is given with $D$ countable, $\mathsf{D}$ infinitely supported over some $\mathsf{W}$ and $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ IO contractible over the same $\mathsf{W}$. Then, letting $\mathsf{H}$ be the directing random conditional of $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ (Definition 3.13) and $\mathsf{Y}^D_{iD} := (\mathsf{Y}^D_{ij})_{j\in D}$, we have for all $i \in \mathbb{N}$, $\mathsf{Y}^D_{iD} \perp\!\!\!\perp^e_{\mathbb{P}_\cdot} (\mathsf{Y}^D_{\mathbb{N}\setminus\{i\}D}, \mathsf{W}, \mathrm{id}_C)|\mathsf{H}$ and

$$\mathbb{P}_C^{\mathsf{Y}^D_{iD}|\mathsf{H}}(A|\nu) \overset{\mathbb{P}_\alpha}{\cong} \nu(A)$$

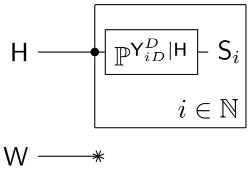
Proof. Fix $w \in W$ and consider $\mathbb{P}^{\mathsf{Y}^D}_{\alpha,w} := \mathbb{P}^{\mathsf{Y}^D|\mathsf{W}}_\alpha(\cdot|w)$. From Lemma 3.16, we have the exchangeability of the sequence $(\mathsf{Y}^D_{1D}, \mathsf{Y}^D_{2D}, ...)$ with respect to $(\mathbb{P}_{\alpha,w}, \Omega, \mathcal{F})$ as a special case of the invariance of $\mathbb{P}_\alpha^{(\mathsf{Y}^D_{ij})_{\mathbb{N}\times D}|\mathsf{W}}$ to permutations of rows. By the column exchangeability of $\mathbb{P}^{\mathsf{Y}^D}_{\alpha,w}$, from Kallenberg [2005, Prop. 1.4] (where $\mathsf{H}$ is precisely what Kallenberg calls the directing random measure)

$$\mathbb{P}^{\mathsf{Y}^D|\mathsf{H}}_{\alpha,w} =$$



Because the right hand side does not depend on $w$, we can say

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



and because it also does not depend on $\alpha$ we have $\mathsf{Y}^D \perp\!\!\!\perp^e_{\mathbb{P}_\cdot} (\mathsf{W}, \mathrm{id}_C)|\mathsf{H}$. Further application of Kallenberg [2005, Prop. 1.4] yields $\mathsf{Y}^D_{iD} \perp\!\!\!\perp^e_{\mathbb{P}_\cdot} (\mathsf{Y}^D_{\mathbb{N}\setminus\{i\}D}, \mathsf{W})|(\mathsf{H}, \mathrm{id}_C)$ and

$$\mathbb{P}_\alpha^{\mathsf{Y}^D_{iD}|\mathsf{H}}(A|\nu) \overset{\mathbb{P}_\alpha}{\cong} \nu(A)$$

Again, the right hand side does not depend on $\alpha$, which yields $\mathsf{Y}^D_{iD} \perp\!\!\!\perp^e_{\mathbb{P}_\cdot} (\mathsf{Y}^D_{\mathbb{N}\setminus\{i\}D}, \mathsf{W}, \mathrm{id}_C)|\mathsf{H}$. □

Theorem 3.17. Suppose a sequential input-output model $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ is given with $D$ countable, $\mathsf{D}$ infinitely supported and for some $\mathsf{W}$, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ is IO contractible for all $\alpha$. 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 $\mathsf{H}_A$, the directing random conditional of $(\mathbb{P}_\cdot, \mathsf{D}_A, \mathsf{Y}_A)$ is almost surely equal to $\mathsf{H}$, the directing random conditional of $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$.

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 $\mathsf{Y}^D$. Then we show that the contracted version of $\mathsf{Y}^D$ 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 $\#^{A,k}_j$ be the count restricted to $A$:

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

then

$$\begin{aligned}\mathsf{Y}^{D,A}_{ij} &:= \sum_{k\in A} [\![\#^{A,k}_j = i-1]\!][\![\mathsf{D}_k = j]\!]\mathsf{Y}_k \\ &= \sum_{k\in A} [\![\#^{A,k}_j = i-1]\!][\![\mathsf{D}_k = j]\!]\mathsf{Y}^D_{\mathsf{R}_k j}\end{aligned}$$

That is, defining $\mathsf{Q} : \mathbb{N} \to \mathbb{N}$ by $i \mapsto \sum_{k\in A}[\![\#^{A,k}_j = i-1]\!][\![\mathsf{D}_k = j]\!]\mathsf{R}_k$ then

$$\mathsf{Y}^{D,A}_{ij} = \mathsf{Y}^D_{\mathsf{Q}(i)j} \tag{12}$$

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 (12) is what is meant by "the subsequence $(\mathsf{D}_A, \mathsf{Y}_A)$ induces a random contraction over the rows of $\mathsf{Y}^D$". We will now show that $\mathsf{Y}^{D,A}$ has the same distribution as $\mathsf{Y}^D$.

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

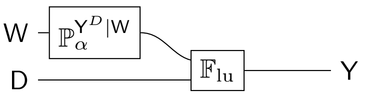
$$\begin{aligned}
\mathbb{P}_\alpha^{\mathsf{Y}^{D,A}|\mathsf{WQ}}(B|w,q) &= \mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}\mathrm{con}_q(B|w) \\
&= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}\mathrm{con}_q(B|w,e) && \text{by Eq.(11)} \\
&= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}(B|w,e) && \text{by Theorem 3.9} \\
&= \mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}(B|w) && \text{by Eq.(11)} \qquad (13)
\end{aligned}$$

Finally, take $\mathsf{H}_A$ the directing random measure of $\mathsf{Y}^{D,A}$. We conclude from the equality Eq. (13) and from the fact that there is a one-to-one map from directing random measures to exchangeable distributions that $\mathsf{H}_A \overset{\mathbb{P}_\alpha}{\cong} \mathsf{H}$. □

The following is a technical lemma that will be used in Theorem 3.18.

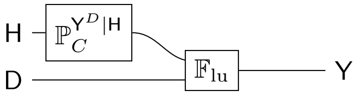
Lemma B.7. Suppose a sequential input-output model $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ is given with $D$ countable, $\mathsf{D}$ infinitely supported over $\mathsf{W}$, for some $\mathsf{W}$, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ is IO contractible for all $\alpha$ and for all $\alpha$

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


W
$\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}$
D
$\mathbb{F}_{\mathrm{lu}}$
Y

then $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_\cdot} \mathsf{W}|(\mathsf{H}, \mathsf{D}, \mathrm{id}_C)$ and for all $\alpha$

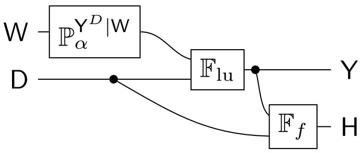
$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{HD}} =$


H
$\mathbb{P}_C^{\mathsf{Y}^D|\mathsf{H}}$
D
$\mathbb{F}_{\mathrm{lu}}$
Y

Proof. We show that the function that maps the variables $\mathsf{Y}$ and $\mathsf{D}$ to $\mathsf{H}$ also maps $\mathsf{Y}^D$ and the constant $e \in D^{\mathbb{N}}$ to $\mathsf{H}'$ with $\mathsf{H}' \overset{\mathbb{P}_\cdot}{\cong} \mathsf{H}$, and the result follows from disintegration along with a conditional independence given by Lemma 3.16.

$\mathsf{Y}^D$ is a function of $\mathsf{Y}$ and $\mathsf{D}$ (see Definition 3.11) and $\mathsf{H}$ is a function of $\mathsf{Y}^D$. Say $f : Y \times D \to H$ is such that $\mathsf{H} = f(\mathsf{Y}, \mathsf{D})$ (see Definition 3.12). Because $\mathsf{H} = f(\mathsf{Y}, \mathsf{D})$, we have $\mathsf{H} \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{W}, \mathrm{id}_C)|(\mathsf{Y}, \mathsf{D})$. Thus

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


W
$\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}$
D
$\mathbb{F}_{\mathrm{lu}}$
Y
$\mathbb{F}_f$
H

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

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

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

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

Let $g := (y^D, d) \mapsto f_d \circ \mathrm{lu}(y^D, d)$ 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 $A \in \mathcal{Y}^D$

$$\mathbb{P}_\alpha(g(\mathsf{Y}^D, d)(A) \bowtie g(\mathsf{Y}^D, d')(A)) = \int_H \mathbb{P}_\alpha^{\mathrm{Id}_\Omega|\mathsf{H}}(g(\mathsf{Y}^D, d)(A) \bowtie g(\mathsf{Y}^D, d')(A)|\nu)\mathbb{P}_\alpha^{\mathsf{H}}(\mathrm{d}\nu)$$

Note that

$$\mathbb{P}_\alpha^{\mathrm{Id}_\Omega|\mathsf{H}}(g(\mathsf{Y}^D, d)(A) \bowtie \nu(A)|\nu) = \mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{H}}(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \prod_{j\in D} \mathbb{1}_{A_j}(y^D_{\#_{[d=j]_i},j}) \bowtie \nu(A)|\nu)\mathbb{P}_\alpha^{\mathsf{H}}(\mathrm{d}\nu)$$

by independent permutability of the rows of $\mathsf{Y}^D$ (Lemma 3.16), for each row we can send $\#_{[d=j]_i}$ to $i$ and obtain

$$\begin{aligned}\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{H}}(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \prod_{j\in D} \mathbb{1}_{A_j}(y^D_{\#_{[d=j]_i},j}) \bowtie \nu(A)|\nu)\mathbb{P}_\alpha^{\mathsf{H}}(\mathrm{d}\nu) &= \mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{H}}(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \prod_{j\in D} \mathbb{1}_{A_j}(y^D_{i,j}) \bowtie \nu(A)|\nu)\\ &= \mathbb{P}_\alpha^{\mathsf{Y}^D_{iD}|\mathsf{H}}(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{A}(y^D_{i,D}) \bowtie \nu(A)|\nu)\end{aligned}$$

but by Lemma B.6, the sequence $(\mathsf{Y}^D_{iD})_{i\in\mathbb{N}}$ are mutually independent conditional on $\mathsf{H}$ and for all $\alpha$, $\mathbb{P}_\alpha^{\mathsf{Y}_{iD}|\mathsf{H}}(A|\nu) \overset{\mathbb{P}_C}{\cong} \nu(A)$. 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

$$\begin{aligned}&\int_H \mathbb{P}_\alpha^{\mathrm{Id}_\Omega|\mathsf{H}}(g(\mathsf{Y}^D, d)(A) \bowtie g(\mathsf{Y}^D, d')(A)|\nu)\mathbb{P}_\alpha^{\mathsf{H}}(\mathrm{d}\nu)\\ &= \int_H \mathbb{P}_\alpha^{\mathrm{Id}_\Omega|\mathsf{H}}(g(\mathsf{Y}^D, d)(A) \bowtie \nu(A) \cap g(\mathsf{Y}^D, d')(A) \bowtie \nu(A)|\nu)\mathbb{P}_\alpha^{\mathsf{H}}(\mathrm{d}\nu)\\ &= 1\end{aligned}$$

Because this holds for all $A$,

$$g(\mathsf{Y}^D, d) \overset{\mathbb{P}_\alpha}{\cong} g(\mathsf{Y}^D, d') \qquad \text{as this holds for all } A$$

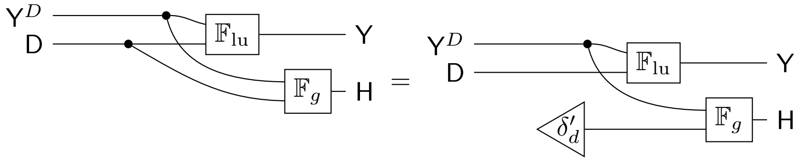
And, as a consequence, defining

$$i : (y^d, d, d') \mapsto (\mathrm{lu}(\mathsf{Y}^D, d), g(\mathsf{Y}^D, d'))$$

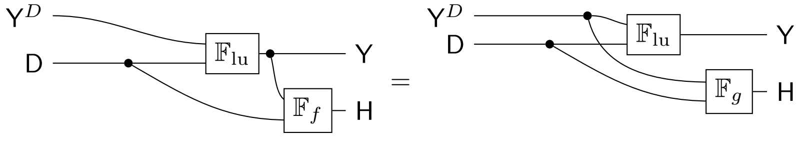
we have

$$i(y^d, d, d) \overset{\mathbb{P}_\alpha}{\cong} i(y^d, d, d')$$

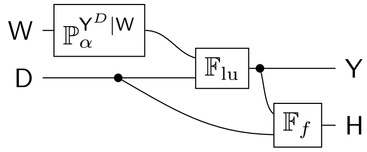
which in turn implies the almost sure equality of the associated Markov kernels:



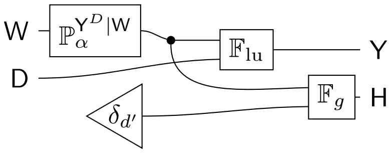
but we also have, by the definitions of $f$ and $g$



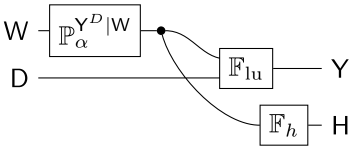
we can therefore write $\mathbb{P}_\alpha^{\mathsf{YH|WD}}$ as



$=$

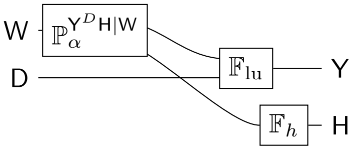


$=:$



because $\mathsf{H}$ is a deterministic function of $\mathsf{Y}^D$, this implies

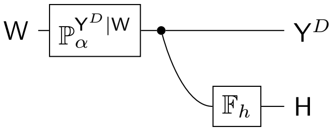
$$\mathbb{P}_\alpha^{\mathsf{YH|WD}} =$$



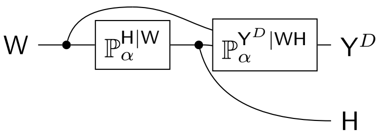
(14)

Noting that $\mathbb{F}_h \otimes \mathrm{Del}_W = \mathbb{P}_\alpha^{\mathsf{H|Y}^D\mathsf{W}}$

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



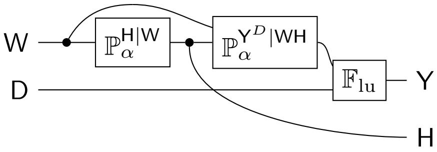
$=$


W
$\mathbb{P}_\alpha^{\mathsf{H}|\mathsf{W}}$
$\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{WH}}$
$\mathsf{Y}^D$
H

(15)

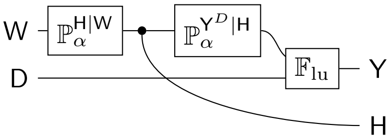
and so by substituting Equation (15) into (14)

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


W
$\mathbb{P}_\alpha^{\mathsf{H}|\mathsf{W}}$
$\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{WH}}$
D
$\mathbb{F}_{\text{lu}}$
Y
H

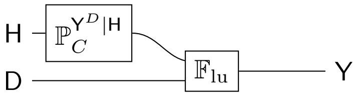
From Lemma 3.16 we also have $\mathsf{Y}^D \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{W}, \mathrm{id}_C)|\mathsf{H}$ , so

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


W
$\mathbb{P}_\alpha^{\mathsf{H}|\mathsf{W}}$
$\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{H}}$
D
$\mathbb{F}_{\text{lu}}$
Y
H

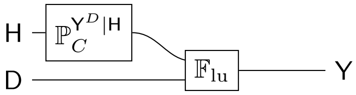
and so by higher order conditionals $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{W}|(\mathsf{H}, \mathsf{D}, \mathrm{id}_C)$ and

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



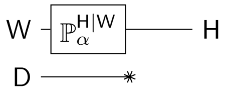
Because the right hand side does not depend on $\alpha$, we finally have $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{W}, \mathrm{id}_C)|(\mathsf{H}, \mathsf{D})$ and the result

$$\mathbb{P}_C^{\mathsf{Y|HD}} =$$



Furthermore, by marginalising the right hand side of Equation B.3 we have

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



Hence $\mathsf{H} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{D}|(\mathsf{W}, \mathrm{id}_C)$. □

B.4 Representation theorem

This is the proof of the main result from Section 3, Theorem 3.18.

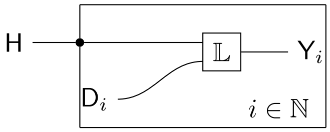
Theorem 3.18. Suppose a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ with sample space $(\Omega, \mathcal{F})$ is given with $D$ countable and $\mathsf{D}$ infinitely supported. Then the following are equivalent:

1. There is some $\mathsf{W}$ such that $\mathbb{P}_\alpha^{\mathsf{Y|WD}}$ is IO contractible for all $\alpha$
2. For all $i$, $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{\neq i}, \mathsf{D}_{\neq i}, \mathrm{id}_C)|(\mathsf{H}, \mathsf{D}_i)$ and for all $i, j$

$$\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{HD}_i} \overset{\mathbb{P}_\alpha^{\mathsf{D}_i|\mathsf{H}}}{\cong} \mathbb{P}_C^{\mathsf{Y}_j|\mathsf{HD}_j}$$

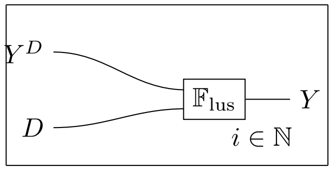
3. There is some $\mathbb{L} : H \times X \rightarrowtriangle Y$ such that

$\mathbb{P}_C^{\mathsf{Y|HD}} =$



Proof. As a preliminary, we will show

$\mathbb{F}_{\text{lu}} =$



(16)

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

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

Recall that lu is the function

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

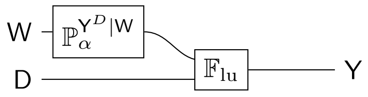
By definition, for any $\{A_i \in \mathcal{Y} | i \in \mathbb{N}\}$

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

which is what we wanted to show.

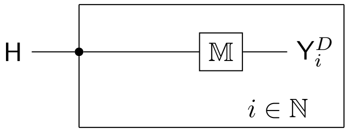
(1) $\Longrightarrow$ (3): From Lemma 3.16, we have some $\mathsf{Y}^D$ such that

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



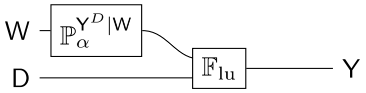
and by Lemma B.6

$$\mathbb{P}_C^{\mathsf{Y}^D|\mathsf{H}} = \tag{17}$$



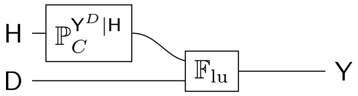
By Lemma 3.16, for each $w \in W$

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



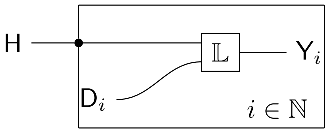
and so by Lemma B.7

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{HD}} = \tag{18}$$



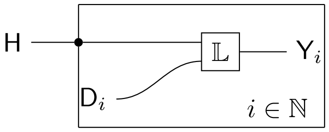
We can substitute Equations (17) and (16) into (18) for

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{HD}} =$$



(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 $\alpha \in C$

$$\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{HD}_i\mathsf{Y}_{\neq i}\mathsf{D}_{\neq i}} \overset{\mathbb{P}_C}{\cong} \mathbb{L} \otimes \mathrm{Del}_{Y^{\mathbb{N}} \times X^{\mathbb{N}}}$$

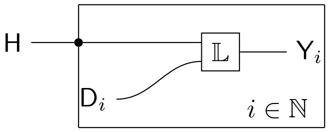
hence $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{\neq i}, \mathsf{D}_{\neq i}, \mathrm{id}_C)|(\mathsf{H}, \mathsf{D}_i)$

(2) $\implies$ (1): Take $\mathsf{W} := \mathsf{H}$. Because we assume $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{[1,i)}, \mathsf{D}_{[1,i),\mathrm{id}_C})|(\mathsf{H}, \mathsf{D}_i)$ we can take $\mathbb{L} := \mathsf{H}_X^Y = \mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{HX}_i}$ for all $i, \alpha$ (existence given by Theorem 3.3) and

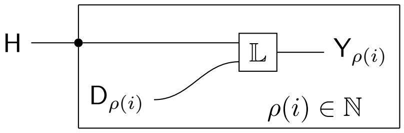
$$\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{HD}_i\mathsf{Y}_{[1,i)}\mathsf{D}_{[1,i)}} \overset{\mathbb{P}_C}{\cong} \mathbb{L} \otimes \mathrm{Del}_{Y^{i-1} \times X^{i-1}}$$

by taking the semidirect product of the conditionals

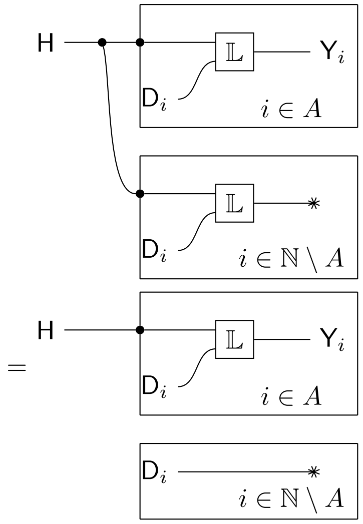
$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{HD}} =$$



$$=$$



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



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

### B.5 Consequences of Theorem 3.18

Theorem 3.18 says that a data independent sequential input-output model $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ features conditionally independent and identical response functions $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{H}\mathsf{D}_i}$ for all $\alpha$ if and only if there is some $\mathsf{W}$ such that $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{W}\mathsf{D}}$ is IO contractible over $\mathsf{W}$ for all $\alpha$.

A simple special case to consider is when $\mathsf{W}$ is single valued – that is, when $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}}$ is IO contractible. As Theorem B.8 shows, this corresponds to the CIIR sequence models where the inputs $\mathsf{D}$ are unconditionally data-independent and independent of the hypothesis $\mathsf{H}$. We can also consider the case where $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ is only exchange commutative over $*$. This corresponds to models where the inputs $\mathsf{D}$ are data-independent and the hypothesis $\mathsf{H}$ depends on a symmetric function of the inputs $\mathsf{D}$ (under some side conditions).

Theorem B.8 (Data-independent IO contractibility). Suppose a sequential input-output model $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ 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 $\alpha$. Then the following are equivalent:

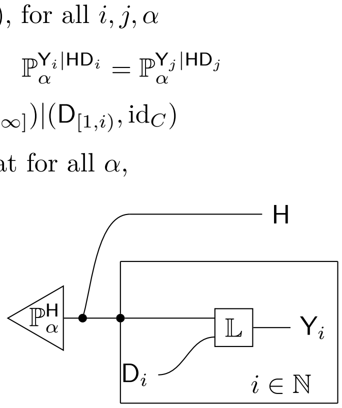
1. $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}}$ is IO contractible for all $\alpha$
2. For all $i$, $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_\cdot} (\mathsf{Y}_{\neq i}, \mathsf{D}_{\neq i}, \mathrm{id}_C)|(\mathsf{H}, \mathsf{D}_i)$, for all $i, j, \alpha$

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

   , $\mathsf{H} \perp\!\!\!\perp^e_{\mathbb{P}_\cdot} \mathsf{D}|\mathrm{id}_C$ and for all $i$ $\mathsf{D}_i \perp\!\!\!\perp^e_{\mathbb{P}_\cdot} \mathsf{D}_{(i,\infty]})|(\mathsf{D}_{[1,i)}, \mathrm{id}_C)$

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

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



Proof. See Appendix B.5. □

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

Theorem B.9. If $\mathbb{P}_\alpha^{\mathsf{Y|D}}$ is exchange commutative, and for each $\alpha$ $\mathbb{P}_\alpha^{\mathsf{D}}$ is absolutely continuous with respect to some exchangeable distribution $\mathbb{Q}_\alpha^{\mathsf{D}}$ in $\Delta(D^{\mathbb{N}})$ with directing random measure $\mathsf{F}$ and $\mathsf{D}$ infinitely supported over $\mathsf{F}$ with respect to $\mathbb{Q}_\alpha$ , then $\mathbb{P}_\alpha^{\mathsf{Y|HD}}$ is IO contractible, where $\mathsf{H}$ 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 $\mathsf{G}$. □

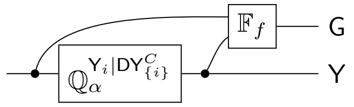
Lemma B.10 (Exchangeably dominated conditionals). Given $(\mathbb{P}_C, \Omega, \mathcal{F})$ and variables $\mathsf{D}, \mathsf{Y}$, if for any $\alpha$ there is some $\mathbb{Q}_\alpha$ such that $\mathbb{Q}_\alpha^{\mathsf{DY}}$ is exchangeable with directing random measure $\mathsf{G}$, $\mathsf{D}$ is infinitely supported over $\mathsf{G}$ with respect to $\mathbb{Q}_\alpha$ and for any $i$, $\mathbb{Q}_\alpha^{\mathsf{Y}_i|\mathsf{DY}_{\{i\}^\complement}} \overset{\mathbb{P}}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{DY}_{\{i\}^\complement}}$ then $\mathbb{P}_\alpha^{\mathsf{Y|HD}}$ is IO contractible (where $\mathsf{H}$ is the directing random conditional for $\mathbb{P}_\alpha^{\mathsf{Y|D}}$).

Proof. By Kallenberg [2005, Prop. 1.4], there is a $\mathsf{G}$ such that $(\mathsf{D}_i, \mathsf{Y}_i) \perp\!\!\!\perp^e_{\mathbb{Q}_C} (\mathsf{D}_{\{i\}^\complement}\mathsf{Y}_{\{i\}^\complement})|(\mathsf{G}, \mathrm{id}_C)$ and for all $i, j$

$$\mathbb{Q}_\alpha^{\mathsf{Y}_i\mathsf{D}_i|\mathsf{G}} = \mathbb{Q}_\alpha^{\mathsf{Y}_j\mathsf{D}_j|\mathsf{G}} \tag{19}$$

There is some function $f : D^{\mathbb{N}} \times Y^{\mathbb{N}}$ such that $\mathsf{G} = f(\mathsf{D}, \mathsf{Y})$, i.e.

$$\mathbb{Q}_\alpha^{\mathsf{Y_iG|DY}_{\{i\}^\complement}} = \mathsf{D}, \mathsf{Y}_{\{i\}^C}$$



$$\begin{aligned} &\overset{P}{\cong} \mathbb{P}_\alpha^{\mathsf{Y_iG|DY}_{\{i\}^\complement}} \\ \implies \mathbb{Q}_\alpha^{\mathsf{Y_i|GDY}_{\{i\}^\complement}} &\overset{P}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}_i\mathsf{|GDY}_{\{i\}^\complement}} \end{aligned} \tag{20}$$

It follows from weak union that

$$\begin{aligned} &\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}_C} (\mathsf{D}_{\{i\}^\complement}\mathsf{Y}_{\{i\}^\complement})|(\mathsf{D}_i, \mathsf{G}, \mathrm{id}_C) \\ \iff &\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{GY}_{\{i\}^\complement}\mathsf{D}_{\{i\}^\complement}}(A|d_i, g, d, y) \overset{P}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{G}}(A|d_i, g) \qquad \forall A, d_i, g, d, y, \alpha \\ \implies &\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{D}_{\{i\}^\complement}\mathsf{Y}_{\{i\}^\complement})|(\mathsf{D}_i, \mathsf{G}, \mathrm{id}_C) \end{aligned} \tag{21}$$

where Eq. (21) follows from Eq. (20).

Finally, from Eq. (19) and Eq. (21)

$$\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{G}} \overset{\mathbb{P}}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}_j\mathsf{D}_j|\mathsf{G}}$$

Thus $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ features independent and identical responses conditioned on $\mathsf{G}$, and by Lemma 3.19 it also has independent and identical responses conditioned on $\mathsf{H}$. Finally, the infinite support of $\mathsf{D}$ over $\mathsf{G}$ with respect to $\mathsf{Q}_\alpha$ implies $\mathsf{D}$ is also infinitely supported over $\mathsf{G}$ with respect to $\mathbb{P}_\alpha$, so by Theorem 3.18 $\mathbb{P}_\alpha^{\mathsf{Y|HD}}$ is IO contractible. □

Theorem 3.20. 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 $\mathsf{D}$ infinitely supported over $\mathsf{G}$ only if for any sets $A, B \subset \mathbb{N}$ such that $\mathsf{D}_A$ and $\mathsf{D}_B$ are also infinitely supported over $\mathsf{G}$ and any $i, j \in \mathbb{N}$ such that $i \notin A$, $j \notin B$,

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

. 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 3.18 and Lemma 3.19, $\mathbb{P}_\alpha^{\mathsf{Y|HD}}$ is IO contractible. By Theorem 3.17, $\mathsf{H}$ 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 $E \in \mathcal{Y}, d_i \in D, d \in D^{\mathbb{N}}, y \in Y^{\mathbb{N}}$

$$\begin{aligned}\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{Y}_A,\mathsf{D}_A}(E|d_i,y,d) &= \mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}}(E|d_i,f(y,d))\\ &= \mathbb{P}_\alpha^{\mathsf{Y}_j|\mathsf{D}_j\mathsf{H}}(E|d_i,f(y,d)) \qquad (22)\\ &= \mathbb{P}_\alpha^{\mathsf{Y}_j|\mathsf{D}_j\mathsf{Y}_B,\mathsf{D}_B}(E|d_i,y,d)\end{aligned}$$

Where Eq. (22) follows from Theorem 3.9.

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.10. □

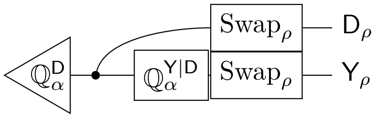
Theorem B.11. Given $(\mathbb{P}_c dot, \mathsf{Y}, \mathsf{D})$, if $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}}$ is exchange commutative for each $\alpha$, and for each $\alpha$ $\mathbb{P}_\alpha^{\mathsf{D}}$ is absolutely continuous with respect to some exchangeable distribution $\mathbb{Q}_\alpha^{\mathsf{D}}$ in $\Delta(D^{\mathbb{N}})$ with directing random measure $\mathsf{F}$, and if $\mathsf{D}$ is infinitely supported over $\mathsf{F}$ with respect to $\mathbb{Q}_\alpha$, then $(\mathbb{P}_c dot, \mathsf{Y}, \mathsf{D})$ is IO contractible.

Proof. For each $\alpha$, extend $\mathbb{Q}_\alpha^{\mathsf{D}}$ to a distribution on $(\mathsf{D}, \mathsf{Y})$ by asserting that $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}} \overset{\mathbb{Q}_\alpha}{\cong} \mathbb{Q}_\alpha^{\mathsf{Y}|\mathsf{D}}$. Because $\mathbb{Q}_\alpha^{\mathsf{D}}$ dominates $\mathbb{P}_\alpha^{\mathsf{D}}$, we have in fact $\mathbb{Q}_\alpha^{\mathsf{Y}|\mathsf{D}} \overset{\mathbb{P}}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}}$

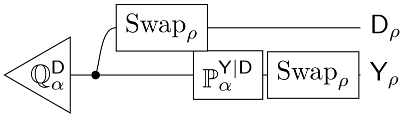
We will show $\mathbb{Q}_\alpha^{\mathsf{DY}}$ is unchanged by finite permutations of $(\mathsf{D}_i, \mathsf{Y}_i)$ pairs. For some finite permutation $\rho : \mathbb{N} \to \mathbb{N}$:

$$\begin{aligned}\mathbb{Q}_\alpha^{\mathsf{D}_\rho\mathsf{Y}_\rho} &= \mathbb{Q}_\alpha^{\mathsf{D}_\rho\mathsf{Y}_\rho}(\mathrm{Swap}_{\rho,D^{\mathbb{N}}} \otimes \mathrm{Swap}_{\rho,Y^{\mathbb{N}}})\\ &= \mathbb{Q}_\alpha^{\mathsf{D}} \odot \mathbb{Q}_\alpha^{\mathsf{Y}|\mathsf{D}}(\mathrm{Swap}_{\rho,D^{\mathbb{N}}} \otimes \mathrm{Swap}_{\rho,Y^{\mathbb{N}}})\end{aligned}$$

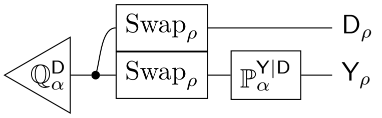
=



=

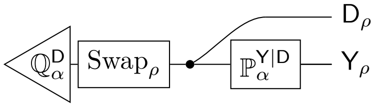


=



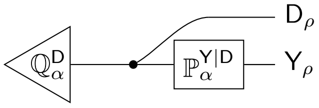
(23)

=



(24)

=



(25)

$$= \mathbb{Q}_\alpha^{\mathsf{DY}}$$

Where line (23) follows from exchange commutativity, (24) follows from Theorem A.6 and the fact that the swap map is deterministic and line (25) comes from the exchangeability of $\mathbb{Q}_\alpha^{\mathsf{D}}$.

Because $\mathbb{P}_\alpha^{\mathsf{D}}$ is dominated by $\mathbb{Q}_\alpha^{\mathsf{D}}$ by assumption, we have $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}} \overset{\mathbb{P}}{\cong} \mathbb{Q}_\alpha^{\mathsf{Y}|\mathsf{D}}$, which implies $\mathbb{Q}_\alpha^{\mathsf{Y}_i|\mathsf{DY}_{\{i\}\complement}} \overset{\mathbb{P}}{\cong} \mathbb{Q}_\alpha^{\mathsf{Y}_i|\mathsf{DY}_{\{i\}\complement}}$ and from Lemma B.10 we therefore have $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{HD}}$ IO contractible over $\mathsf{H}$, and from Theorem 3.18 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 $\mathsf{H}$ also. □

# C Precedented options

## C.1 IO contractibility from diverse precedent

This is the proof of Theorem 4.7 in Section 4.

Definition 4.6. Given a latent CIIR see-do model $(\mathbb{P}_\cdot, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i, \mathsf{Z}_i)_{i\in\mathbb{N}\cup\{c\}})$ with $E, X, Y$ and $Z$ all discrete, recall $\mathsf{G}$ is the directing random conditional of $(\mathbb{P}_\cdot, \mathsf{Z}_\mathbb{N}, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i)_{i\in\mathbb{N}})$.

We say that the options $C$ have diverse precedent with respect to $(\mathbb{P}_\cdot, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i, \mathsf{Z}_i)_{i\in\mathbb{N}\cup\{c\}})$ if $\mathbb{P}_\cdot$ satisfies the diversity condition:

$$\mathbb{P}_\alpha^{\mathsf{G}_Z^{EX}|\mathsf{G}^Y_{EXZ}}(\cdot|g^Y_{EXZ}) \ll U_{\Delta(E)} \qquad \forall\alpha, z, \mathbb{P}_\alpha - \text{almost all } g^Y_{EXZ}$$

as well as the precedent condition:

$$\mathbb{P}_\alpha^{\mathsf{E}_c|\mathsf{G}} \ll \sum_{z\in Z} \mathbb{P}_\alpha^{\mathsf{E}_i|\mathsf{G}}(\cdot|g) \qquad \mathbb{P}_\alpha - \text{almost all } g$$

Where $U_{\Delta(E)}$ is the uniform measure on the $|E-1|$ simplex of discrete probability distributions with $|E|$ outcomes.

Theorem 4.7. Given a see-do model $(\mathbb{P}_\cdot, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i, \mathsf{Z}_i)_{i\in\mathbb{N}\cup\{c\}})$ with $E, X, Y$ and $Z$ all discrete sets, suppose among the observations $i \in \mathbb{N}$ the pairs $(\mathsf{Z}_i, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i))$ share conditionally independent and identical responses and, for all observations and consequences $i \in \mathbb{N} \cup \{c\}$, pairs $(\mathsf{E}_i, (\mathsf{X}_i, \mathsf{Y}_i))$ also share conditionally independent and identical responses. Take $\mathsf{G}$ to be the directing random conditional of $(\mathbb{P}_\cdot, \mathsf{Z}_\mathbb{N}, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i)_{i\in\mathbb{N}})$.

Let $I \subset \Delta(Y)^{XZ}$ be the event $\mathsf{G}^Y_{Xz} = \mathsf{G}^Y_{Xz'}$ for all $z, z' \in Z$; i.e. the event that $\mathsf{Y}_i$ is independent of $\mathsf{Z}_i$ conditional on $\mathsf{X}_i$ and $\mathsf{G}$. Define $\mathbb{Q}_\alpha \in \Delta(\Omega)$ to be the probability measure such that, for all $A \in \mathcal{F}$

$$\mathbb{Q}_\alpha(A) := \mathbb{P}_\alpha^{\mathrm{id}_\Omega|\mathbb{1}_I\circ\mathsf{G}}(A|1)$$

i.e. $\mathbb{Q}_\alpha$ is $\mathbb{P}_\alpha$ conditioned on $\mathsf{G}^Y_{XZ} \in I$, so $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}_\cdot} \mathsf{Z}_i|(\mathsf{X}_i, \mathrm{id}_C)$.

If the options $C$ have diverse precedent with respect to $(\mathbb{Q}_\cdot, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i, \mathsf{Z}_i)_{i\in\mathbb{N}\cup\{c\}})$, then $(\mathbb{Q}_\cdot, \mathsf{X}, \mathsf{Y})$ is also IO contractible.

Proof. We apply the diversity condition to show that $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{E}_i|(\mathsf{Z}_i, \mathsf{X}_i, \mathsf{G}, \mathrm{id}_C)$ for $i \in \mathbb{N}$. We then apply the precedent condition to extend this independence to $\mathsf{Y}_c \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{E}_c|(\mathsf{Z}_c, \mathsf{X}_c, \mathsf{G}, \mathrm{id}_C)$ to complete the proof.

Note that by construction of $\mathbb{Q}_\alpha$ we have $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{Z}_i|(\mathsf{X}_i, \mathsf{G}, \mathrm{id}_C)$. This in turn implies, for all $\alpha$ the following holds $\mathbb{Q}_\alpha$-almost surely:

$$\sum_{e\in E} \mathsf{G}^y_{exz} \frac{\mathsf{G}^x_{ez}\mathsf{G}^e_z}{\sum_{e'\in E}\mathsf{G}^x_{e'z}\mathsf{G}^{e'}_z} \overset{\mathbb{Q}_\alpha}{\cong} \sum_{e\in E} \mathsf{G}^y_{exz'} \frac{\mathsf{G}^x_{ez'}\mathsf{G}^e_{z'}}{\sum_{e'\in E}\mathsf{G}^x_{e'z'}\mathsf{G}^{e'}_{z'}}$$

Conditioning on $\mathsf{G}^Y_{EXZ} = g^Y_{EXZ}$

$$\sum_{e\in E} g^y_{exz} \frac{\mathsf{G}^x_{ez}\mathsf{G}^e_z}{\sum_{e'\in E}\mathsf{G}^x_{e'z}\mathsf{G}^{e'}_z} \overset{\mathbb{P}_C}{\cong} \sum_{e\in E} \mathsf{g}^y_{exz'} \frac{\mathsf{G}^x_{ez'}\mathsf{G}^e_{z'}}{\sum_{e'\in E}\mathsf{G}^x_{e'z'}\mathsf{G}^{e'}_{z'}} \tag{26}$$

Eq. (26) defines a polynomial constraint on $\mathsf{G}^{\mathsf{Ex}}_{\{z,z'\}}$ for each $x, z, z'$. If $g^y_{exz} = g^y_{e'xz}$ for all $e, e'$ and likewise $g^y_{exz'} = g^y_{e'xz'}$, then this constraint is trivial; it is satisfied for every possible value of $\mathsf{G}^x_{E\{z,z'\}}$.

We will show that, unless $g^y_{exz} = g^y_{e'xz}$ for all $e, e'$ and $z$, that this constraint is nontrivial for some $z$. Consequently, the set of solutions for $\mathsf{G}^x_{EZ}$ subject to the restriction $g^y_{exz} \neq g^y_{e'xz}$ has Lebesgue measure 0. We will do this by showing that, assuming $g^y_{exz} > g^y_{e^<xz}$ for some $e, e^<$, we can find alternative realisations of $\mathsf{G}^e_z$ that lead to unequal values of the left hand side of Eq (26) without affecting the right hand side.

Let $g^x_{ez}$ and $g^e_z$ be a possible realisation of $\mathsf{G}^x_{ez}$ and $\mathsf{G}^e_z$. Assuming $g^y_{exz} > g^y_{e^<xz}$, either $g^x_{ez} = g^x_{e^<z}$, $g^x_{ez} < g^x_{e^<z}$ or $g^x_{ez} > g^x_{e^<z}$. Consider the first case, and take $g'$ such that $g'^e_z = 0.5g^e_z$ and $g'^{e^<}_z = g^{e^<}_z + 0.5g^e_z$ and equal to

$g_z^{e''}$ for all other $e'' \in E$. Note that $g_z^{\prime E}$ is also a possible realisation of $\mathsf{G}_z^e$, as it is everywhere positive and sums to 1, and $g_z^{\prime\prime e} < g_z^e$ almost surely as $g_z^e > 0$ almost surely. Then

$$\frac{g^x_{ez} g^e_z}{\sum_{e'\in E} g^x_{e'z} g^{e'}_z} > \frac{g^x_{ez} g^{\prime e}_z}{\sum_{e'\in E} g^x_{e'z} g^{\prime e'}_z}$$
$$\frac{g^x_{e^< z} g^{e^<}_z}{\sum_{e'\in E} g^x_{e'z} g^{e'}_z} < \frac{g^x_{e^< z} g^{\prime e^<}_z}{\sum_{e'\in E} g^x_{e'z} g^{\prime e'}_z}$$

because by assumption the denominator remains the same. But then

$$\sum_{e\in E} g^y_{exz} \frac{g^{x\;\;e}_{ezz}}{\sum_{e'\in E} g^{x\;\;e'}_{e'zz}} > \sum_{e\in E} g^y_{exz'} \frac{g^{x\;\;\prime e}_{ezz'}}{\sum_{e'\in E} g^{x\;\;\prime e'}_{e'z'z'}} \tag{27}$$

because on the right side a smaller term in the sum receives more weight, a larger term receives less weight and all other terms are weighted equally.

Consider $g^x_{ez'} > g^x_{e^< z'}$. Then we still have

$$\frac{g^x_{ez} g^e_z}{\sum_{e'\in E} g^x_{e'z} g^{e'}_z} > \frac{g^x_{ez} g^{\prime e}_z}{\sum_{e'\in E} g^x_{e'z} g^{\prime e'}_z}$$
$$\frac{g^x_{e^< z} g^{e^<}_z}{\sum_{e'\in E} g^x_{e'z} g^{e'}_z} < \frac{g^x_{e^< z} g^{\prime e^<}_z}{\sum_{e'\in E} g^x_{e'z} g^{\prime e'}_z}$$

For the second inequality, the right hand numerator grows and the denominator shrinks. For the first, note that

$$\frac{g^x_{ez} g^{\prime e}_z}{\sum_{e'\in E} g^x_{e'z} g^{\prime e'}_z} = \frac{0.5 g^x_{ez} g^e_z}{\sum_{e'\in E} g^x_{e'z} g^{e'}_z - 0.5 g^e_z (g^x_{ez} - g^x_{e^< z})}$$

$g^e_z g^x_{ez} < 1$ (an almost sure event) implies that the right hand denominator is greater than $0.5 \sum_{e'\in E} g^x_{e'z} g^{e'}_z$, and hence the right hand side is less than $\frac{g^x_{ez} g^e_z}{\sum_{e'\in E} g^x_{e'z} g^{e'}_z}$.

Thus the conclusion in Eq. (27) follows for the same reasons as before. Considering $g^x_{ez'} < g^x_{e^< z'}$, analogous reasoning implies Eq. (27) once again.

Thus, unless $g^y_{exz} = g^y_{e'xz}$ for all $e, e'$ and $z$, Eq. (26) implies a nontrivial constraint on $\mathsf{G}^x_{Ez}$ for some $z$. Thus for some $e, e'$,$z$, $x$ and $y$ the set of solutions $S := \{g^X_{EZ} | \mathsf{G}^X_{EZ} = g^X_{EZ}$ satisfies Eq. (26) for all $x, z \wedge g^y_{exz} \neq g^y_{e'xz}\}$ has Lebesgue measure 0 [Okamoto, 1973], and so by domination

$$\mathbb{Q}_\alpha^{\mathsf{G}^X_{EZ}|\mathsf{G}^{XY}_{EZ}}(S|g^{XY}_{EZ}) = 0$$

On the other hand, by assumption, the set $T := \{g^E_z | \mathsf{G}^E_z = g^E_z$ satisfies Eq. (26)$\}$ has measure 1. Thus we conclude that with the exception of a $\mathbb{Q}_\alpha$ measure 0 set, $g^y_{exz} = g^y_{e'xz}$. That is, $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{E}_i | (\mathsf{Z}_i, \mathsf{X}_i, \mathsf{G}, \mathrm{id}_C)$.

From the diversity condition, we have $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{E}_i | (\mathsf{Z}_i, \mathsf{X}_i, \mathsf{G}, \mathrm{id}_C)$. By contraction with $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{Z}_i | (\mathsf{X}_i, \mathsf{G}, \mathrm{id}_C)$, we have $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}} (\mathsf{Z}_i, \mathsf{E}_i) | (\mathsf{X}_i, \mathsf{G}, \mathrm{id}_C)$.

By CIIR of the $(\mathsf{E}_i | (\mathsf{X}_i, \mathsf{Y}_i))$ pairs, we have for all $i$, $\mathbb{Q}_\alpha^{\mathsf{Y}_i \mathsf{X}_i | \mathsf{E}_i \mathsf{G}} \overset{\mathbb{Q}_\alpha^{\mathsf{E}_i|\mathsf{G}}}{\cong} \mathbb{Q}_\alpha^{\mathsf{Y}_c \mathsf{X}_c | \mathsf{E}_c \mathsf{G}}$. Because we have a representative version $\mathsf{G}^{XY}_E$ of $\mathbb{Q}_\alpha^{\mathsf{Y}_i \mathsf{X}_i | \mathsf{E}_i \mathsf{G}} forall i \in \mathbb{N}$ (Theorem 3.3) and precedent implies that any set of measure 0 with respect to $\mathbb{Q}_\alpha^{\mathsf{E}_i|\mathsf{G}}$ for all $i \in \mathbb{N}$ also has measure 0 with respect to $\mathbb{Q}_\alpha^{\mathsf{E}_c|\mathsf{G}}$, we have

$$\mathsf{G}^{XY}_E \overset{\mathbb{Q}_\alpha^{\mathsf{E}_c|\mathsf{G}}}{\cong} \mathbb{Q}_\alpha^{\mathsf{Y}_c \mathsf{X}_c | \mathsf{E}_c \mathsf{G}}$$

and thus

$$\mathsf{G}^Y_X \overset{\mathbb{Q}_\alpha^{\mathsf{X}_c|\mathsf{G}}}{\cong} \mathbb{Q}_\alpha^{\mathsf{Y}_c | \mathsf{X}_c \mathsf{G}}$$

completing the proof. □

### C.2 Diverse precedent from independent causal mechanisms

Here we prove Theorem ??.

Theorem ??. Consider a latent CIIR see-do model $(\mathbb{P}_{\cdot}, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i, \mathsf{Z}_i)_{i\in\mathbb{N}\cup\{c\}})$ and define $\mathbb{Q}_{\cdot}$ as $\mathbb{P}_{\cdot}$ conditioned on $\mathbb{1}_I = 1$ where $\mathbb{1}_I := [\![\mathsf{G}^Y_{Xz} = \mathsf{G}^Y_{Xz'}]\!]$ for all $z, z' \in Z$.

If either of the following hold:

$$\mathbb{P}^{\mathsf{G}^E_z|\mathsf{G}^E_{z'}}(\cdot|g^E_{z'}) \ll U_{\Delta(E)} \text{ almost all } g^E_{z'} \qquad \text{and } \mathsf{G}^E_Z \perp\!\!\!\perp^e_{\mathbb{P}} \mathsf{G}^{XY}_{EZ}|\mathrm{Id}_C \tag{28}$$

$$\mathbb{P}^{\mathsf{G}^X_{Ez}|\mathsf{G}^X_{Ez'}}(\cdot|g^X_{Ez'}) \ll U_{\Delta(X)} \text{ almost all } g^X_{Ez'} \qquad \text{and } \mathsf{G}^X_{EZ} \perp\!\!\!\perp^e_{\mathbb{P}} \mathsf{G}^{\mathsf{Y}}_{\mathsf{EXZ}}|\mathrm{Id}_C \tag{29}$$

then the options $C$ have diverse precedent with respect to $(\mathbb{Q}_{\cdot}, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i, \mathsf{Z}_i)_{i\in\mathbb{N}\cup\{c\}})$.

Proof. If the conditions on line (28) hold, then we require $(\mathsf{G}^E_{z'}, \mathsf{G}^E_z) \perp\!\!\!\perp^e_{\mathbb{P}} (\mathsf{G}^{XY}_{EZ}, 1_I)|\mathrm{id}_C$ for diverse precedent. $\mathbb{1}_I = [\![\mathsf{G}^Y_{Xz} = \mathsf{G}^Y_{Xz'}]\!]$ is determined by $\mathsf{G}^{\mathsf{XY}}_{\mathsf{EZ}}$, so by decomposition it is sufficient to show $(\mathsf{G}^E_{z'}, \mathsf{G}^E_z) \perp\!\!\!\perp^e_{\mathbb{P}} \mathsf{G}^{XY}_{EZ}|\mathrm{id}_C$, but this follows directly from the assumption $\mathsf{G}^{\mathsf{E}}_{\mathsf{Z}} \perp\!\!\!\perp^e_{\mathbb{P}} \mathsf{G}^{\mathsf{XY}}_{\mathsf{EZ}}|\mathrm{Id}_C$.

If the conditions on line (29) hold, then we require $(\mathsf{G}^X_{Ez'}, \mathsf{G}^X_{Ez}) \perp\!\!\!\perp^e_{\mathbb{P}} (\mathsf{G}^{XY}_{EZ}, 1_I)|\mathrm{id}_C$ for diverse precedent. As $\mathbb{1}_I$ is also determined by $\mathsf{G}^Y_{EXZ}$, this follows by an argument analogous to the above. □