Research Article Open Access

David Johnston*, Cheng Soon Ong, and Robert C. Williamson

Learning Consequences Without Interventions

DOI: DOI, Received ..; revised ..; accepted ..

Abstract: Broadly valid causal inference is a tantalizing but elusive prospect. Assumptions are required to make any progress at all, and these assumptions cannot all be washed out by sufficiently large datasets. We argue that the use of structured interventions is a convention that frequently embodies assumptions not entailed by prior knowledge or by the given data. Instead of modelling structured interventions, the decision theoretic approach to causal modelling grounds causal models in the problem of making good decisions. A key role of structural interventions is to identify relationships that are invariant between the observations and the consequences of an action. We show how invariant relationships can be analysed using decision theoretic approach with fewer additional commitments. Our first result is an equivalence between models with "invariant conditionals" and 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 somewhat mysterious assumption of an "unknown true distribution". We discuss the fact that IO contractibility often seems to be an unreasonable assumption. Our second positive result is that IO contractibility may be implied by a combination of an observed conditional independence and a weaker assumption of "precedence", 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

MSC: 62C99

1 Introduction

Judea Pearl's causal hierarchy [37] distinguishes between three types of problems: prediction problems, intervention problems, and counterfactual problems. 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 is an important insight into the differences between causal inference and classical statistics, we feel the terminology is confusing. In Pearl's theory, *structural interventions* are used to model intervention problems. Structural interventions are operations that transform a probability distribution according to a *graphical causal model*. In our view, problems involving structural interventions should not be considered synonymous to "intervention problems" which, as we see it, 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?". This kind 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 put forward as the appropriate tool for modelling the consequences of different options (that is, "intervention problems" broadly understood). However, making use of structural interventions in contexts where a decision maker's

***Corresponding Author: David Johnston:** Australian National University; E-mail: davidoj@fastmail.com.au
Cheng Soon Ong: Data61; E-mail; chengsoon.ong@anu.edu.au
Robert C. Williamson: Universität Tübingen; Email: bob.williamson@uni-tuebingen.de

background causal knowledge is unreliable or incomplete poses a conundrum. Our decision maker has some set of options C to consider, and selecting an option α from this set is their decision problem. Given a dataset, they may attempt to infer causal relationships using causal discovery, or to estimate certain structural intervention-based effects of interest from the data along with whatever causal assumptions they supplied at the outset. In standard practice, they do *not* attempt to learn a correspondence between each option α and a structural intervention – thus, this correspondence must apparently be provided by their prior knowledge. However, such a correspondence seemingly goes beyond the kind of prior knowledge a decision maker can be expected to have.

The most common kind of structural intervention is known as a *perfect* or *hard* intervention. A perfect intervention is often denoted with the symbol $\mathrm{do}(\mathsf{X} = x)$. Given a probability distribution $\mathbb{P}$, a variable X and a causal graphical model $\mathcal{G}$ which, among other things, specifies a set of *causal parents* $\mathrm{Pa}(\mathsf{X})$ of X, the intervention $\mathrm{do}(\mathsf{X} = x)$ yields a new probability distribution $\mathbb{P}'$ such that the conditional probability $\mathbb{P}'^{\mathsf{X}|\mathrm{Pa}(\mathsf{X})}$ becomes the function $\cdot \mapsto \delta_x$, while all other conditional distributions of a "child" conditional on its "parents" according to $\mathcal{G}$ match their counterparts in $\mathbb{P}$ [35, Sec. 1.3.1]

Thus identifying a perfect intervention $\mathrm{do}(\mathsf{X} = x)$ with an option α embodies a collection of assumptions – first, it embodies the assumption that selecting the option α will force future observations of the variable X to take the value x. This information may sometimes (though not always) be available to decision makers. However, the identification of options with perfect interventions also embodies the assumption that selecting the option α will leave all "parental conditionals" with respect to $\mathcal{G}$ unchanged with the exception of $\mathbb{P}'^{\mathsf{X}|\mathrm{Pa}(\mathsf{X})}$. It's harder to see how a decision maker could know this.

It's possible that the decision maker might choose the graph $\mathcal{G}$ carefully just so that these additional conditions hold. If so, the decision maker must know exactly which conditionals are invariant a priori, and the graphical model $\mathcal{G}$ is merely a convenient shorthand for representing this knowledge. Typically a decision maker's prior knowledge will not be so extensive. Furthermore, adapting the graph to the set of options under consideration conflicts with normal practice in causal discovery where the learned graph does not depend on the options under consideration. Common measures of success in causal discovery are the structural intervention distance [38] and the structural hamming distance, neither of which depend on the options under consideration [2, 4, 14, 42, 46]. The idea that the relationships captured by a causal graph are independent of the options under consideration is also defended in Pearl [36].

On the other hand, it is not obvious to us how a decision maker could know a priori that their options correspond to perfect interventions on some unknown "objectively correct" graphical model. Beyond this, there are multiple different ways to influence many variables we are interested in measuring and they cannot all be perfect interventions. Hernán and Taubman [22], Hernán [24] considers 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. These will all plausibly affect an individual's risk of death differently, and so they cannot all be modeled by the same intervention on body mass index. Further, it seems to us (as well as other authors in this exchange [25, 36, 44]) that none of these options stand out as a strong candidate to be identified with a perfect intervention on body mass index. The difficulty is twofold: first, we don't know what makes the "true graph" true, and as a result it is not clear how to decide which option if any should be be identified with perfect interventions on this graph. Second, it is not even obvious which action represents the "canonical" way to alter a person's body mass index so it is not clear how we could verify that any procedure for learning a causal graph actually yields the correct interventions as a result.

Hernán argues that limb removal can be dismissed as an option because it is not interesting from a scientific or public health standpoint. While this is a reasonable contention, demanding that a causal graph correctly represent options that are interesting from a scientific or public health standpoint is a version of the strategy of of choosing the graph so that the options of interest are correctly represented, which as we've mentioned is not the standard practice in causal discovery.

Perhaps in recognition of the fact that many actions are not modeled by perfect interventions, there is a diverse array of structural interventions that can be found in the literature: a non-exhaustive review reveals perfect interventions or "hard" interventions [20, 35, ch. 1], soft interventions [7, 11], general or fat-hand interventions [11, 16, 47] and general interventions with unknown targets [2]. However, as in the case of perfect interventions, none of these generalised families of structural interventions make provisions for learning the option-intervention correspondence.

The assumptions of invariant parental conditional distributions made by structural interventions play a very important role in inference based on graphical causal models. 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. In this paper, we show how similar assumptions can specified and analysed without using the framework of structural interventions.

We investigate whether certain assumptions of symmetry are a viable alternative to structured interventions for this task of using features of observations to forecast consequences of different options. In particular, we are interested in assumptions about the future being like the past. In probability models, an assumption of this type is the assumption of *exchangeability*, which holds that reordering a sequence of observed variables leaves a forecaster with exactly the same prediction problem. This assumption cannot be applied as-is to decision models because, in decision problems, future events are affected by the choices made by the decision maker and therefore cannot be in all respects similar to previously observed events.

We investigate two different assumptions inspired by the idea that "the future is in some sense like the past". 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 [10] 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 3.17). Unlike exchangeability, however, IO contractibility is not an appealing assumption in many data-driven decision problems.

We subsequently explore the assumption of *precedent*, which can be informally stated as the assumption that everything a decision maker can do has been done before. Though precedent is a weak assumption, we show that under some additional side conditions it can support the stronger conclusion of IO contractibility and with it the possibility of estimating an input output map from the given observational data (Theorem 5.6). A key assumption of Theorem 5.6 is that the posterior of the parametrisation of a certain conditional distribution is dominated by the uniform measure. We discuss, speculatively, how this assumption may be related to causal structures, noting similar assumptions that appear in Meek [33] and Janzing [28].

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 [6], 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 3.17. Section 5 explains the assumption of precedent and proves Theorem 5.6 and discusses the interpretation of this assumption and its connection to structural causal models.

1.1 Previous work on symmetries in causal inference

The approach that we take assumes that decision making is the fundamental problem that requires 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* [8, 9, 21]. Lattimore and Rohde [30, 31] 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 [32] 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. Imbens and Rubin [27], Rubin [40] made use of the assumption of models with exchangeable potential outcomes to prove several identification results. Saarela et al. [41], 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. Banerjee et al. [1], Dawid [8], Greenland and Robins [19], Hernán [23], Hernán and Robins [26] 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 [8] and the stable unit treatment value assumption (SUTVA) in [40]. This regularity is similar to the condition of *locality* introduced here.

Theorem 5.6 was inspired by the idea of *causal inference by invariant prediction* [39]. While both the assumptions and the conclusions drawn in that work differ from the assumptions and conclusion of Theorem 5.6, 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 X and Y such that the distribution of Y given X doesn't change after actions are taken. Finally, the variable described in that work as "the environment" is similar to the variable Z in Theorem 5.6 in that neither variable needs to be IID, and both variables are only necessarily of interest in the observation set, and need not be of any interest for the consequences of actions.

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.

2.1 Probability Theory

2.1.1 Measurable spaces

Definition 2.1 (Sigma algebra)**.** Given a set A, a σ-algebra $\mathcal{A}$ is a collection of subsets of A where

- $A \in \mathcal{A}$ and $\emptyset \in \mathcal{A}$
- $B \in \mathcal{A} \implies B^{\complement} \in \mathcal{A}$
- $\mathcal{A}$ 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 σ-algebra $\mathcal{A}$.

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

2.1.1.1 Common σ algebras

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

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

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

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

If A is a separable, completely metrizable topological space, then $(A, \mathcal{B}(A))$ is a *standard measurable set*. All standard measurable sets are isomorphic to either $(\mathbb{R}, B(\mathbb{R}))$ or $(C, \mathcal{P}(C))$ for denumerable C [3, 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 σ-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 $x \in E$, $B \in \mathcal{F}$.

2.1.2.1 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 Ω* 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 Y is higher than X if X is given by application of a function to Y. For example, $\mathsf{Y} \preccurlyeq (\mathsf{W}, \mathsf{Y})$ as Y can be obtained by composing a projection with (W, Y).

Definition 2.15 (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 X with respect to μ, $\mu^{\mathsf{X}} : \mathcal{X} \to [0, 1]$ by $\mu^{\mathsf{X}}(A) := \mu(\mathsf{X}^{-1}(A))$ for any $A \in \mathcal{X}$.

Definition 2.17 (Conditional distribution)**.** Given a probability space $(\mu, \Omega, \mathcal{F})$ and variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, the *conditional distribution* of Y given X is any Markov kernel $\mu^{\mathsf{Y}|\mathsf{X}} : X \rightarrowtriangle Y$ such that

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

Definition 2.18 (Trivial variable)**.** We let $*$ stand for any 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 α.

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 $\mathbb{K} \overset{\mathbb{P}_.^{\mathsf{X}}}{\cong} \mathbb{L}$.

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 $\mathbb{P}_\alpha^{\mathsf{X}}$ 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 [6]. 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 exists, is equivalent to other common definitions todo cite

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 Y is conditionally independent of X given 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 [6] is defined using "non-stochastic variables" on the option set C. For our purposes, it is sufficient to use only the special non-stochastic variable $\mathrm{id}_C : C \to C$.

Our two notions are *global conditional independence* and *uniform conditional independence*. The former can be understood as meaning "conditional independence for every $\alpha \in C$", while the latter means "conditional independence for every $\alpha \in C$ and moreover not dependent on α".

Definition 2.22 (Global conditionally independence)**.** Given a decision model $(\mathbb{P}_\cdot, (\Omega, \mathcal{F}), (C, \mathcal{C}))$ and variables X, Y and Z, Y is globally independent of X given 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 2.23 (Uniform conditional independence)**.** Given a decision model $(\mathbb{P}_\cdot, (\Omega, \mathcal{F}), (C, \mathcal{C}))$ and variables X, Y and 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, as shown by Constantinou and Dawid [6], we can reason with collections of extended conditional independence statements as if they were regular conditional independence statements. In the following rules, ϕ and ξ refer to complementary variables on the set C (see Constantinou and Dawid [6] for details), but for our purposes we only consider the cases where either $\phi = \mathrm{id}_C$ and $\xi = *$ or $\phi = *$ and $\xi = \mathrm{id}_C$, where $*$ is the trivial variable $\cdot \mapsto *$. In the rest of this text, we will omit the trivial variable from extended conditional independence statements.

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

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

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

If we have the extended conditional independence $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_\cdot} \mathrm{id}_C|\mathsf{X}$, 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 α

Definition 2.24 (Uniform conditional distribution)**.** Given a decision model $(\mathbb{P}_\cdot, (\Omega, \mathcal{F}), (C, \mathcal{C}))$ and variables X, Y, if $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_\cdot} \mathrm{id}_C|\mathsf{X}$ then

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

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

3 Input Output contractibility

Suppose a 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 X_i is an individual's body mass index and 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* 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 Y_i is independent of $(\mathsf{X}_j, \mathsf{Y}_j)_{j\neq i}$ conditional on X_i and H; for all i, $\mathsf{Y}_i \perp\!\!\!\perp^{e}_{\mathbb{P}.} (\mathrm{id}_C, \mathsf{X}_{[m+n]\setminus\{i\}}, \mathsf{Y}_{[m+n]\setminus\{i\}})|(\mathsf{X}_i, \mathsf{H})$

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

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

The key assumption is that the same response H is shared by both the observations $(\mathsf{X}_i, \mathsf{Y}_i)$, $i \in [m]$ and the consequences $(\mathsf{X}_j, \mathsf{Y}_j)$, $j > m$.

A famous theorem of de Finetti [10] shows that sequential probability models where each variable in the sequence shares an unknown distribution are equivalent to probability models of *exchangeable* sequences. In this section, we introduce input-output contractibility (Definition 3.7) as an analogue of exchangeability for sequences of pairs, and show that it is equivalent to the assumption discussed here of a shared but unknown response H.

3.1 Conditionally independent and identical responses

We formalise the general kind of model sketched above as a model of sequences of pairs of variables 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 H taking values in $\Delta(Y)^D$ – i.e. in the set of maps from D to probability distributions over Y – such that every output Y_i "responds to" D_i according to the same H. While above we discussed an example where the decision maker has prior knowledge about how to control some of the inputs D_i, this is a separate assumption and is not required by the assumption of CIIR pairs.

We define the following shorthand for a decision model incorporating a sequence of pairs.

Definition 3.1 (Sequential input-output model)**.** A decision model $(\mathbb{P}., (C, \mathcal{C}), (\Omega, \mathcal{F}))$ 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 $(\mathbb{P}., \mathsf{D}, \mathsf{Y})$.

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

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

Definition 3.3 (Weakly data-independent)**.** A sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ is weakly data-independent if $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{D}_{(i,\infty]}|(\mathsf{H}, \mathsf{D}_{[1,i]}, \mathrm{id}_C)$.

3.2 Symmetries of sequential conditional probabilities

Given the previously mentioned sequences D and 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 D and 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 D_ρ similarly. Then symmetry to paired permutations means for all α, ρ

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

This symmetry is conceptually very similar to exchangeability, and as we will show that it implies that the $(\mathsf{D}_i, \mathsf{Y}_i)$ share conditionally independent and identical responses. However, the converse is not true.

Example 3.4. 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 hypotheses H are "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

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

Even though $(\mathsf{D}_1, \mathsf{Y}_1)$ and $(\mathsf{D}_2, \mathsf{Y}_2)$ have a shared unknown response, swapping these pairs leads to a different model. What's going on here is that D_1 and D_2 are offering the DM different evidence about which response H is the true one. If the DM could observe a long enough sequence of pairs then the evidence imparted by the inputs on their own would be screened off, but if the DM is only considering the observation of a single pair then they weigh this evidence heavily in their assessment of the probability of different values of H.

Example 3.4 motivates the weaker symmetry we call *exchange commutativity.* A a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ is exchange commutative if there is some variable W such that the conditional $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ is symmetric to paired swaps of Y and D.

Definition 3.5 (Exchange commutativity)**.** Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ along with some $\mathsf{W} : \Omega \to W$, we say $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *commutes with exchange* over 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 W such that $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ commutes with exchange over W.

A second regularity condition we will impose can be roughly understood as the idea that Y_i doesn't "depend on" D_j for $j \neq i$. As Example 3.4 suggests, this cannot be an assumption that Y_i doesn't depend on D_j unconditionally; D_j could, after all, offer some evidence about the state of the shared response H. Instead, we assume that Y_i doesn't depend on non-corresponding X_j after conditioning on some auxiliary W.

Definition 3.6 (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 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 W such that $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ is local over 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 3.8.

Definition 3.7 (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 W if it is local and commutes with exchange.

Theorem 3.8 (Equality of equally sized subsequence conditionals)**.** *Given a sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *and some* W*,* $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ *is IO contractible over* W *if and only if for all subsequences* $A, B \subset \mathbb{N}$ *with* $|A| = |B|$ *and for every* α

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

Proof. Appendix **??** □

Appendix **??** sets out two additional properties of these symmetries. Example C.1 shows that neither locality nor exchange commutativity is implied by the other, and Example C.2 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 3.17, 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 W.

The proof of the theorem is involved, and can be found in its entirety Appendix D. Here we just introduce enough to explain the key terms in the theorem statement.

Definition 3.9 (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 Y into a table Y^D such that Y_{11}^D is equal to the value of the first Y_i such that $\mathsf{D}_i = 1$, Y_{21}^D is equal to the value of the second such 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 Y^D.

Definition 3.10 (Tabulated conditional distribution)**.** Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ on $(\Omega, \mathcal{F})$, define the *tabulated conditional distribution* $\mathsf{Y}^D : \Omega \to Y^{\mathbb{N} \times D}$ 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 Y^D is equal to the value of Y_k for which the corresponding D_k is the ith 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 [29]. 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 3.11 (Directing random measure)**.** Given a probability set $(\mathbb{P}_C, \Omega, \mathcal{F})$ and a sequence $\mathsf{X} := (\mathsf{X}_i)_{i \in \mathbb{N}}$, the directing random measure of X written $\mathsf{H} : \Omega \to \Delta(X)$ 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 D and Y we define the *directing random conditional* as the directing random measure of the tabulated conditional Y^D interpreted as a sequence of column vectors $((\mathsf{Y}_{1j}^D)_{j \in D}, (\mathsf{Y}_{2j}^D)_{j \in D}, ...)$.

Definition 3.12 (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

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

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 3.13 (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 3.15 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 [40]), but we do not assume that the tabulated conditional 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 3.17.

To prove Lemma 3.15, 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 D_is taking some value $j \in D$ preventing us from filling out the "tabulated conditional" Y^D correctly. We call this side condition *infinite support.*

Definition 3.14 (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 α, w, then we say D is *infinitely supported over* 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 3.15. *Suppose a sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *is given with* D *countable and* D *infinitely supported over* W. *Then for some* W, α, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ *is IO contractible if and only if*

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

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

If: We construct a conditional probability according to Definition 3.10 and verify that it satisfies IO contractibility.

The full proof can be found in Appendix D.1. □

Because the distribution $\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}$ from Lemma 3.15 is row-exchangeable, the limit in the definition of the directing random conditional H exists almost surely (see Lemma D.1). In fact, we do not need the full sequence of pairs (D, Y) to calculate H; any subsequence $A \subset \mathbb{N}$ that satisfies the condition that D_A is infinitely supported over W is sufficient.

Theorem 3.16. *Suppose a sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *is given with* D *countable,* D *infinitely supported over* W *and for some* W, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ *is IO contractible for all* α. *Consider an infinite set* $A \subset \mathbb{N}$, *and let* $\mathsf{D}_A := (\mathsf{D}_i)_{i\in A}$ *and* $\mathsf{Y}_A := (\mathsf{Y}_i)_{i\in A}$ *such that* D_A *is also infinitely supported over* W. *Then* H_A, *the directing random conditional of* $(\mathbb{P}_C, \mathsf{D}_A, \mathsf{Y}_A)$ *is almost surely equal to* H, *the directing random conditional of* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$.

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

The proof is in Appendix D.1. □

We are now ready to state the main result, Theorem 3.17. Assuming a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ (Definition 3.1) with inputs D infinitely supported (Definition 3.14) over some random variable W, $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ is IO contractible over the same 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 H (Definition 3.12) and $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ is weakly data-independent.

Theorem 3.17 (Representation of IO contractible models)**.** *Suppose a sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *with sample space* $(\Omega, \mathcal{F})$ *is given with* D *countable and* D *infinitely supported over* W*. Then the following are equivalent:*

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

Proof. (1) $\implies$ (3): We apply Lemma 3.15 followed by Lemma D.1 followed by Lemma D.2.

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

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

See Appendix D.2 for the full proof. □

The arbitrary conditioning variable W presents some obstacles to interpreting Theorem 3.17; we are not merely looking for IO contractibility, but IO contractibility after conditioning on some W. To help us understand what is going on, we observe that without loss of generality we can consider the conditioning variable to be the directing random conditional H.

Corollary 3.18. *If a sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *has independent and identical responses conditional on some variable* W *and* D *has infinite support over the same* W*, then letting* H *be the directing random conditional with respect to inputs* D *and outputs* Y*, it follows that for for all* i, $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{W}|(\mathsf{D}_i, \mathsf{H}, id_C)$ *and for all* α, i, j, $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}} = \mathbb{P}_\alpha^{\mathsf{Y}_j|\mathsf{D}_j\mathsf{H}}$.

Proof. We have by Theorem 3.17 that $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{W}\mathsf{D}}$ is IO contractible over W. The conclusion follows by applying Theorem 3.17 a second time. □

Building on Corollary 3.18, Theorem 3.19 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 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 $(\mathbb{P}_\cdot, \mathsf{D}_A, \mathsf{Y}_A)$ 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 H.

Theorem 3.19. *A data-independent sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *with directing random conditional* H *and* D *infinitely supported over* 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* D_A *and* D_B *are also infinitely supported over* 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 D.2. □

4 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* [40]. Both assumptions hold that, given a treatment D_i, covariates X_i and an outcome Y_i, there is an unknown but common conditional distribution of Y_i given D_i and 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 3.19, 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 4.1 makes a similar point.

Example 4.1. Suppose we have two sequences of binary pairs $((\mathsf{D}, \mathsf{X}), \mathsf{Y}) := ((\mathsf{D}_i, \mathsf{X}_i), \mathsf{Y}_i)_{i \in \mathbb{N}}$ the D_is represent whether patient i was given a particular medicine. The D_is were assigned uniformly according to some source of randomness for even $i \geq 2$, while what exactly determined the D_j for odd j is not known and is likely to have involved patient or doctor discretion. The X_is are covariates, and the Y_is record binarized outcomes of the treatment. D_0 is up to the decision maker, set deterministically according to $\alpha \in 0, 1$. Within both the even and the odd indices of D both options are taken infinitely often with probability 1.

According to Theorem 3.19, the assumption of conditionally independent and identical responses applied to $((\mathsf{D}, \mathsf{X}), \mathsf{Y})$ implies

$$\begin{aligned}
\mathbb{P}_\alpha^{\mathsf{Y}_0|\mathsf{D}_0\mathsf{X}_0\mathsf{D}_{\text{odds}}\mathsf{X}_{\text{odds}}\mathsf{Y}_{\text{odds}}} &= \mathbb{P}_\alpha^{\mathsf{Y}_0|\mathsf{D}_0\mathsf{D}_{\text{evens}\setminus\{0\}}\mathsf{X}_{\text{evens}\setminus\{0\}}\mathsf{Y}_{\text{evens}\setminus\{0\}}} \\
&= \mathbb{P}_\alpha^{\mathsf{Y}_2|\mathsf{D}_2\mathsf{X}_2\mathsf{X}_{\text{evens}\setminus\{0,2\}}\mathsf{Y}_{\text{evens}\setminus\{0,2\}}} \\
&= \mathbb{P}_\alpha^{\mathsf{Y}_2|\mathsf{D}_2\mathsf{X}_2\mathsf{X}_{\text{odds}}\mathsf{Y}_{\text{odds}}}
\end{aligned}$$

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

- Predicting a held-out experimental outcome from the experimental data
- Predicting a held-out experimental outcome from the observational data
- Predicting the outcome of the decision maker's input from the experimental data
- Predicting the outcome of the decision maker's input from the observational data

But the proposition that these problems are *identical* is hard to swallow: despite the obvious differences in the procedures used to obtain the various sequences of pairs, such an assumption nevertheless holds that these differences cannot possibly lead to any differences between the problems discussed.

In practice, when both experimental and observational data are available, they are *not* assumed to be interchangeable in this sense – in fact, the question of how well the observational data predicts experimental outputs is one of substantial interest Eckles and Bakshy [12], Gordon et al. [17, 18].

5 Precedent

We have suggested that IO contractibility 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 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 E_i are never observed these interchanges do not imply that we would use the same model for different experiments.

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

We can relate the minimum size of the set E of possible values of the latent inputs to the number of different options available to the decision maker. In particular it is always possible to construct a see-do model with latent conditionally independent and identical responses E_i from a see-do model $(\mathbb{P}_., \Omega, \mathcal{F})$ of $(\mathsf{X}_i)_{i\in\mathbb{N}\cup\{c\}}$ where the range E of the variables E_i has size at least equal to the number of linearly independent options and observations.

Definition 5.1 (Dimension)**.** Given a collection of probability distributions $A = \{\mathbb{P}_i | i \in B\}$ on a discrete space X, let $p_i := (\mathbb{P}_i(\{x\}))_{x\in X}$. Then $\dim(A) = \dim(\mathrm{span}(\{p_i | i \in B\}))$.

Theorem 5.2 (construction of latent inputs)**.** *Suppose a decision model $(\mathbb{P}_., C, \Omega)$ and observable variables $\mathsf{X} := (\mathsf{X}_i)_{i\in\mathbb{N}\cup\{c\}}$ with X discrete, $\mathsf{X}_\mathbb{N}$ exchangeable, $\mathsf{X}_\mathbb{N} \perp\!\!\!\perp^e \mathrm{id}_C$ and $\mathsf{X}_c \perp\!\!\!\perp^e \mathsf{X}_\mathbb{N}|(\mathsf{G}, \mathrm{id}_C)$ where G is the directing random measure of $\mathsf{X}_\mathbb{N}$. Let*

$$A_g := \{\mathbb{P}_C^{\mathsf{X}_1\mathsf{G}}(\cdot|g)\} \cup \{\mathbb{P}_\alpha^{\mathsf{X}_c|\mathsf{G}}(\cdot|g)|\alpha \in C\}$$

and take $A := \arg\max_{\{A_g|g\in\Delta(X)\}}(\dim(A_g))$; assume A is a finite set.

Then there exists a sequence $\mathsf{E} := (\mathsf{E}_i)_{i\in\mathbb{N}\cup\{c\}}$ on a refinement Ω' of Ω with $|E| = \dim(A)$ such that $(\mathbb{P}'_., \mathsf{E}, \mathsf{X})$ is IO contractible and for all α, $\mathbb{P}'^{\mathsf{X}}_\alpha = \mathbb{P}^{\mathsf{X}}_\alpha$.

Moreover, for any such sequence, $|E| \geq \dim(A)$.

Proof. See Appendix D.3 □

Example 5.3. More concretely, suppose we have an infinite set of observations $(\mathsf{X}_i)_{i\in\mathbb{N}}$ and one "consequence" X_c. X is binary, and the control we can exert is to choose either $\mathbb{P}^{\mathsf{X}}_0 = \frac{1}{4}\delta_0 + \frac{3}{4}\delta_1$ or $\mathbb{P}^{\mathsf{X}}_1 = \frac{3}{4}\delta_0 + \frac{1}{4}\delta_1$, independent of all other observations. Suppose further that for $i \in \mathbb{N}$, $\mathbb{P}^{\mathsf{X}_i}_\alpha = \delta_1$ independent of all other observations for all $\alpha \in \{0, 1\}$. Then, because the dimension of

$$A = \{\frac{1}{4}\delta_0 + \frac{3}{4}\delta_1, \frac{3}{4}\delta_0 + \frac{1}{4}\delta_1, \delta_1\}$$

is 2, we can consider this model to be IO contractible with inputs $\mathsf{E}_i \in \{0, 1\}$ such that

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

in this simple example, the directing measure G and the directing conditional H are trivial.

A special case of a see-do model with conditionally independent and identical responses is when, among the observations, $\mathbb{P}_\alpha^{\mathsf{E}_1|\mathsf{G}}(\cdot|g)$ has full support almost surely. In such a case, roughly speaking, the consequences of anything the decision maker can do have already been seen. We refer to this as a model in which the decision maker's actions have *precedent.*

Theorem 5.6 shows that a slightly strengthened version of this assumption of precedent can have signigicant implications for a decision maker who wants to infer consequences from their observations. This theorem is motivated by the following example:

Example 5.4. Suppose we have a collection of doctors who each see a series of patients, offer a treatment X_i and report their results Y_i. Each doctor may decide whether or not to prescribe based on any number of unobserved factors, and may offer additional unrecorded treatments, vary in their bedside manner and so forth, and these decisions could be stochastic. The decision maker is *also* a doctor, and is reviewing the data contained in the sequences $(\mathsf{Z}_i, \mathsf{X}_i, \mathsf{Y}_i)_{i\in[n]}$, where Z_i identifies the doctor involved in the ith treatment interaction. The decision maker supposes that whatever overall treatment plan they will adopt (which could and probably does also involve features not listed in this set of variables), the same plan has probably been executed at least sometimes by some of these other doctors. Because the other doctors have some variation in their treatment behaviour, it stands to reason that different doctors making the same prescription decisions should see different results *if, conditional on the prescription, the different treatment plans actually lead to different results.* Conversely, if there is *no* variation in results different doctors obtain, then whether or not treatment occurred is presumably the *only* important feature of any treatment plan.

This story might fail if the doctors all knew exactly the long-run probabilistic outcomes of different treatment plans and coordinated with one another to mask any variation they induced. For example, doctor 1 picks a medium effectiveness unobserved plan 100% of the time, while doctor 2 picks a highly effective unobserved plan 50% of the time and a low effectiveness unobserved plan 50% of the time, leading to the same distribution over outcomes. Approximate coordination is plausible – everyone is likely to be aiming for similar goals, and may therefore make choices that are similarly effective. In order to conclude that the lack of variation between doctors is indicative of the importance of prescription decisions, our decision maker must somehow rule out coordination of this kind.

Note that X_i needn't be limited to a particular treatment; in principle, the decision maker might explore many different candidates for a variable X_i which renders Y_i conditionally independent of Z_i.

Theorem 5.6 establishes formal conditions for the informal deduction described in Example 5.4.

Definition 5.5 (Index notation for discrete conditionals)**.** Given a joint probability distribution μ^{XY} with X and Y discrete, let $\mu_x^y := \mu^{\mathsf{Y}|\mathsf{X}}(\{y\}|x)$ and $\mu_X^Y := (x, y) \mapsto \mu_x^y$

Theorem 5.6 (Latent to observable IO contractibility)**.** *Given a decision model $(\mathbb{P}_\cdot, (C, \mathcal{C}), (\Omega, \mathcal{F})$ and sequences $(\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i)_{i\in\mathbb{N}\cup\{c\}}$, $(\mathsf{Z}_i)_{i\in\mathbb{N}}$ all taking values in discrete sets, suppose among the observations $i \in \mathbb{N}$, the pairs $(\mathsf{E}_i, (\mathsf{X}_i, \mathsf{Y}_i, \mathsf{Z}_i))$ share conditionally independent and identical responses and for all $i \in \mathbb{N} \cup \{c\}$ pairs $(\mathsf{E}_i, (\mathsf{X}_i, \mathsf{Y}_i))$ share conditionally independent and identical responses. Take G to be the directing random conditional of $(\mathbb{P}_\cdot, \mathsf{E}_\mathbb{N}, (\mathsf{X}_i, \mathsf{Y}_i, \mathsf{Z}_i)_{i\in\mathbb{N}})$ and H to be the directing random conditional of $(\mathbb{P}_\cdot, \mathsf{E}_{\mathbb{N}\cup\{c\}}, (\mathsf{X}_i, \mathsf{Y}_i)_{i\in\mathbb{N}\cup\{c\}})$.*

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 Y_i is independent of Z_i conditional on X_i and G. For arbitrary α, $\mathbb{Q}_\alpha \in \Delta(\Omega)$ be the probability measure such that, for all $A \in \mathcal{F}$

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

i.e. $\mathbb{Q}_\alpha$ *is* $\mathbb{P}_\alpha$ *conditioned on* $\mathsf{G}^Y_{XZ} \in I$*.*

Thus $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{Z}_i|(\mathsf{X}_i, \mathrm{id}_C)$*. Suppos for all* α*,* $\mathbb{Q}_\alpha$*-almost all* $z, z' \in Z$*,* $e \in E$*,* $g^E_z \in \Delta(E)$*,* $g^{XY}_{EZ} \in \Delta(X \times Y)^{E\times Z}$*,* $\mathbb{Q}_\alpha$ *satisfies the* dominated posterior *assumption:*

$$\mathbb{Q}_\alpha^{\mathsf{G}^E_{z'}|\mathsf{G}^{XY}_{EZ}\mathsf{G}^E_z}(\cdot|g^{XY}_{EZ}, g^e_z) \ll U_{\Delta(D)}$$

Where $U_{\Delta(D)}$ *is the uniform measure on the* $|D-1|$ *simplex of discrete probability distributions with* $|D|$ *outcomes. Then* $(\mathbb{Q}., \mathsf{X}, \mathsf{Y})$ *is also IO contractible.*

Proof. We show that the assumption of conditional independence imposes a polynomial constraint on G^d_z 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 D.3. □

6 Under what circumstances are latent IO contractible models appropriate?

The crucial assumption in Theorem 5.6 – apart from latent IO contractibility – is the assumption that the distribution of the conditional distribution of the latent variable is dominated by the Lebesgue measure. To see why this is critical, consider that every sequence $(\mathsf{X}_i, \mathsf{Y}_i)_{i\in\mathbb{N}}$ can be transformed to the IO contractible sequence $((\mathsf{X}_i, \mathsf{Y}_i), (\mathsf{X}_i, \mathsf{Y}_i))_{i\in\mathbb{N}}$. Thus, were the dominated posterior assumption not required by Theorem 5.6, *any* nontrivial conditional independence would imply observable IO contractibility. However, the sequence $((\mathsf{X}_i, \mathsf{Y}_i), (\mathsf{X}_i, \mathsf{Y}_i))_{i\in\mathbb{N}}$ does not satisfy the dominated posterior assumption. In particular, if $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{Z}_i|(\mathsf{X}_i, \mathsf{G}^Y_{XZ}, \mathrm{id}_C)$ then fixing $\mathsf{G}^{XY}_z = g^{XY}_z$ for some z implies $\mathsf{G}^{XY}_{z'}$ must be such that $\mathsf{G}^Y_{Xz} = \mathsf{G}^Y_{Xz'}$, a Lebesgue measure 0 event.

If $\mathbb{P}^{\mathsf{G}}_\alpha$ is dominated by the uniform measure on $\Delta(EXYZ)$, then $\mathbb{P}_\alpha^{\mathsf{G}^E_Z|\mathsf{G}^{XY}_{EZ}}(\cdot|g^{XY}_{EZ})$ is dominated by the uniform measure on $\Delta(E)^Z$ for almost all (g^{XY}_{EZ}, g^Z) [3, pg. 155]. However, this is not enough for Theorem 5.6 – we condition on $I \subset \Delta(Y)^{XZ}$, which is a measure 0 event with respect to the uniform measure on $\Delta(EXYZ)$.

In light of this, it would be very useful to extend Theorem 5.6 to an approximate result. Specifically, in the event Y_i is approximately independent of Z_i given X_i and G, under what conditions is Y_i also approximately independent of E_i given X_i and G?

For theorem 5.6 to hold, the latent inputs must support the assumption of a dominated posterior for the conditional G^E_Z, and for an approximate result along the same lines we posit that a stronger requirement of diversity for the posterior over conditional distributions $\{\mathsf{G}^E_z | z \in Z\}$ will be necessary. We don't know in general how these requirements should be understood.

The dominated posterior assumption also has a connection to the theory of causal graphical models. Meek [33] 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 Z_i is an ancestor of E_i and $(\mathsf{E}_i, \mathsf{Z}_i)$ is an ancestor of $(\mathsf{X}_i, \mathsf{Y}_i)$ (so that G^E_Z and G^{XY}_{EZ} are associated with forward edges in the causal model) then the dominated posterior 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 Z_i and 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* [28, 45]. 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 independence $\mathsf{E}_i \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{Z}_i|(\mathsf{G}, \mathrm{id}_C)$ and Z_i is an ancestor of 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 dominated posterior assumption. Note that this (as well as the implication suggested in the previous paragraph) are highly speculative.

7 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?" 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 3.17 nor Theorem 5.6 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 3.17 are unlikely to be compelling in many situations. While Theorem 5.6 may be more broadly plausible, we've identified the "dominated posterior" assumption as a particularly difficult one to evaluate. We've suggested that there might be a connection between this assumption and causal structure assumptions. If this is so, one might also want to ask how often the relevant structural assumptions are transparent to a decision maker.

For any practical inference, a generalisation of Theorem 5.6 to approximate independence is in order. Such a generalisation may bring additional clarity to the dominated posterior 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 5.6 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 5.6 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

[1] 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. 10.1016/bs.hefe.2016.08.005. URL https://www.sciencedirect.com/science/article/pii/S2214658X16300071.

[2] 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.

[3] Erhan Çinlar. *Probability and Stochastics*. Springer, 2011.

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

[5] 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. 10.1017/S0960129518000488.

[6] 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.

[7] 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. 10.1609/aaai.v34i06.6567. URL https://ojs.aaai.org/index.php/AAAI/article/view/6567. Number: 06.

[8] 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.

[9] 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. 10.1002/9781119945710.ch4. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119945710.ch4.

[10] 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. 10.1007/978-1-4612-0919-5_10. URL https://doi.org/10.1007/978-1-4612-0919-5_10.

[11] Frederick Eberhardt and Richard Scheines. Interventions and Causal Inference. *Philos. Sci.*, 74, December 2007. 10.1086/525638.

[12] 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. 10.1080/01621459.2020.1796393. URL https://doi.org/10.1080/01621459.2020.1796393.

[13] 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.

[14] 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.

[15] 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. 10.1016/j.aim.2020.107239. URL https://www.sciencedirect.com/science/article/pii/S0001870820302656.

[16] M. Maria Glymour and Donna Spiegelman. Evaluating Public Health Interventions: 5. Causal Inference in Public Health ResearchDo Sex, Race, and Biological Factors Cause Health Outcomes? *American Journal of Public Health*, 107(1):81–85, January 2017. ISSN 0090-0036. 10.2105/AJPH.2016.303539. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5308179/.

[17] 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.

[18] 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.

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

[20] 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.

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

[22] 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. 10.1038/ijo.2008.82. URL https://www.nature.com/articles/ijo200882.

[23] 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. 10.1177/0962280211398037. URL https://doi.org/10.1177/0962280211398037.

[24] 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. 10.1016/j.annepidem.2016.08.016. URL http://www.sciencedirect.com/science/article/pii/S1047279716302800.

[25] 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. 10.1093/aje/kwp293. URL https://academic.oup.com/aje/article/170/8/959/145135. Publisher: Oxford Academic.

[26] 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. 10.1136/jech.2004.029496. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2652882/.

[27] 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. 10.1017/CBO9781139025751. URL https://www.cambridge.org/core/books/causal-inference-for-statistics-social-and-biomedical-sciences/71126BE90C58F1A431FE9B2DD07938AB.

[28] 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. 10.1515/jci-2021-0022. URL https://www.degruyter.com/document/doi/10.1515/jci-2021-0022/html. Publisher: De Gruyter.

[29] 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. 10.1007/0-387-28861-9_2. URL https://doi.org/10.1007/0-387-28861-9_2.

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

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

[32] 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.

[33] 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.

[34] 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.

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

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

[37] 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.

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

[39] 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. 10.1111/rssb.12167. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12167.

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

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

[42] 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].

[43] 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. 10.1007/978-3-642-12821-9_4. URL https://doi.org/10.1007/978-3-642-12821-9_4.

[44] 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. 10.1093/aje/kwp292.

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

[46] 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].

[47] 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.

A Appendix

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 B.

B String Diagrams

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

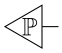
A coherence theorem exists for string diagrams and Markov categories. Applying planar deformation or any of the commutative comonoid axioms to a string diagram yields an equivalent string diagram. The coherence theorem establishes that any proof constructed using string diagrams in this manner corresponds to a proof in any Markov category [43]. More comprehensive introductions to Markov categories can be found in Cho and Jacobs [5], Fritz [15].

B.1 Elements of string diagrams

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

$$\mathbb{K} := \;-\!\boxed{\mathbb{K}}\!-$$

$\mu :=$

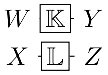


Given two Markov kernels $\mathbb{L} : X \rightharpoonup Y$ and $\mathbb{M} : Y \rightharpoonup Z$, the product $\mathbb{L}\mathbb{M}$ is represented by drawing them side by side and joining their wires:

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

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

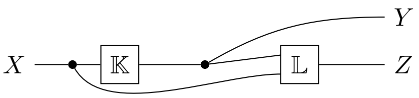
$\mathbb{K} \otimes \mathbb{L} :=$



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

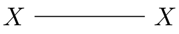
$$\mathbb{K} \odot \mathbb{L} := \mathrm{Copy}_X(\mathbb{K} \otimes \mathrm{id}_X)(\mathrm{Copy}_Y \otimes \mathrm{id}_X)(\mathrm{id}_Y \otimes \mathbb{L})$$

$=$



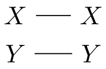
A space X is identified with the identity kernel $\mathrm{id}^X : X \to \Delta(\mathcal{X})$. A bare wire represents the identity kernel:

$\mathrm{Id}^X :=$

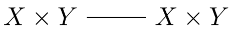


Product spaces $X \times Y$ are identified with tensor product of identity kernels $\mathrm{id}^X \otimes \mathrm{id}^Y$. These can be represented either by two parallel wires or by a single wire representing the identity on the product space $X \times Y$:

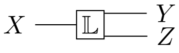
$X \times Y \cong \mathrm{Id}^X \otimes \mathrm{Id}^Y :=$



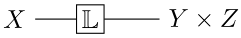
$=$



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



$\equiv$

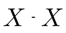


We read diagrams from left to right (this is somewhat different to Cho and Jacobs [5], Fong [13], Fritz [15] but in line with Selinger [43]), and any diagram describes a set of nested products and tensor products of Markov kernels. There are a collection of special Markov kernels for which we can replace the generic "box" of a Markov kernel with a diagrammatic elements that are visually suggestive of what these kernels accomplish.

B.2 Special maps

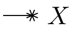
Definition B.1 (Identity map). The identity map $\mathrm{Id}_X : X \rightarrow X$ defined by $(\mathrm{id}_X)(A|x) = \delta_x(A)$ for all $x \in X$, $A \in \mathcal{X}$, is represented by a bare line.

$\mathrm{id}_X :=$



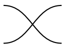
Definition B.2 (Erase map). Given some 1-element set $\{*\}$, the erase map $\mathrm{Del}_X : X \rightarrow \{*\}$ is defined by $(\mathrm{Del}_X)(*|x) = 1$ for all $x \in X$. It "discards the input". It looks like a lit fuse:

$\mathrm{Del}_X :=$



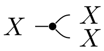
Definition B.3 (Swap map)**.** The swap map $\mathrm{Swap}_{X,Y} : X \times Y \rightarrowtriangle Y \times X$ is defined by $(\mathrm{Swap}_{X,Y})(A \times B|x,y) = \delta_x(B)\delta_y(A)$ for $(x,y) \in X \times Y$, $A \in \mathcal{X}$ and $B \in \mathcal{Y}$. It swaps two inputs and is represented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

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



Definition B.4 (Copy map)**.** The copy map $\mathrm{Copy}_X : X \rightarrowtriangle X \times X$ 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:

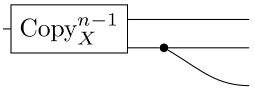
$$\mathrm{Copy}_X :=$$



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

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

$$\mathrm{Copy}^n_X = \qquad\qquad n > 1$$

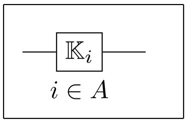


B.2.0.1 Plates

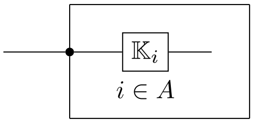
In a string diagram, a plate that is annotated $i \in A$ means the tensor product of the $|A|$ elements that appear inside the plate. A wire crossing from outside a plate boundary to the inside of a plate indicates an $|A|$-fold copy map, which we indicate by placing a dot on the plate boundary. For 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 $i \in A$,

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

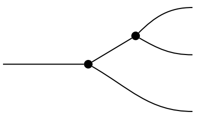


$$\mathrm{Copy}^{|A|}_X(\bigotimes_{i \in A} \mathbb{K}_i) \qquad :=$$

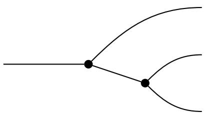


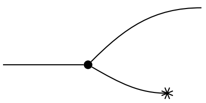
B.3 Commutative comonoid axioms

Diagrams in Markov categories satisfy the commutative comonoid axioms.

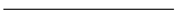


$$= \tag{1}$$

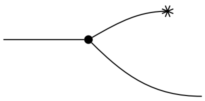


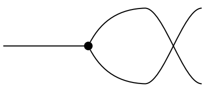


$$=$$

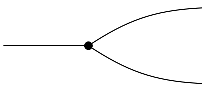


$$= \tag{2}$$

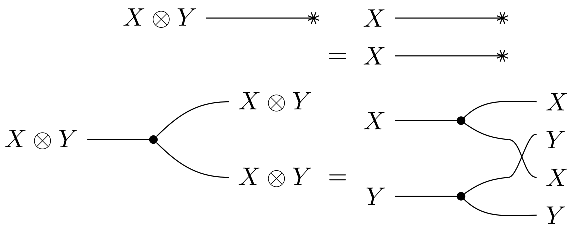




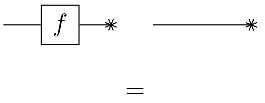
$$=$$



as well as compatibility with the monoidal structure



and the naturality of *Del*, which means that

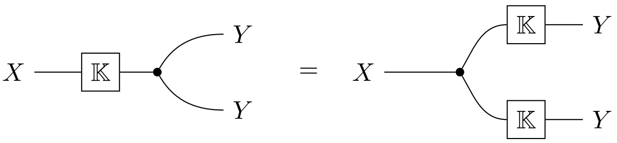


(3)

B.4 Manipulating String Diagrams

Planar deformations along with the applications of Equations (1) through to Equation (3) are almost the only rules we have for transforming one string diagram into an equivalent one. One further rule is given by Theorem B.6.

Theorem B.6 (Copy map commutes for deterministic kernels [13]). *For* $\mathbb{K} : X \rightarrowtriangle Y$

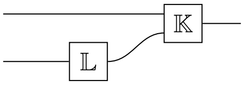


holds iff $\mathbb{K}$ *is deterministic.*

B.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 $\mathbb{K} : X \times Y \rightarrowtriangle Z$, $\mathbb{L} : W \rightarrowtriangle Y$



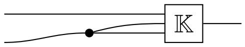
$$= (\mathrm{id}_X \otimes \mathbb{L})\mathbb{K}$$

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

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

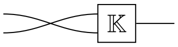
That is, an identity map “passes its input directly to the next kernel”.

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



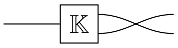
$$
\begin{aligned}
&= (\mathrm{id}_X \otimes \mathrm{Copy}_Y)\mathbb{K} \\
[(\mathrm{id}_X \otimes \mathrm{Copy}_Y)\mathbb{K}](A|x,y) &= \int_Y \int_Y \mathbb{K}(A|x,y',y'')\delta_y(\mathrm{d}y')\delta_y(\mathrm{d}y'') \\
&= \mathbb{K}(A|x,y,y)
\end{aligned}
$$

That is, the copy map “passes along two copies of its input” to the next kernel in the product.
For arbitrary $\mathbb{K} : X \times Y \rightarrow Z$



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

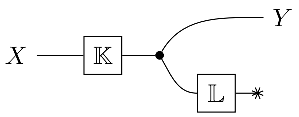
The swap map before a kernel switches the input arguments.
For arbitrary $\mathbb{K} : X \rightarrow Y \times Z$



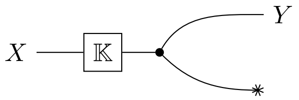
$$
\begin{aligned}
&= \mathbb{K}\mathrm{Swap}_{YZ} \\
(\mathbb{K}\mathrm{Swap}_{YZ})(A \times B|x) &= \int_{Y\times Z} \delta_y(B)\delta_z(A)\mathbb{K}(\mathrm{d}y \times \mathrm{d}z|x) \\
&= \int_{B\times A} \mathbb{K}(\mathrm{d}y \times \mathrm{d}z|x) \\
&= \mathbb{K}(B \times A|x)
\end{aligned}
$$

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

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

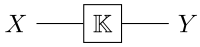


$=$



by Eq. (3)

$=$



by Eq. (2)

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

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

C Symmetries of conditional probabilities

Example C.1 shows that neither locality nor exchange commutativity is implied by the other.

Example C.1. 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 ρ

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

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 C.2.

Example C.2. Consider an experiment where I first flip a coin and record the results of this flip as the outcome Y_1 of "step 1". Subsequently, I can either copy the outcome from step 1 to the result for "step 2" (this is the input $\mathsf{D}_1 = 0$), or flip a second coin use this as the input for step 2 (this is the input $\mathsf{D}_1 = 1$). D_2 is an arbitrary single-valued variable. Then for all d_1, d_2

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

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

D Representation of IO contractible models

This is the proof of Lemmas 3.15 and D.1 and Theorem 3.16. The following definitions are reproduced for convenience. Note that these proofs use the string diagram notation explained in Appendix B.

Definition 3.9. *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 3.10. *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.*

The proof of the theorem follows.

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

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

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

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

$$\begin{aligned}\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{WD}_{\mathsf{R}-1}}(A|w,d) &= \mathbb{P}_\alpha^{\mathsf{Y}_{\mathsf{R}-1}|\mathsf{WD}_{\mathsf{R}-1}}(A|w,d)\\ &\overset{\mathbb{P}_\cdot}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}_{\mathsf{R}-1}|\mathsf{WD}_{\mathsf{R}-1}}(A|w,e)\\ &= \mathbb{P}_\alpha^{\mathsf{Y}^D}(A|w) \end{aligned} \tag{4}$$

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

Now,

$$\mathbb{P}_\alpha^{\mathsf{Y}_{\mathsf{R}-1}|\mathsf{WD}_{\mathsf{R}-1}}(A|w,d) = \int\limits_R \mathbb{P}_\alpha^{\mathsf{Y}_\rho|\mathsf{WD}_\rho}(A|d)\mathbb{P}_\alpha^{\mathsf{R}^{-1}|\mathsf{WD}_{\mathsf{R}-1}}(\mathrm{d}\rho|w,d) \tag{5}$$

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

$$\begin{aligned}\mathbb{P}^{\mathsf{Y}_{\rho^n([n])}|\mathsf{WD}_{\rho^n([n])}} &= \mathbb{P}^{\mathsf{Y}_{\rho([n])}|\mathsf{WD}_{\rho([n])}}\\ &= \mathbb{P}^{\mathsf{Y}_{[n]}|\mathsf{WD}_{[n]}}\end{aligned}$$

By Corollary **??**, it must therefore be the case that

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

Then from Equation (5)

$$\begin{aligned}\mathbb{P}_\alpha^{\mathsf{Y}_{\mathsf{R}-1}|\mathsf{WD}_{\mathsf{R}-1}}(A|w,d) &\overset{\mathbb{P}_\cdot}{\cong} \int\limits_R \mathbb{P}_\alpha^{\mathsf{Y}_\rho|\mathsf{WD}_\rho}(A|d)\mathbb{P}_\alpha^{\mathsf{R}^{-1}|\mathsf{WD}_{\mathsf{R}-1}}(\mathrm{d}\rho|w,d)\\ &\overset{\mathbb{P}_\cdot}{\cong} \int\limits_R \mathbb{P}_\cdot^{\mathsf{Y}|\mathsf{WD}}(A|w,d)\mathbb{P}_\alpha^{\mathsf{R}^{-1}|\mathsf{WD}_{\mathsf{R}-1}}(\mathrm{d}\rho|w,d)\\ &\overset{\mathbb{P}_\cdot}{\cong} \mathbb{P}_\cdot^{\mathsf{Y}|\mathsf{WD}}(A|w,d)\end{aligned} \tag{6}$$

for all $i, j \in \mathbb{N}$. Then by Equation (4) and Equation (6)

$$\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}(A|w) = \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}(A|w,e) \tag{7}$$

Take some $d \in D^{\mathbb{N}}$. From Equation (7) and IO contractibility of $\mathbb{P}_\cdot^{\mathsf{Y}|\mathsf{WD}}(A|e)$,

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

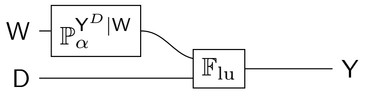
It remains to be shown that Y^D is invariant to finite permutations within rows. Consider some finite permutation within columns $\eta : \mathbb{N} \times D \to \mathbb{N} \times D$, note that $e_{\eta(i,j)} = j$ and hence $(e_{\eta(i,j)})_{i\in\mathbb{N},j\in D} = e$. Thus

$$\begin{aligned}
\mathbb{P}_\alpha^{(\mathsf{Y}^D_{\eta(i,j)})_{\mathbb{N}\times D}|\mathsf{W}}(A|w) &= \mathbb{P}_\alpha^{(\mathsf{Y}^D)_{\mathbb{N}\times D}|\mathsf{W}}\mathrm{Swap}_\eta(A|w) \\
&= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}\mathrm{Swap}_\eta(A|w,e) && \text{from Eq. (7)} \\
&= \mathbb{P}_\alpha^{\mathsf{Y}_\eta|\mathsf{WD}}(A|w,e) \\
&= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}_{\eta^{-1}}}(A|w,e) && \text{by exchange commutativity} \\
&= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}(A|w,(e_{\eta^{-1}(i,j)})_{i\in\mathbb{N},j\in D}) \\
&= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}(A|w,e) \\
&= \mathbb{P}_\alpha^{(\mathsf{Y}^D_{ij})_{\mathbb{N}\times D}|\mathsf{W}}(A|w) && \text{from Eq. (7)}
\end{aligned}$$

If: We construct a conditional probability according to Definition 3.10 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 (**??**).

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

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

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

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

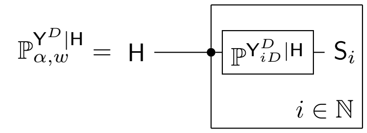
so $\mathbb{K}$ is IO contractible by Theorem 3.8. □

As a consequence of Lemma 3.15 along with De Finetti's representation theorem, we can say that given $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ IO contractible, conditioning on H renders the columns of Y^D independent and identically distributed.

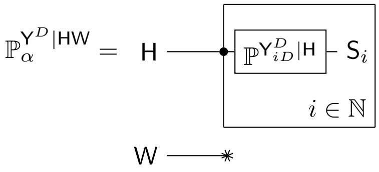
Lemma D.1. *Suppose a sequential input-output model* $(\mathbb{P}_{\cdot}, \mathsf{D}, \mathsf{Y})$ *is given with* D *countable,* D *infinitely supported over some* W *and* $(\mathbb{P}_{\cdot}, \mathsf{D}, \mathsf{Y})$ *IO contractible over the same* W*. Then, letting* H *be the directing random conditional of* $(\mathbb{P}_{\cdot}, \mathsf{D}, \mathsf{Y})$ *(Definition 3.12) and* $\mathsf{Y}^D_{iD} := (\mathsf{Y}^D_{ij})_{j\in D}$*, we have for all* $i \in \mathbb{N}$*,* $\mathsf{Y}^D_{iD} \perp\!\!\!\perp^e_{\mathbb{P}_{\cdot}} (\mathsf{Y}^D_{\mathbb{N}\setminus\{i\}D}, \mathsf{W}, id_C)|\mathsf{H}$ *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 3.15, 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 [29, Prop. 1.4] (where H is precisely what Kallenberg calls the directing random measure)


$\mathbb{P}^{\mathsf{Y}^D|\mathsf{H}}_{\alpha,w}$ =
H
$\mathbb{P}^{\mathsf{Y}^D_{iD}|\mathsf{H}}$
S_i
$i \in \mathbb{N}$

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


$\mathbb{P}^{\mathsf{Y}^D|\mathsf{HW}}_{\alpha}$ =
H
$\mathbb{P}^{\mathsf{Y}^D_{iD}|\mathsf{H}}$
S_i
$i \in \mathbb{N}$
W

and because it also does not depend on α we have $\mathsf{Y}^D \perp\!\!\!\perp^e_{\mathbb{P}_{\cdot}} (\mathsf{W}, \mathrm{id}_C)|\mathsf{H}$. Further application of Kallenberg [29, 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 α, 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 3.16. *Suppose a sequential input-output model* $(\mathbb{P}_{\cdot}, \mathsf{D}, \mathsf{Y})$ *is given with* D *countable,* D *infinitely supported and for some* W*,* $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ *is IO contractible for all* α*. Consider an infinite set* $A \subset \mathbb{N}$*, and let* $\mathsf{D}_A := (\mathsf{D}_i)_{i\in A}$ *and* $\mathsf{Y}_A := (\mathsf{Y}_i)_{i\in A}$*. Then* H_A*, the directing random conditional of* $(\mathbb{P}_{\cdot}, \mathsf{D}_A, \mathsf{Y}_A)$ *is almost surely equal to* 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 Y^D. Then we show that the contracted version of Y^D has the same distribution as the original, and consequently the normalised partial sums converge to the same limit.

Define $\mathsf{Y}^{D,A}$ as the tabulated conditional of $(\mathsf{D}_A, \mathsf{Y}_A)$, i.e. let $\#^{A,k}_j$ 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

$$\mathsf{Y}^{D,A}_{ij} = \mathsf{Y}^D_{\mathsf{Q}(i)j} \tag{8}$$

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

Equation (8) is what is meant by "the subsequence $(\mathsf{D}_A, \mathsf{Y}_A)$ induces a random contraction over the rows of Y^D". We will now show that $\mathsf{Y}^{D,A}$ has the same distribution as Y^D.

Let $\mathrm{con}_q : Y^{\mathbb{N}\times D} \to Y^{\mathbb{N}\times D}$ 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:

$$\begin{aligned}
\mathbb{P}_\alpha^{\mathsf{Y}^{D,A}|\mathsf{WQ}}(B|w,q) &= \mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}\mathrm{con}_q(B|w) \\
&= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}\mathrm{con}_q(B|w,e) && \text{by Eq.(7)} \\
&= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}(B|w,e) && \text{by Theorem 3.8} \\
&= \mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}(B|w) && \text{by Eq.(7)} \qquad (9)
\end{aligned}$$

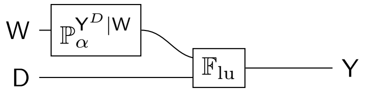
Finally, take H_A the directing random measure of $\mathsf{Y}^{D,A}$. We conclude from the equality Eq. (9) and from the fact that there is a one-to-one map from directing random measures to exchangeable distributions that $\mathsf{H}_A \overset{\mathbb{P}_\alpha}{\cong} \mathsf{H}$. □

D.1 IO contractibility proofs

The following is a technical lemma that will be used in Theorem 3.17.

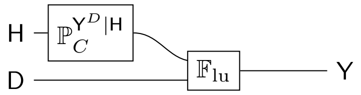
Lemma D.2. *Suppose a sequential input-output model* $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ *is given with* D *countable,* D *infinitely supported over* W*, for some* W*,* $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ *is IO contractible for all* α *and for all* α

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


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

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

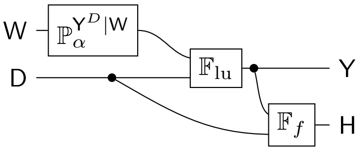
$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{HD}} =$


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 Y and D to H also maps Y^D and the constant $e \in D^{\mathbb{N}}$ to H' with $\mathsf{H}' \overset{\mathbb{P}_\cdot}{\cong} \mathsf{H}$, and the result follows from disintegration along with a conditional independence given by Lemma 3.15.

Y^D is a function of Y and D (see Definition 3.10) and H is a function of Y^D. Say $f : Y \times D \to H$ is such that $\mathsf{H} = f(\mathsf{Y}, \mathsf{D})$ (see Definition 3.11). 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 ith coordinate of d equal to $j \in D$ and $\#_{[d=j]_i}$ to be the position in d of $[d = j]_i$. Concretely, f is given by

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

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

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

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 Y^D (Lemma 3.15), 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 D.1, the sequence $(\mathsf{Y}^D_{iD})_{i\in\mathbb{N}}$ are mutually independent conditional on H and for all α, $\mathbb{P}_\alpha^{\mathsf{Y}_{iD}|\mathsf{H}}(A|\nu) \overset{\mathbb{P}_C}{\cong} \nu(A)$. Thus, by the law of large numbers

$$\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{H}}(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\prod_{j\in D} A_j}(y^D_{i,D}) \bowtie \nu(A)|\nu) = 1$$

which implies

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

Because this holds for all A,

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

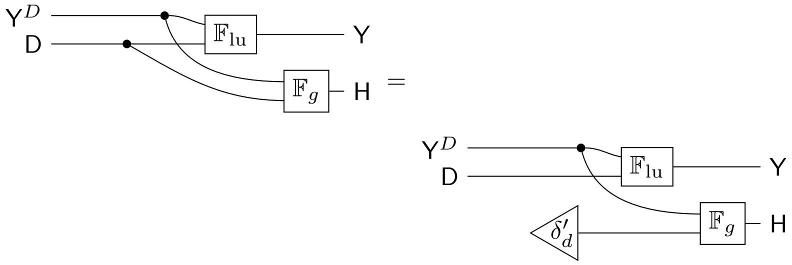
And, as a consequence, defining

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

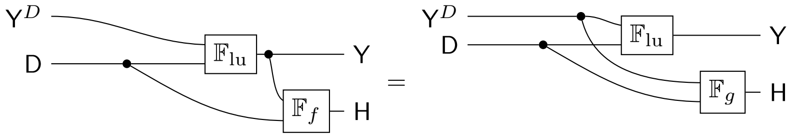
we have

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

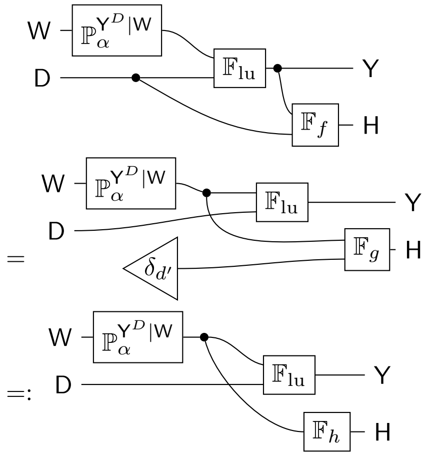
which in turn implies the almost sure equality of the associated Markov kernels:



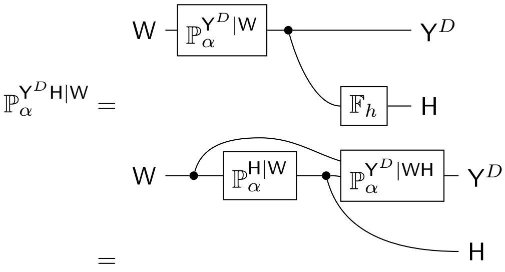
but we also have, by the definitions of f and g



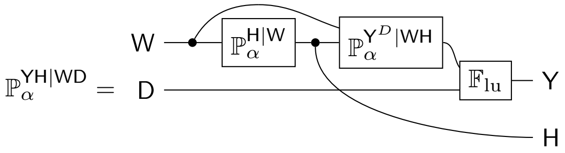
finally



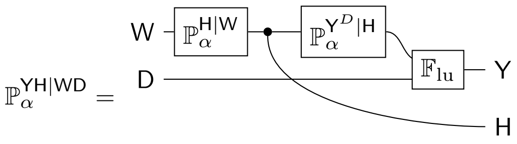
Noting that $\mathbb{F}_h \otimes \mathrm{Del}_W = \mathbb{P}_\alpha^{\mathsf{H}|\mathsf{Y}^D\mathsf{W}}$


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

and so

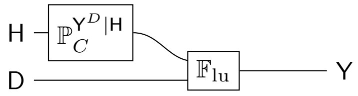

$\mathbb{P}_\alpha^{\mathsf{YH}|\mathsf{WD}} =$
W
$\mathbb{P}_\alpha^{\mathsf{H}|\mathsf{W}}$
$\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{WH}}$
D
$\mathbb{F}_{\text{lu}}$
Y
H

From Lemma 3.15 we also have $\mathsf{Y}^D \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{W}, \mathrm{id}_C)|\mathsf{H}$, so


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

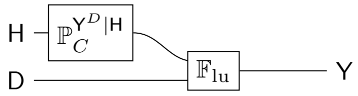
and so by higher order conditionals $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{W}|(\mathsf{H}, \mathsf{D}, \mathrm{id}_C)$ and

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


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

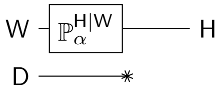
Because the right hand side does not depend on α, we finally have $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{W}, \mathrm{id}_C)|(\mathsf{H}, \mathsf{D})$ and the result

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


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

Furthermore, by marginalising the right hand side of Equation D.1 we have

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



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

D.2 Data-independent models

Theorem 3.17 says that a data independent sequential input-output model $(\mathbb{P}_., \mathsf{D}, \mathsf{Y})$ features conditionally independent and identical response functions $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{HD}_i}$ for all α if and only if there is some W such that $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ is IO contractible over W for all α. The variable W is something of a nuisance; rather than thinking only about whether IO contractibility holds, we must consider whether there's *any* variable that licenses the assumption of IO contractibility.

A simple special case to consider is when W is single valued – that is, when $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}}$ is IO contractible. As Theorem D.3 shows, this corresponds to the CIIR sequence models where the inputs D are unconditionally data-independent and independent of the hypothesis H. We can also consider the case where $(\mathbb{P}_., \mathsf{D}, \mathsf{Y})$ is only exchange commutative over $*$. This corresponds to models where the inputs D are data-independent and the hypothesis H depends on a symmetric function of the inputs D (under some side conditions).

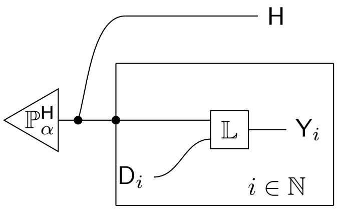
Theorem D.3 (Data-independent IO contractibility)**.** *Suppose a sequential input-output model* $(\mathbb{P}_., \mathsf{D}, \mathsf{Y})$ *with sample space* $(\Omega, \mathcal{F})$ *is given with* D *countable and, letting* $E \subset D^{\mathbb{N}}$ *be the set of all sequences for which each* $j \in D$ *occurs infinitely often,* $\mathbb{P}_\alpha^{\mathsf{D}}(E) = 1$ *for all* α*. Then the following are equivalent:*

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

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

, $\mathsf{H} \perp\!\!\!\perp^e_{\mathbb{P}_.} \mathsf{D}|id_C$ *and for all* i $\mathsf{D}_i \perp\!\!\!\perp^e_{\mathbb{P}_.} \mathsf{D}_{(i,\infty]})|(\mathsf{D}_{[1,i)}, id_C)$
3. *There is some* $\mathbb{L} : H \times X \rightarrowtriangle Y$ *such that for all* α*,*

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



Proof. See Appendix D.2. □

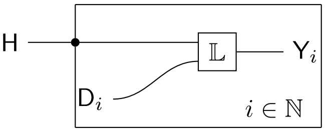
While $\mathbb{P}_\cdot^{\mathsf{Y}|\mathsf{D}}$ exchange commutative is not necessarily IO contractible, exchange commutativity of this conditional implies IO contractibility over the directing random conditional H, and thus is sufficient for conditionally independent and identical responses.

Theorem D.4. *If $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}}$ is exchange commutative, and for each α $\mathbb{P}_\alpha^{\mathsf{D}}$ is absolutely continuous with respect to some exchangeable distribution $\mathbb{Q}_\alpha^{\mathsf{D}}$ in $\Delta(D^{\mathbb{N}})$ with directing random measure F and D infinitely supported over F with respect to $\mathbb{Q}_\alpha$, then $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{HD}}$ is IO contractible, where H is the directing random conditional for $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}}$.*

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

Corollary D.5. *If $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ is exchange commutative over $*$, and for each α $\mathbb{P}_\alpha^{\mathsf{D}}$ is absolutely continuous with respect to some exchangeable distribution in $\Delta(D^{\mathbb{N}})$ then*

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



Proof. By Theorem D.4, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ is IO contractible over some W for all α, so the result follows immediately from Theorem 3.17. □

D.3 Precedent

Theorem 5.2. *Suppose a decision model $(\mathbb{P}_\cdot, C, \Omega)$ and observable variables $\mathsf{X} := (\mathsf{X}_i)_{i\in\mathbb{N}\cup\{c\}}$ with X discrete, $\mathsf{X}_{\mathbb{N}}$ exchangeable, $\mathsf{X}_{\mathbb{N}} \perp\!\!\!\perp^e \mathrm{id}_C$ and $\mathsf{X}_c \perp\!\!\!\perp^e \mathsf{X}_{\mathbb{N}}|(\mathsf{G}, \mathrm{id}_C)$ where G is the directing random measure of $\mathsf{X}_{\mathbb{N}}$. Let*

$$A_g := \{\mathbb{P}_C^{\mathsf{X}_1\mathsf{G}}(\cdot|g)\} \cup \{\mathbb{P}_\alpha^{\mathsf{X}_c|\mathsf{G}}(\cdot|g)|\alpha \in C\}$$

and take $A := \arg\max_{\{A_g|g\in\Delta(X)\}}(\dim(A_g))$; assume A is a finite set.

Then there exists a sequence $\mathsf{E} := (\mathsf{E}_i)_{i\in\mathbb{N}\cup\{c\}}$ on a refinement Ω' of Ω with $|E| = \dim(A)$ such that $(\mathbb{P}'_\cdot, \mathsf{E}, \mathsf{X})$ is IO contractible and for all α, $\mathbb{P}_\alpha^{\prime\mathsf{X}} = \mathbb{P}_\alpha^{\mathsf{X}}$.

Moreover, for any such sequence, $|E| \geq \dim(A)$.

Proof. Take $E = [dim(A)]$.

Take $B_g := \mathrm{EP}(\mathrm{span}(A_g) \cap \Delta(X))$ where EP is the function that returns the extreeme points of a convex set (note that A_g is the intersection of two convex sets). $(b_{ig})_{i\in|B_g|}$ some enumeration of the elements of B_g. For each $g \in \Delta(X)$, $i \in \mathbb{N} \cup \{c\}$ set $\mathbb{P}_\alpha^{\mathsf{X}_i|\mathsf{E}_i\mathsf{G}}(\cdot|e, g) = b_{eg}$.

Define

$$d_{g\alpha} := \begin{cases} \mathbb{P}_C^{\mathsf{X}_1|\mathsf{G}}(\cdot|g) & \alpha \notin C \\ \mathbb{P}_\alpha^{\mathsf{X}_c|\mathsf{G}}(\cdot|g) & \alpha \in C \end{cases}$$

By construction, for each $d_{g\alpha}$ there is some set of weights $\{0 \leq w^e g\alpha | e \in E\}$, $\sum_{e\in E} w_{g\alpha}^e = 1$ such that

$$d_{g\alpha} = \sum_{e\in E} w_{g\alpha}^e b_{eg}$$

Let o be some index that is not an element of C. Set

$$\mathbb{P}_\alpha^{\prime\mathsf{E}_i|\mathsf{G}}(e|g) = \begin{cases} w_{go}^e & i \in \mathbb{N} \\ w_{g\alpha}^e & i = c \end{cases}$$

Then $i \in \mathbb{N}$ implies

$$\begin{aligned}\sum_{e\in E} \mathbb{P}_\alpha^{\prime \mathsf{X}_i|\mathsf{E}_i\mathsf{G}}(\{x\}|e,g)\mathbb{P}_\alpha^{\prime \mathsf{E}_1|\mathsf{G}}(\{e\}|g) &= \sum_{e\in E} b_{eg}^x w_{go}^e \\ &= \mathbb{P}_\alpha^{\mathsf{X}_i|\mathsf{G}}(\{x\}|g)\end{aligned}$$

and

$$\begin{aligned}\sum_{e\in E} \mathbb{P}_\alpha^{\mathsf{X}_c|\mathsf{E}_c\mathsf{G}}(J|e,g)\mathbb{P}_\alpha^{\mathsf{E}_c|\mathsf{G}}(\{e\}|g) &= \sum_{e\in E} b_{eg} w_{g\alpha}^e \\ &= \mathbb{P}_\alpha^{\mathsf{X}_c|\mathsf{G}}(\{x\}|g)\end{aligned}$$

Finally, choose $\mathbb{P}_\alpha^{\prime \mathsf{E}|\mathsf{X}\mathsf{G}}$ such that $\mathsf{E} \perp\!\!\!\perp^e \mathsf{X}|(\mathsf{G},\mathrm{id}_C)$ and $\mathbb{P}_\alpha^{\prime \mathsf{G}} = \mathbb{P}_\alpha^{\mathsf{G}}$.
Then

$$\begin{aligned}\mathbb{P}_\alpha^{\prime \mathsf{X}}(\{x_{\mathbb{N}\cup\{c\}}\}) &= \int_G \sum_{e\in E} \mathbb{P}_\alpha^{\prime \mathsf{E}_i|\mathsf{G}}(\{e\}|g)\mathbb{P}_\alpha^{\prime \mathsf{X}_i|\mathsf{E}_i\mathsf{G}}(\{x_i\}|e,g)\mathbb{P}_\alpha^{\prime \mathsf{G}}(\mathrm{d}g) \\ &= \int_G \mathbb{P}_\alpha^{\mathsf{X}_i|\mathsf{G}}(\{x_i\}|g)\mathbb{P}_\alpha^{\mathsf{G}}(\mathrm{d}g) \\ &= \mathbb{P}_\alpha^{\mathsf{X}}\end{aligned}$$

□

Theorem 5.6. *Given a decision model* $(\mathbb{P}_., (C,\mathcal{C}), (\Omega,\mathcal{F})$ *and sequences* $(\mathsf{E}_i,\mathsf{X}_i,\mathsf{Y}_i)_{i\in\mathbb{N}\cup\{c\}}$, $(\mathsf{Z}_i)_{i\in\mathbb{N}}$ *all taking values in discrete sets, suppose the observations* $(\mathsf{E}_i,\mathsf{X}_i,\mathsf{Y}_i,\mathsf{Z}_i)_{i\in\mathbb{N}}$ *are exchangeable for all options* $\alpha \in C$ *and pairs* $(\mathsf{E}_i,(\mathsf{X}_i,\mathsf{Y}_i))$ *share conditionally independent responses. Take* G *to be the directing random measure of the observations* $\mathbb{P}_\alpha^{(\mathsf{E}_i,\mathsf{X}_i,\mathsf{Y}_i,\mathsf{Z}_i)_{i\in\mathbb{N}}}$ *and* H *to be the directing random conditional of* $(\mathbb{P}_., \mathsf{E}_i, (\mathsf{X}_i,\mathsf{Y}_i))$.

Fix some $g_{EZ}^{XY} \in \Delta(X \times Y)^{EZ}$, *and let* $\mathbb{Q}_\alpha := \mathbb{P}_\alpha^{\mathrm{id}_\Omega|\mathsf{G}_{EZ}^{XY}}(\cdot|g_{EZ}^{XY})$ – *i.e.* $\mathbb{Q}_\alpha$ *is* $\mathbb{P}_\alpha$ *conditioned on* $\mathsf{G}_{EZ}^{XY} = g_{EZ}^{XY}$.

Suppose for any $i \in \mathbb{N}$, $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{Q}}^e \mathsf{Z}_i|(\mathsf{X}_i,\mathsf{G}_{XZ}^Y,\mathrm{id}_C)$ *and for all* α, $\mathbb{Q}_\alpha$*-almost all* $z,z' \in Z$, $e \in E$, $g_z^e \in [0,1]$, $g_{EXZ}^Y \in$ *satisfies the* dominated posterior *assumption:*

$$\mathbb{Q}_\alpha^{\mathsf{G}_{z'}^E|\mathsf{G}_z^E}(\cdot|g_z^e) \ll U_{[0,1]^{|E|}}$$

then $(\mathbb{Q}_., \mathsf{X}, \mathsf{Y})$ *is also IO contractible.*

Proof. For abitrary e, z', fix $\mathsf{G}_{z'}^E$ to some arbitrary $g_{z'}^E$. Note that $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{Q}}^e \mathsf{Z}_i|(\mathsf{X}_i,\mathrm{id}_C)$ implies $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{Q}}^e \mathsf{Z}_i|(\mathsf{X}_i,\mathsf{G}_{z'}^e,\mathrm{id}_C)$. This in turn implies, for all α and $\mathbb{Q}_\alpha$-almost all e,z,x,y

$$\sum_{e\in E} g_{exz}^y \frac{g_{ez}^x \mathsf{G}_z^e}{\sum_{e'\in E} g_{e'z}^x \mathsf{G}_z^{e'}} \overset{\mathbb{P}_C}{\cong} \sum_{e\in E} g_{exz'}^y \frac{g_{ez'}^x g_{z'}^e}{\sum_{e'\in E} g_{e'z'}^x g_{z'}^{e'}} \tag{10}$$

Note that both g_{exz}^y and g_{ez}^x are fixed by the choice of g_{EZ}^{XY}.

For fixed g_{exz}^y, g_{ez}^x and $\mathsf{G}_{z'}^e$, Eq. (10) defines a polynomial constraint on G_z^e. We will show that, unless $g_{exz}^y = g_{e'xz}^y$ for all e,e' and z, that this constraint is nontrivial for some z. Consequently, the set of solutions subject to the restriction $g_{exz}^y \neq g_{e'xz}^y$ has Lebesgue measure 0.

For arbitrary e, assume $g_{exz'}^y > g_{e^<xz'}^y$ for some $e^<$.

Then either $g_{ez'}^x = g_{e^<z'}^x$, $g_{ez'}^x < g_{e^<z'}^x$ or $g_{ez'}^x > g_{e^<z'}^x$. Consider the first case, and take G' such that $\mathsf{G}_{z'}^{\prime\prime e} = 0.5\mathsf{G}_{z'}^e$ and $\mathsf{G}_{z'}^{\prime e^<} = \mathsf{G}_{z'}^{e^<} + 0.5\mathsf{G}_{z'}^e$ and equal to $\mathsf{G}_z^{e''}$ for all other $e'' \in E$. Note that $\mathsf{G}_z^{\prime E}$ is a Markov kernel as it is everywhere positive and sums to 1, and that $\mathsf{G}_{z'}^{\prime\prime e} < \mathsf{G}_{z'}^e$ almost surely as $\mathsf{G}_{z'}^e > 0$ almost surely. Then

$$\begin{aligned}\frac{g_{ez}^x \mathsf{G}_z^e}{\sum_{e'\in E} g_{e'z}^x \mathsf{G}_z^{e'}} &> \frac{g_{ez}^x \mathsf{G}_z^{\prime e}}{\sum_{e'\in E} g_{e'z}^x \mathsf{G}_z^{\prime e'}} \\ \frac{g_{e^<z}^x \mathsf{G}_z^{e^<}}{\sum_{e'\in E} g_{e'z}^x \mathsf{G}_z^{e'}} &< \frac{g_{e^<z}^x \mathsf{G}_z^{\prime e^<}}{\sum_{e'\in E} g_{e'z}^x \mathsf{G}_z^{\prime e'}}\end{aligned}$$

because by assumption the denominator remains the same. But then

$$\sum_{e\in E} g^{y}_{exz} \frac{g^{x}_{ez}\mathsf{G}^{e}_{z}}{\sum_{e'\in E} g^{x}_{e'z}\mathsf{G}^{e'}_{z}} > \sum_{e\in E} g^{y}_{exz'} \frac{g^{x}_{ez}\mathsf{G}'^{e}_{z'}}{\sum_{e'\in E} g^{x}_{e'z'}\mathsf{G}'^{e'}_{z'}} \tag{11}$$

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

$$\frac{g^{x}_{ez}\mathsf{G}^{e}_{z}}{\sum_{e'\in E} g^{x}_{e'z}\mathsf{G}^{e'}_{z}} > \frac{g^{x}_{ez}\mathsf{G}'^{e}_{z}}{\sum_{e'\in E} g^{x}_{e'z}\mathsf{G}'^{e'}_{z}}$$
$$\frac{g^{x}_{e<z}\mathsf{G}^{e^<}_{z}}{\sum_{e'\in E} g^{x}_{e'z}\mathsf{G}^{e'}_{z}} < \frac{g^{x}_{e<z}\mathsf{G}'^{e^<}_{z}}{\sum_{e'\in E} g^{x}_{e'z}\mathsf{G}'^{e'}_{z}}$$

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

$$\frac{g^{x}_{ez}\mathsf{G}'^{e}_{z}}{\sum_{e'\in E} g^{x}_{e'z}\mathsf{G}'^{e'}_{z}} = \frac{0.5 g^{x}_{ez}\mathsf{G}^{e}_{z}}{\sum_{e'\in E} g^{x}_{e'z}\mathsf{G}^{e'}_{z} - 0.5\mathsf{G}^{e}_{z}(g^{x}_{ez} - g^{x}_{e<z})}$$

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

Thus the conclusion in Eq. (11) follows for the same reasons. Considering $g^{x}_{ez'} < g^{x}_{e<z'}$, analogous reasoning implies Eq. (11) once again.

Thus, unless $g^{y}_{exz} = g^{y}_{e'xz}$ for all e, e' and z, Eq. (10) implies a nontrivial constraint on G^{t}_{z} for some z. Thus for some e, e',z, x and y the set of solutions $S := \{g^{E}_{z} | \mathsf{G}^{E}_{z} = g^{E}_{z}$ satisfies Eq. (10) $\wedge\, g^{y}_{exz} \neq g^{y}_{e'xz}\}$ has Lebesgue measure 0 in the set $[0,1]^{E}$ [34], and so by domination

$$\mathbb{Q}_{\alpha}^{\mathsf{G}^{E}_{z}|\mathsf{G}^{E}_{z'}}(S|g^{E}_{z'}) = 0$$

On the other hand, by assumption, the set $T := \{g^{E}_{z} | \mathsf{G}^{E}_{z} = g^{E}_{z}$ satisfies Eq. (10)$\wedge g^{y}_{exz} \neq g^{y}_{e'xz}\}$ has measure 1. Thus we conclude that $g^{y}_{exz} = g^{y}_{e'xz}$ with probability 1. That is, $\mathsf{Y}_i \perp\!\!\!\perp^{e}_{\mathbb{Q}} \mathsf{E}_i | \mathsf{Z}_i, \mathsf{X}_i, \mathrm{id}_C$. By contraction with $\mathsf{Y}_i \perp\!\!\!\perp^{e}_{\mathbb{Q}} \mathsf{Z}_i | (\mathsf{X}_i, \mathrm{id}_C)$, we have $\mathsf{Y}_i \perp\!\!\!\perp^{e}_{\mathbb{Q}} (\mathsf{Z}_i, \mathsf{E}_i) | (\mathsf{X}_i, \mathrm{id}_C)$.

By IO contractibility, we have $\mathbb{Q}_{\alpha}^{\mathsf{Y}_i|\mathsf{E}_i\mathsf{X}_i\mathsf{G}^{Y}_{EX}} = \mathbb{Q}_{\alpha}^{\mathsf{Y}_i|\mathsf{E}_i\mathsf{X}_i\mathsf{H}}$. But then, by independence, $\mathbb{Q}_{\alpha}^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{G}^{Y}_{EX}} = \mathbb{Q}_{\alpha}^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{H}}$. □