Causal Inference Without Interventions

A Preprint

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

Cheng Soon Ong
Data61

Robert C. Williamson
Universität Tübingen

November 27, 2023

Abstract

If we have data and want to evaluate the consequences of an action, we can use a causal model. In this model, actions are represented by structural interventions. However, for many variables the connection between actions and structural interventions is not obvious. This raises the question: if we learn a causal model but do not know how actions correspond to interventions, have we learned anything useful? We show, eventually, that the answer may be "yes". We begin with decision models, which map options to distributions over outcomes but are otherwise agnostic to causality. We first analyse inference under the assumption of conditionally independent and identical responses (CIIR), an assumption of a consistent relationship between pairs of input and output variables. For CIIR models, we prove that arbitrary sequences of input-output pairs can be exchanged given sufficient data. On this basis, we argue for a negative result: it is usually unreasonable to assume causal effects are identified in observational data (this is a common view, but our argument is novel). However, we also show a positive result: precedent is a version of the CIIR assumption in which inputs are unobserved. We show that precedent plus the right causal structure and conditional independence implies pairs of observed variables also satisfy CIIR. Here. causal structures encode the assumption of independent causal mechanisms but not intervention operations. That is, even without interventions, causal structure and precedent can be enough to learn about the consequences of your actions.

Keywords causal inference · decision theory

1 Introduction

Sometimes we want to make decisions supported by data. Structural causal models are a standard framework for addressing this kind of problem. Roughly speaking, there are two different ways to use the structural modelling framework to help make decisions: in some situations, we are confident in our ability to construct the correct structural model from prior information, and the model helps us draw valid deductions from this prior information. On the other hand, in some situations we have little prior information about the correct structural model, and we might attempt causal discovery in order to learn it from the given data. We can then use the learned model to draw valid inferences as before.

It is a well-known difficulty of causal inference that the available data may not be sufficient to identify the causal effects of interest. A further difficulty is that the options a decision maker has available often do not have a simple correspondence to interventions on the learned causal graph. This means that, even if a structural model is known, additional knowledge is needed to determine how the consequences of actions should be modeled. This may generally be particularly problematic in settings that call for causal discovery, because in such situations we do not have detailed prior knoweldge about causal relationships.

[*]

Consider, for example, an author who wants to know what genre to pick for their next book in order to maximise sales – science fiction or romance. Suppose that, using a large dataset from a bookseller, they are able to learn a structural model. The exact nature of the model is not critical for the example, but for concreteness we will suppose it contains three variables: sales S, genre G (as judged by the bookseller) and "covariates" C, which is actually a pair of variables: a statistic based on the author's historical sales record, and a statistic based on recent trends in genre sales.

The theory of perfect interventions tells us that, given a distribution $\mathbb{P}(\mathsf{S}, \mathsf{G}, \mathsf{C})$ estimated from the data, an intervention on genre will yield the intervened distribution

$$\mathbb{P}(\mathsf{S}, \mathsf{C}|\text{do}(\mathsf{G} = x)) = \mathbb{P}(\mathsf{C})\mathbb{P}(\mathsf{S}|\mathsf{C}, \mathsf{G} = x)$$

Suppose that the conditional $\mathbb{P}(\mathsf{S}|\mathsf{C}, \mathsf{G})$ tells us that at any level of author skill, given observed recent sales trends, romance novels are expected to outsell science fiction novels. This seems to suggest that the author is well-advised to write a romance novel.

However, there is a difference between deciding to write a romance novel and actually writing one. Having made the decision, one is not guaranteed to end up having written a romance novel (at least, according to the bookseller) – or even having written a novel at all. Furthermore, there are an enormous number of different ways to end up with a book classed as "romance". The author could, for example, write any novel at all and press booksellers to label it "romance". Even if they do write a book that is honestly classed as "romance", there are an enormous number of different books that qualify, and many of these different books are likely to have different sales prospects.

Inspection of the model (??) does not tell us which, if any, of these options is a "canonical intervention on genre". This problem is somewhat reminiscent of the problem of ambiguous manipulations raised by Spirtes and Scheines [2004]. Spirtes and Scheines note that when a variable is composed of multiple variables with clear intervention semantics, the composite variable often fails to have clear intervention semantics of its own. Pearl [2009, Ch. 11] states "there is no way a model can predict the effect of an action unless one specifies which variables are affected by the action and how."

We could propose is that causal models are not the right tool for judging the consequences of "coarse" actions like choosing the genre of a book to write – perhaps they are meant to give more precise answers to more precise questions. However, on the face of it, it seems reasonable to ask: on the basis of given data, what genre should I aim to write? Perhaps our author could do better by deliberating over more detailed plans, but picking the genre suggested by (??) seems better than picking randomly.

There are many other decision problems for which the relationship between actions and interventions is ambiguous. A notable example is the idea of an intervention on body mass index examined by [Hernán and Taubman, 2008, Hernán, 2016]. Hernán and Taubman consider the example of different options that are known a priori to affect a person's body mass index, including diet plans, gastric bypass surgery and limb removal. As with actions affecting genre, none of these options seem to be viable candidates for a "canonical intervention on body mass index", and opinion shared by other authors [Pearl, 2018, Hernán and Cole, 2009, Shahar, 2009]).

In this paper, we investigate an approach to modelling decision problems that takes supporting decisions, rather than representing causal relationships, to be the primary objective. Decision models represent actions we can take and consequences we want to evaluate, while representing causal relationships is entirely optional. Such models have been studied previously by numerous authors, including Heckerman and Shachter [1995], Dawid [2012, 2020], Lattimore and Rohde [2019a,b]. An advantage of decision models is that, by construction, it is clear how actions correspond to features of the model. A disadvantage is that, unlike causal models, decision models do not come equipped with inference rules "by default". Thus the main objective of this paper is to explore inference rules in decision models.

Independent and identically distributed (IID) sequences of variables are a foundational concept in statistics. The first inference rule we introduce is a generalisation of the IID assumption. Variable sequences with independent and identical responses (IIR) feature pairs of "input" and "output" variables related by identical stochastic response functions. For IIR decision models with unknown response functions, we prove a generalisation of De Finetti's representation theorem [de Finetti, [1937] 1992]. We show that an IIR decision model with an unknown response function is equivalent to an input-output contractible decision model (Theorem 4.18). IO-contractibility is a somewhat complex symmetry of decision models which implies (among other things) the interchangeability of sufficiently large samples of experimental and observational data.

IIR sequences, however, is not a promising inference rule in general. We argue that sufficiently large samples of experimental and observational data are usually not interchangeable, and thus it is usually unreasonable to use IIR decision models. This is not a novel view; it is analogous to the standard view that causal effects are typically not known to be identified in observational data. However, our argument for this conclusion is new, which shows that the decision model framework can at least offer a different perspective on familiar problems.

We then turn to a more promising inference rule. This inference rule has a number of conditions which are somewhat complex, and will be explained in more detail in Section 5. Briefly, the first condition is precedent. In the context of our book writing example, this is the assumption that, given either of the available options (write a science fiction or romance book), the distribuiton over sales is given by some unknown reweighting of the observations. The next condition is an observed conditional independence – in our example, this could be the observation that book sales are independent of the author's identity given the genre and the identified covariates. The final condition is absolute continuity of conditionals. This requirement is not easy to explain without the formalism we introduce, and so we will save the explanation for the relevant section. This condition is implied by certain causal structures along with the assumption of independent causal mechanisms, such as the following:

These three conditions, together, justify assuming that the consequences of (successfully) writing a romance novel are the interventional consequences as given by the original model (??) (Theorem 5.8). Thus, rather than needing to assume a correspondence between actions and interventions – which we've argued can be problematic – we can derive it from the previously mentioned conditions.

The assumption of independent causal mechanisms underpins the assumption of faithfulness that facilitates conditional independence based causal discovery algorithms [Meek, 1995], as well as many alternative approaches to causal discovery [Lemeire and Janzing, 2013]. Theorem 5.8 also suggests that it could also underpin the interventional interpretation of structural causal models, without needing to introduce interventions as an additional assumption (though this is only a preliminary suggestion, and it may fail to play out under further investigation).

1.1 Connections to previous work in causal inference

Our approach starts with the assumption that we are trying to model options and consequences, and we do not demand that our models capture any other notion of causation. This assumption motivates the formalism of "decision models". This approach is in the tradition of the the decision theoretic approach to causal inference that has been formalised in slightly different ways by Heckerman and Shachter [1995] and Dawid [2012, 2020]. While Lattimore and Rohde [2019a,b] do not explicitly call their approach "decision theoretic", it is also similar to our approach in that options are explicitly incorporated into the model.

Lindley and Novick [1981] discussed of sequences of exchangeable observations along with "one more (possibly non-exchangeable) observation". This construction is very similar to our "see-do models" (Definition 5.1). 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.

There have been a number of works on symmetries in causal inference reminiscent of our work on input-output contractibility. Rubin [2005], Imbens and Rubin [2015] made use of the assumption of models with exchangeable potential outcomes to prove several identification results. Saarela et al. [2020], used graphical causal models to propose conditional exchangeability, defined as the exchangeability of the non-intervened causal parents of a target variable under intervention on its remaining parents. Sareela et. al. suggested that this could be interpreted as a symmetry of an experiment involving administering treatments to patients with respect to exchanging the patients in the experiment. Hernán and Robins [2006], Hernán [2012], Greenland and Robins [1986], Banerjee et al. [2017], Dawid [2020] all discuss similar experimental symmetries. A key difference between all of these causal symmetries and input-output contractibility is that these are all counterfactual symmetries – they say that, had the experiment been performed differently (say, if different treatments had been administered to different patients), the same model would be used to analyse it. Input-output contractibility, on the other hand, is a data symmetry – it holds that there are certain transformations that do not affect the choice of appropriate model. Our work on input-output contractibility is also distinguished by the fact that we prove the equivalence of input-output contractible decision models and decision models with conditionally independent and identical responses, which is required in any case where any conditionals that arise “as a consequence of my actions” are thought to be identical to conditionals in previously observed data.

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

Theorem 5.8 was inspired by causal inference by invariant prediction [Peters et al., 2016]. While both the assumptions and the conclusions drawn in that work differ from the assumptions and conclusion of Theorem 5.8, both look for variable pairs X and Y such that the distribution of Y given X doesn't change when actions are taken. Unlike Peters et. al., our result does not make use of structural interventions, and the connection to the principle of independent causal mechanisms is original to this work.

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

1.2 Outline

Section 2 outlines our mathematical framework and provides a brief reference on notation. Section 4.2 introduces decision models with conditionally independent and identical responses, a generalisation of conditionally independent and identicallly distributed variables. We then introduce and explains Theorem 4.18, a “decision model analogue” for De Finetti's represention theorem, and finally argue, on the basis of this theorem, that the assumption of conditionally independent and identical responses is often unreasonable.

Section 5 looks at a more promising inference rule. It introduces the notion of precedent and then proves Theorem 5.8, which establishes that precedent along with some additional conditions implies conditionally independent and identical responses. We then examine these additional conditions in more detail, and show that they are supported by structural assumptions, if those assumptions are interpreted as expressing the independence of certain conditional probabilities.

2 Technical Prerequisites

Our approach to causal inference is based on probability theory. This section gathers some necessary technical definitions, and is included for reference. A reader who wishes to follow the arguments of the paper may skip to Section 4 and refer back to this section as required.

Section 2.1 introduces the notation used in this paper. Because decision models are stochastic functions rather than probability measures (Section ??), we make use a generalisation of conditional independence called extended conditional independence, explained in Section ?? – the content of these sections may be less familiar.

2.1 Probability Notation

We refer a reader to Çinlar [2011] chapters 1, 2 and 4 for an introduction to the probability theory we use use in this paper. Here we offer a brief overview of our notation.

We denote a measurable space $(A, \mathcal{A})$. Given a collection U of subsets of A, $\sigma(U)$ is the smallest σ-algebra containing U.

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

We write random variables $\mathsf{X} : A \to X$; the variable and its codomain share names but are written with different fonts. Given a probability measure $\mathbb{P}$ on $(A, \mathcal{A})$, $\mathbb{P}^{\mathsf{X}}$ is the marginal distribution of X and $\mathbb{P}^{\mathsf{X}|\mathsf{Y}}$ is the conditional distribution of X given Y.

A sequence of random variables (X, Y) is itself a random variable $\omega \mapsto (\mathsf{X}(\omega), \mathsf{Y}(\omega))$. We denote by $*$ a random variable $* : A \mapsto \{*\}$ where $(\{*\}, \{\emptyset, \{*\}\})$ is equipped with the indiscrete or “trivial” σ-algebra.

We denote by $\Delta(A)$ the set of all probability measures on $(A, \mathcal{A})$.

We use the Iverson bracket $[\![\text{condition}]\!]$ for the function that evaluates to 1 if condition is true and 0 if it is false. The Dirac measure $\delta_x \in \Delta(X)$ is the probability measure for which $\delta_x(A) = [\![x \in A]\!]$.

3 Decision Models

We are interested in modelling decision making rather than prediction. The key difference is that different decisions can lead to different outcomes. This is not the case for prediction problems, where the outcome is unaffected by the prediction offered.

Decisions differ from ordinary random variables in that a decision maker does not require a probability distribution over their options. A decision maker requires a model that accepts any proposed option and offers a forecast of its consequences, and on the basis of this model they make their decision. A probability distribution over options does not contribute to this decision making process, and so it is not required in a decision model (in fact, it is generally accepted that – whatever their merits – probability distributions over options should not contribute to decision making Liu and Price [2020]).

Instead of a probability distribution which offers unconditional forecasts of outcomes, a decision maker requires a function that maps their options to outcomes. We model this with a Markov kernel, a function that maps options to probability distributions.

Definition 3.1 (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 $x \in E$

We use an alternative notation for the signature of a Markov kernel to stress the fact that we can consider it a map from a measurable set to a set of probability distributions.

Notation 3.2 (Signature of a Markov kernel). Given measurable spaces $(E, \mathcal{E})$ and $(F, \mathcal{F})$ and a Markov kernel $\mathbb{M} : E \times \mathcal{F} \to [0, 1]$, we write $\mathbb{M} : E \rightarrowtriangle F$, which we read as "$\mathbb{M}$ maps from E to probability measures on F".

A decision model is a generalisation of a probability space. A decision model combines an option set, a sample space and a Markov kernel that maps from the option set to probability measures on the sample space.

Definition 3.3 (Decision model). A decision model is a triple $(\mathbb{P}_\cdot, (\Omega, \mathcal{F}), (C, \mathcal{C}))$ where $\mathbb{P}_\cdot : C \rightarrowtriangle \Omega$ is a Markov kernel, $(\Omega, \mathcal{F})$ is the sample space and $(C, \mathcal{C})$ is the set of choices (or options).

As in standard probability spaces, we take random variables to be measurable functions on the sample space.

Definition 3.4 (Random variable). Given a decision model $(\mathbb{P}_\cdot, (\Omega, \mathcal{F}), (C, \mathcal{C}))$, an X-valued random variable is a measurable function $\mathsf{X} : (\Omega, \mathcal{F}) \to (X, \mathcal{X})$.

3.1 Conditional distributions and conditional independence in decision models

Decision models differ from regular probability models in that we always need a choice α from the set C. There is in general no unconditional "$\mathbb{P}$" nor any distribution over choices "$\mathbb{P}^{\mathsf{C}}$". This is due to the fact that we noted above – a decision maker does not make use of a distribution over C in the course of their deliberation, and this is reflected by the absence of any distribution not conditioned on a choice.

If we have two random variables X and Y, for each $\alpha \in C$ we have a conditional distribution $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}}$.

The only notion of conditional independence we require is conditional independence for each $\alpha \in C$. We say $\mathsf{X} \perp\!\!\!\perp \mathsf{Y}|(\mathsf{Z}, \mathsf{C})$ – read "X is independent of Y given Z and C" if, for each $\alpha \in C$, $\mathsf{X} \perp\!\!\!\perp_{\mathbb{P}_\alpha} \mathsf{Y}|\mathsf{Z}$ – that is, relative to $\mathbb{P}_\alpha$, X is independent of Y given Z.

There is an extended notion of conditional independence for decision models given by Constantinou and Dawid [2017], which is substantially more general than the notion we use.

Our notion of conditional independence satisfies the standard properties:

1. Symmetry: $\mathsf{X} \perp\!\!\!\perp \mathsf{Y}|(\mathsf{Z}, \mathsf{C})$ iff $\mathsf{Y} \perp\!\!\!\perp \mathsf{X}|(\mathsf{Z}, \mathsf{C})$

2. $\mathsf{X} \perp\!\!\!\perp \mathsf{Y}|(\mathsf{Y}, \mathsf{C})$
3. Decomposition: $\mathsf{X} \perp\!\!\!\perp (\mathsf{Z}, \mathsf{Y})|(\mathsf{W}, \mathsf{C})$ implies $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}_\cdot} \mathsf{Z}|(\mathsf{W}, \mathsf{C})$ and $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}_\cdot} \mathsf{Y}|(\mathsf{W}, \mathsf{C})$
4. Weak union:
   (a) $\mathsf{X} \perp\!\!\!\perp (\mathsf{Y}, \mathsf{Z})|(\mathsf{W}, \mathsf{C})$ implies $\mathsf{X} \perp\!\!\!\perp (\mathsf{Y}, \phi)|(\mathsf{Z}, \mathsf{W}, \mathsf{C})$
5. Contraction: $\mathsf{X} \perp\!\!\!\perp \mathsf{Z}|(\mathsf{W}, \mathsf{C})$ and $\mathsf{X} \perp\!\!\!\perp \mathsf{Y}|(\mathsf{Z}, \mathsf{W}, \mathsf{C})$ implies $\mathsf{X} \perp\!\!\!\perp (\mathsf{Y}, \mathsf{Z})|(\mathsf{W}, \mathsf{C})$

In a decision model, we say that two random variables are almost surely equal if they are almost surely equal for every $\alpha \in C$. That is, given X, Y

$$\mathsf{X} \overset{C}{\cong} \mathsf{Y}$$
$$\iff$$
$$\mathbb{P}_\alpha(\mathsf{X} \neq \mathsf{Y}) = 0 \qquad \forall \alpha \in C$$

We say that two random variables are almost surely equal on a particular $\alpha' \in C$ if they are almost surely equal for α' but not necessarily for any other element of O:

$$\mathsf{X} \overset{\alpha'}{\cong} \mathsf{Y}$$
$$\iff$$
$$\mathbb{P}_{\alpha'}(\mathsf{X} \neq \mathsf{Y}) = 0$$

3.2 Directed graphs

Definition 3.5 (Directed graph). A directed acyclic graph $\mathcal{G}$ is a set of nodes $\mathcal{V}$ and a set of edges $\mathcal{E}$. Each edge is an ordered pair of nodes $(V_i, V_j) \in \mathcal{V}^2$, with V_i the source and V_j the destination. An acyclic graph must have no directed path that begins and ends at V_i for any $V_i \in \mathcal{V}$.

Definition 3.6 (Directed path). Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a directed path is a sequence of edges $((V_k^1, V_k^2))_{k\in[n]}$ from $\mathcal{E}$ such that for any $k \in [n]$, $V_k^2 = V_{k+1}^1$. A directed path begins as V_1^1 and ends at V_k^2.

Definition 3.7 (Directed acyclic graph). A directed graph $\mathcal{G}$ is a directed acyclic graph if it contains no directed paths beginning and ending at the same node.

Definition 3.8 (Parents, ancestors, nondescendents). Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the parents of a node V_i are all the nodes V_j such that there is an edge $(V_j, V_i) \in \mathcal{E}$: $\mathrm{Pa}(V_i) = \{V_j | (V_j, V_i) \in \mathcal{E}\}$.

The ancestors of V_i are all nodes V_j which are either parents of V_i or parents of ancestors of V_i (this is a recursive definition).

Nondescendents of V_i, $\mathrm{ND}(V_i)$, are nodes V_j such that V_i is not an ancestor of V_j.

Definition 3.9 (Model graph association). Given a set of variables $(\mathsf{V}_i)_{i\in[k]}$, an associated directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a graph with a node V_i for each variable V_i. We define the parents of a variable via this association: $\mathrm{Pa}_{\mathcal{G}}(\mathsf{V}_i) = \{\mathsf{V}_j | (V_j, V_i) \in \mathcal{E}\}$.

parameters

4 Inferring consequences when observations and consequences share responses

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

In more detail, suppose our decision maker has a decision model $(\mathbb{P}_\cdot, (C, \mathcal{C}), (\Omega, \mathcal{F}))$ 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, $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_\cdot} (\mathrm{id}_C, \mathsf{X}_{[m+n]\setminus\{i\}}, \mathsf{Y}_{[m+n]\setminus\{i\}})|(\mathsf{X}_i, \mathsf{H})$

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

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

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

While engineered systems might be designed just so that their components exhibit repeatable responses, we might suppost that most systems exhibit regular response relationships if we knew exactly where to look for these relationships. We frequently observe predictable behaviour in non-engineered systems from the weather to people's online shopping. This predictable behaviour is a result, perhaps, of observing systems that respond to inputs in regular ways, primed with regular inputs. This might be so, but it tells us little about whether particular pairs of variables share identical probabilistic responses. We will discuss this intuition further in Section 5.

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

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

In the example presented here – concerning the relationship between body mass index and mortality – we consider such an assumption to be unreasonable. In fact, we argue that this result suggests that an assumption of independent and identical responses is untenable in many situations. If, for example, an analyst believes that there would be any benefit in checking their results against experimental data, then our result indicates that this attitude amounts to a rejection of the assumption of independent and idential responses. While we argue on this basis that the assumption of independent and identical responses is often untenable, we argue in Section 5 that a special case of this assumption we refer to as precedent is more plausible.

4.1 Conditionally independent and identical responses

We formalise decision models with "shared but unknown responses" as sequential models of variable pairs with conditionally independent and identical responses (CIIRs). A sequence of pairs $(\mathsf{D}_i, \mathsf{Y}_i)_{i\in\mathbb{N}}$ share conditionally independent and identical responses if there is an unknown stochastic function H taking values in $\Delta(Y)^D$ – 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 our example featured a decision maker who 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 sequential input-output models as a shorthand for a decision model along with a sequence of variable pairs.

Definition 4.1 (Sequential input-output model). A decision model $(\mathbb{P}_\cdot, (C, \mathcal{C}), (\Omega, \mathcal{F}))$ 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 $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$.

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

Definition 4.2 (Conditionally independent and identical responses). Given a sequential input-output model $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$, the $(\mathsf{D}_i, \mathsf{Y}_i)$ pairs are related by independent and identical responses conditional on H if for all i,

$\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{D}_{[1,i)}, \mathsf{Y}_{[1,i)})|(\mathsf{D}_i, \mathsf{H}, \mathrm{id}_C)$ and $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}} \overset{\mathbb{P}_\alpha^{\mathsf{D}_i|\mathsf{H}}}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}_j|\mathsf{D}_j\mathsf{H}}$ for all i, j.

Definition 4.2 asserts that there are versions of all the conditional distributions $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}}$ that are pairwise congruent, and Theorem 4.3 shows that this is sufficient for the existence of a single conditional distribution that is a version of $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}}$ for all i.

Theorem 4.3 (Existence of representative conditional distribution). Given a sequential input-output model $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$, if the $(\mathsf{D}_i, \mathsf{Y}_i)$ pairs are related by independent and identical responses conditional on H, then for every α, $\mathbb{P}_\alpha$-almost all $h \in H$ there is a representative conditional distribution h_X^Y such that $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{H}}(\cdot|h) \overset{\mathbb{P}_\alpha^{\mathsf{D}_i|\mathsf{H}}(\cdot|h)}{\cong} h_X^Y$ for all i.

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

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

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

Define

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

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

In general, the definition for conditionally independent and identical responses only requires that the outputs Y_i are 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 4.4 (Weakly data-independent). A sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ with independent and identical responses conditional on H is weakly data-independent if $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{D}_{\mathbb{N}\setminus\{i\}}|(\mathsf{D}_i, \mathsf{H}, \mathrm{id}_C)$.

4.2 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 reminiscent of exchangeability, and in Theorem B.9 we show that it implies that the $(\mathsf{D}_i, \mathsf{Y}_i)$ share conditionally independent and identical responses. However, the converse is not true, as Example 4.5 shows.

Example 4.5. Suppose there is a machine with two arms $D = \{0, 1\}$, one of which always pays out \$100 50% of the time and nothing otherwise, and the other that pays out nothing. A decision maker (DM) doesn't know which is which, but DM watches a sequence of people operate the machine. The first person in the sequence was told yesterday exactly which arm is good, and most likely remembers. The second one has no idea which arm is good, and does not observe the first person's choice. The DM is sure that they all want the money, so the first person will pull the good arm $1 - \epsilon$ of the time, while the second person will pull the good arm 50% of the time. The response H takes values that can be summarised as "0 is good" and "1 is good" (which we'll just refer to as $\{0, 1\}$), and the DM assigns 50% probability to each initially. Then

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

while

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

Example 4.5 motivates the weaker symmetry we call exchange commutativity. The key difference is that exchange commutativity allows for the permutability of pairs after conditioning on some arbitrary variable W. 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 4.6 (Exchange commutativity). Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ along with some $\mathsf{W} : \Omega \to W$, we say $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ commutes with exchange over 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 consider can be roughly understood as the idea that Y_i doesn't "depend on" D_j for $j \neq i$. As Example 4.5 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 4.7 (Locality). Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ along with some $\mathsf{W} : \Omega \to W$, the model is local over 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 4.9.

Definition 4.8 (Input-output contractibility). A sequential input-output model $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ along with some $\mathsf{W} : \Omega \to W$ is input-output contractible (IO contractible) over W if it is local and commutes with exchange.

Theorem 4.9 (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

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

Proof. Appendix B.1 □

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

4.3 Representation of IO contractible models

In this section, we state Theorem 4.18, which shows that a sequential input output model $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ features pairs $(\mathsf{D}_i, \mathsf{Y}_i)$ related by conditionally independent and identical responses if and only if it is IO contractible over some variable W.

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

4.4 Preliminaries

Definition 4.10 (Input count variable). Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ with countable D, $\#_j^k$ is the variable

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

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

If we have an infinite sequence of pairs $(\mathsf{D}_i, \mathsf{Y}_i)$, we can wrap the sequence 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 4.11 (Tabulated conditional distribution). Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ on $(\Omega, \mathcal{F})$, define the tabulated conditional distribution $\mathsf{Y}^D : \Omega \to Y^{\mathbb{N}\times D}$ by

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

That is, the (i, j)-th coordinate of 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 [Kallenberg, 2005]. The directing random measure is a probability measure. For completeness, we also define a directing random measure in the case that the relevant limit does not exist, although we are only practically interested in using the definition where the limit does exist.

Definition 4.12 (Directing random measure). Given a probability set $(\mathbb{P}_C, \Omega, \mathcal{F})$ and a sequence $\mathsf{X} := (\mathsf{X}_i)_{i\in\mathbb{N}}$, the directing random measure of X written $\mathsf{H} : \Omega \to \Delta(X)$ 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 $((\mathsf{Y}^D_{1j})_{j\in D}, (\mathsf{Y}^D_{2j})_{j\in D}, ...)$.

Definition 4.13 (Directing random conditional). Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$, we will say the directing random conditional $\mathsf{H} : \Omega \to \Delta(Y^D)$ is the function

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

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

Definition 4.14 (Permutation of rows). Given a sequence of indices $(i,j)_{i\in\mathbb{N},j\in D}$ a finite permutation of rows is a function $\eta : \mathbb{N}\times D \to \mathbb{N}\times D$ such that for each $j \in D$, $\eta_j := \eta(\cdot, j)$ is a finite permutation $\mathbb{N} \to \mathbb{N}$ and $\eta(i,j) = (\eta_j(i), j)$.

Lemma 4.16 shows that an IO contractible conditional distribution can be represented as the product of a column exchangeable probability distribution and a "lookup function" or "switch". This lookup function is also used in the representation of potential outcomes models (see, for example, Rubin [2005]), but we do not assume that the tabulated conditional 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 4.18.

To prove Lemma 4.16, we assume that the set of input sequences in which each value appears infinitely often has measure 1 for every option in C. Without this assumption, we would have to accept positive probability that we run out of 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 4.15 (Infinite support). Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ with D countable if, letting $E \subset D^{\mathbb{N}}$ be the set of all sequences such that for all $j \in D$

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

we have $\mathbb{P}_\alpha^{\mathsf{D}|\mathsf{W}}(E|w) = 1$ for all 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 4.16. 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 4.11 and verify that it satisfies IO contractibility.

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

Because the distribution $\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}$ from Lemma 4.16 is row-exchangeable, the limit in the definition of the directing random conditional H exists almost surely (see Lemma B.6). 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 4.17. 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 B.3. □

We are now ready to state the main result, Theorem 4.18. Assuming a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ (Definition 4.1) with inputs D infinitely supported (Definition 4.15) 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 4.2), given by the directing random conditional H (Definition 4.13) and $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ is weakly data-independent.

4.5 Statement of the representation theorem

Theorem 4.18 (Representation of IO contractible models). Suppose a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ with sample space $(\Omega, \mathcal{F})$ is given with D countable and D infinitely supported over W. Then the following are equivalent:

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

Proof. (1) $\implies$ (3): We apply Lemma 4.16 followed by Lemma B.6 followed by Lemma B.7.

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

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

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

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

Corollary 4.19. 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, $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{W}|(\mathsf{D}_i, \mathsf{H}, \mathrm{id}_C)$ 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 4.18 that $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ is IO contractible over W. The conclusion follows by applying Theorem 4.18 a second time. □

Building on Corollary 4.19, Theorem 4.20 shows the assumption that the pairs $(\mathsf{D}_i, \mathsf{Y}_i)$ are related by conditionally independent and identical responses implies that, for the purposes of learning the response function 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 4.20. 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 B.5. □

4.6 Does IO contractibility help us understand identification?

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

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

The particular assumption we have in mind is, in the world of causal graphical models, the assumption that backdoor adjustment is possible and in the world of potential outcomes it is the assumption of conditional ignorability [Rubin, 2005]. Both assumptions hold that, given a treatment 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 4.20, which holds that, if the sequences of observations and consequences are both infinite, then for the purpose of learning the response function the problem is unchanged by swapping any subset of the indices corresponding to observations with any subset of those corresponding to consequences. That is, there is no difference between predicting the response function of the passive observations from an infinite sequence of passive observational data and predicting the response function of the consequences of the decision makers actions from the same sequence of passive observational data.

In practice, we propose that it would be very rare to have both of these datasets and treat them as interchangeable in this manner. Example 4.21 makes a similar point.

Example 4.21. Suppose an experiment is done which assigns some medical treatment D_i uniformly according to some random signal to patients for even i, and allows assignment by patient and doctor discretion for odd i. Y_i is a binary variable recording some health outcome of interest and X_i is some vector of covariates. The sequence $(\mathsf{D}_c, \mathsf{X}_c, \mathsf{Y}_c)$ is associated with the consequences of a decision maker's choices, where c is some special character not in $\mathbb{N}$.

According to Theorem 4.20, the assumption of conditionally independent and identical responses applied to $((\mathsf{D},\mathsf{X}),\mathsf{Y})$ implies

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

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

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

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

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

5 Inferring consequences when options have precedent

We have suggested that conditionally independent and identical responses is usually an unreasonably strong assumption for a decision maker to make, on the grounds that it implies overly strong interchangeability properties between different datasets. One way to get around this objection is to suppose that conditionally independent and identical responses are shared by pairs $(\mathsf{E}_i,\mathsf{X}_i)$ where the 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 simplify the presentation, we will consider a specific kind of decision model featuring long sequence of exchangeable observations indexed by natural numbers that are unresponsive to the decision maker's choice and "one more" variable representing the "consequences of action" indexed by the special character c that may be responsible ot the decision maker's choice. That is, we have $(\mathsf{X}_i)_{i\in\mathbb{N}}$ unresponsive to the decision maker and (X_c) responsive to the decision maker. Call this setup a "see-do model".

Definition 5.1 (See-do model). A see-do model is an decision model $(\mathbb{P}_\cdot,\Omega,C)$ along with a sequence of variables $\mathsf{X}_{\mathbb{N}\cup\{c\}}$ where $\mathsf{X}_\mathbb{N} \perp\!\!\!\perp^e_{\mathbb{P}_\cdot} \mathrm{id}_C$. Variables indexed with $i\in\mathbb{N}$ are referred to as observations and variables indexed with the special index c are referred to as consequences. We specify a see-do model with the shorthand $(\mathbb{P}_\cdot,\mathsf{X}_{\mathbb{N}\cup\{c\}})$.

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

Recall that in Section 4 we suggested that many systems might exhibit (probabilistically) regular input-output behaviours, but where we might not know or observe the right "inputs". The assumption that the pairs $(\mathsf{E}_i,(\mathsf{X}_i,\mathsf{Y}_i))_{i\in\mathbb{N}\cup\{c\}}$ share conditionally indpendent and identical responses can be viewed as a formalisation of this intuition; there is some unknown and unobserved state E_i which X_i and Y_i respond to in a regular manner no matter what else is happening.

Note that we make no assumptions about the distribution of Z_c.

Definition 5.2 (Latent CIIR see-do model). A latent CIIR see-do model is a see-do model $(\mathbb{P}_\cdot, (\mathsf{E}_i, \mathsf{Z}_i, \mathsf{X}_i, \mathsf{Y}_i)_{i\in\mathbb{N}\cup\{c\}})$ such that the observation pairs $(\mathsf{Z}_i, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i))_{i\in\mathbb{N}}$ share conditionally independent and identical responses and the pairs $(\mathsf{E}_i, (\mathsf{X}_i, \mathsf{Y}_i))_{i\in\mathbb{N}\cup\{c\}}$ also share conditionally independent and identical responses. We say the E_is are "latent" variables, which informally means that we typically do not get to observe them. We adopt the convention that the directing random conditional of $(\mathbb{P}_\cdot, \mathsf{Z}_\mathbb{N}, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i)_{i\in\mathbb{N}})$.

We can take any see-do model $(\mathbb{P}_\cdot, \mathsf{X}_{\mathbb{N}\cup\{c\}})$ with exchangeable observations and turn it into a latent CIIR see-do model by setting $\mathsf{Z}_i = *$ and $\mathsf{E}_i = (\mathsf{X}_i, \mathsf{Y}_i)$. This trivial construction typically isn't very helpful, though. One particular feature we might want is for a latent CIIR model to express the fact that "things we can do have been done before"; that is, any setting of the unobserved state E_c that our actions might yield has positive probability in the observed data. Example 5.3 illustrates model constructions with and without this property.

Example 5.3. Suppose we have a see-do model $(\mathbb{P}_\cdot, \mathsf{X}_{\mathbb{N}\cup\{c\}})$ where each X_i takes values in a binary set, and the control we can exert is to choose either $\mathbb{P}_0^{\mathsf{X}_c} = \frac{1}{4}\delta_0 + \frac{3}{4}\delta_1$ or $\mathbb{P}_1^{\mathsf{X}-c} = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$, independent of all other observations. Suppose further that for $i \in \mathbb{N}$, $\mathbb{P}_C^{\mathsf{X}_i} = \frac{3}{4}\delta_0 + \frac{1}{4}\delta_1$ independent of all other observations. Then we can consider this model to be IO contractible with latent binary inputs E_i such that

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

This is not the only way to construct such a model. We could instead choose latent binary inputs E_i' such that

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

On the other hand, the choice E_i'' with

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

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

In the first construction in Example 5.3, but not the following two, we have $\mathbb{P}_C^{\mathsf{E}_i} \gg \mathbb{P}_\alpha^{\mathsf{E}_c}$ for all α. We say under this construction the options have precedent; they have, in a sense, "been done before". The assumption of precedent by itself has some implications – for example, if a decision maker considers precedent a reasonable assumption and they have access to a lot of data, they they should not expect any of their actions to lead to consequences that have never appeared before in the observational data. In Theorem 5.8, we will make use a stronger version of this assumption where the conditional distribution over E_i is different for each value of Z_i, which leads to stronger conclusions. We will discuss the plausibility of the stronger assumption afterwards.

Theorem 5.8 is motivated by the following example:

Example 5.4. Suppose a decision maker collects data about a group of peope who have variously engaged the services of dietiticians, sporting coaches, general practitioners, bariatric surgeons and none of the above, with practitioner choice recorded under the variable Z_i. The decision maker has also collected data on each person's body mass index X_i at the beginning of the study and followed mortality outcomes Y_i for a considerable period of time. A decision maker is reviewing this data, and in particular is wondering if steps they take to manage their weight X_c are likely to improve their own mortality prospects Y_c.

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

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

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

Theorem 5.8 establishes some sufficient conditions for the informal deduction described in Example 5.4. We assume that all variables of interest are discrete, and make use of an alternative notation for discrete conditional probabilities.

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

The first assumption we make is precedent. This assumption strengthens our latent CIIR model by asserting that any distributions over the latent variable that can be accomplished by the decision maker's actions are dominated by the "observational" distribution. That is, loosely speaking, "anything the decision maker can do has been done before and its consequences observed" (though the decision maker cannot necessarily recognise when it was done before by inspecting the data).

Definition 5.6 (Precedent).

A further key assumption for Theorem 5.8 is an assumption we call conditional absolute continuity. The basic idea is that, after a decision maker decides that Y_i is independent of Z_i given X_i, the DM's model still maintains that the distribution of $\mathsf{G}_{\mathsf{Z}}^{\mathsf{EX}}$ conditional on $\mathsf{G}_{\mathsf{EXZ}}^{\mathsf{Y}}$ is still absolutely continuous with respect to the Lebesgue measure. That is, after deciding on the conditional indepedence – itself a measure 0 event with respect to the Lebesgue measure – the conditional distribution of $\mathsf{G}_{\mathsf{Z}}^{\mathsf{EX}}$ does not also concentrate on a measure zero set. As we will see, this assumption can follow from some structural causal assumptions along with a version of the principle of independent causal mechanisms Lemeire and Janzing [2013].

Definition 5.7 (Conditional absolute continuity). Given a latent CIIR see-do model $(\mathbb{P}_\cdot, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i, \mathsf{Z}_i)_{i\in\mathbb{N}\cup\{c\}})$ with E, X, Y and Z all discrete, recall G is the directing random conditional of $(\mathbb{P}_\cdot, \mathsf{Z}_\mathbb{N}, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i)_{i\in\mathbb{N}})$.

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

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

as well as the precedent condition:

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

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

For Theorem 5.8, we assume that on the basis of observations we condition the probability on some event I (in particular, we are interested in the case where I is the event that a certain conditional independence holds).

Theorem 5.8 (Latent to observable IO contractibility). Given a latent CIIR see-do model $(\mathbb{P}_\cdot, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i, \mathsf{Z}_i)_{i\in\mathbb{N}\cup\{c\}})$ with E, X, Y and Z all discrete, recall G is the directing random conditional of $(\mathbb{P}_\cdot, \mathsf{Z}_\mathbb{N}, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i)_{i\in\mathbb{N}})$.

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

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

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

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

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

Full proof in Appendix C. □

6 Continuity and discovery

Causal discovery and conditional absolute continuity

We've already discussed the assumption of precedent, and suggested it may be plausible when we can accept that the outcomes of interest exhibit a probabilistically regular response to some unobserved state, and when the values of this unobserved state that our actions may bring about are represented in the observed data. We will now turn our attention to the assumption of conditional absolute continuity. An intuitive explanation of Theorem 5.8 is that, if we observe Y_i independent of Z_i given X_i, then we can either explain this via "precise alignment" of G_{EZ}^X and G_Z^E or by Y_i also being independent of E_i given X_i. Conditional absolute continuity rules out the "precise alignment" hypothesis.

When is such an assumption justified? We do not have a definitive answer to this question. However, standard assumptions made in the causal discovery literature imply conditional absolute continuity of parental conditional distributions. Thus, under standard interpretations of causal structure, conditional absolute continuity can be justified by structural assumptions. The assumption that causal structures imply conditional absolute continuity of parental conditionals is independent of the other standard interpretation of structural causal models – that they are "atlases for interventions". What we have shown is that if one is willing to accept the assumption of precedent then conditional absolute continuity is enough to draw conclusions about the consequences of actions from causal structures without additionally assuming some correspondence between actions and any kind of structural interventions.

We offer a formal definition of the conditional absolute continuity interpretation of causal structures in the form of an absolutely continuous structural decision model.

Definition 6.1 (Absolutely continuous structural decision model). An absolutely continuous structural decision model is a decision model $(\mathbb{P}_., \Omega, C)$ along with a collection of variable sequences $(\mathsf{X}_j^i)_{i\in A\subset\mathbb{N}, j\in B\subset\mathbb{N}}$ (where X_j^i and X_k^i are successive observations of the "same kind of variable") and a corresponding directed acyclic graph $\boldsymbol{G}$ with nodes $\{X^i | i \in A\}$. Letting $X^F := \{X^i \| i \in F \subset A\}$ be the nodes with no parents in $\boldsymbol{G}$, we require that the corresponding variables $(\mathsf{X}^F, \mathsf{X}^{F^\complement})$ share conditionally independent and identical responses with parameter H. We consider assuming $\boldsymbol{G}$ to impose further restrictions on the distribution of the parameter H, which we will abuse notation to render as "conditioning on G" (this is an abuse of our notation as $\boldsymbol{G}$ is not a random variable):

- (Markov condition) for all $U \subset A$, $\mathsf{X}_j^U \perp\!\!\!\perp^e_{\mathbb{P}_.} \mathrm{ND}_{\boldsymbol{G}}(\mathsf{X}_j^U)|(\mathrm{Pa}_{\boldsymbol{G}}(\mathsf{X}_j^U), \mathrm{Id}_C, \boldsymbol{G})$

- (Absolute continuity) the joint distribution of parameters $\mathbb{P}_.^{(\mathsf{H}^{\mathsf{X}^i}{}_{\mathrm{Pa}_{\boldsymbol{G}}(\mathsf{X}^i)})_{i\in A}|\boldsymbol{G}}$ is absolutely continuous with respect to the uniform joint distribution on the space of parameters $(\mathsf{H}^{\mathsf{X}^i}_{\mathrm{Pa}_{\boldsymbol{G}}(\mathsf{X}^i)})_{i\in A}$.

The absolute continuity condition implies that, for arbitrary $U, V \subset A$, $\mathbb{P}_.^{(\mathsf{H}^{\mathsf{X}^i}{}_{\mathrm{Pa}_{\boldsymbol{G}}(\mathsf{X}^i)})_{i\in U}|(\mathsf{H}^{\mathsf{X}^i}\,\mathrm{Pa}_{\boldsymbol{G}}(\mathsf{X}^i))_{i\in V}\boldsymbol{G}}$ is almost surely absolutely continuous with respect to the uniform joint distribution on the space of parameters $(\mathsf{H}^{\mathsf{X}^i}_{\mathrm{Pa}_{\boldsymbol{G}}(\mathsf{X}^i)})_{i\in U}$.

Suppose we have an absolutely continuous structural decision model $(\mathbb{P}_., \Omega, C)$ where we have observed Z_i independent of Y_i given X_i, and for which we have an arbitrary graph $\boldsymbol{G}$ with nodes E, X, Y, Z satisfying the following three assumptions

1. E_i is a parent of X_i and not independent of X_i
2. Z_i is a parent of X_i and not independent of X_i
3. X_i, E_i and Z_i are nondescendants of Y_i

Under these assumptions, $\mathrm{Pa}_{\boldsymbol{G}}(\mathsf{X}_i) = (\mathsf{E}_i, \mathsf{Z}_i)$, while $\mathrm{Pa}_{\boldsymbol{G}}(\mathsf{E}_i)$ may be Z_i or empty, $\mathrm{Pa}_{\boldsymbol{G}}(\mathsf{Z}_i)$ may be E_i or empty and $\mathrm{Pa}_{\boldsymbol{G}}(\mathsf{Y}_i)$ may be any subsequence of $(\mathsf{E}_i, \mathsf{X}_i, \mathsf{Z}_i)$.

(or together with structural assumptions imply conditional absolute continuity, which we will explain here. Causal structures have two useful properties for this purpose: first, they can "explain" conditional independences like Y_i is independent of Z_i given X_i. Second, under standard interpretations, given a structural causal model, parental conditional probabilities are mutually continuous (in our terminology). This property has been noted before, and has been informally stated as the principle of independent causal mechanisms (in spite of the name, parental conditional probabilities need not be probabilistically independent) [Lemeire and Janzing, 2013, Peters et al., 2017].

To illustrate why we need to explain the conditional independence, suppose we pick a "prior" distribution of G_Z^{EXY} (by which we just mean $\mathbb{P}_{\cdot}^{\mathsf{G}_Z^{EXY}}$) such that for each $z, z' \in Z$, the distribution of G_z^{EXY} conditional on $\mathsf{G}_{z'}^{EXY}$ is absolutely continuous with respect to the uniform measure on $E \times X \times Y$. In this case, the event $\mathsf{G}_{Xz}^Y = \mathsf{G}_{Xz'}^Y$ would have probability 0 for all $z, z' \in Z$. As such, this "prior" does not tell us whether diverse precedent holds after conditioning on this event, as conditional probabilities are not determined by the joint distribution on measure 0 events. Thus this prior does not tell us whether diverse precedent holds after conditioning on the independence of Y_i from Z_i given X_i.

Bayesian causal discovery approaches prior specification in a different way. This approach considers a mixture of graphical hypotheses which each imply certain conditional independences [Heckerman et al., 1995]. Because each graphical hypothesis is given positive probability, independences like Y_i independent of Z_i given X_i also have positive probability, in contrast with the approach that sets G_z^{EXY} absolutely continuous with respect to the uniform measure.

Another conventional feature of structure learning is the assumption that the parameters associated with parental conditional distributions are mutually continuous. That is, for any X, fixing a hypothesised structure $\mathcal{G}$, $\mathbb{P}_{\cdot}^{\mathsf{G}^X_{\mathrm{Pa}_\mathcal{G}(X)} | \mathsf{G}^{X^\mathsf{C}}_{\mathrm{Pa}_\mathcal{G}(X^\mathsf{C})}}$ is absolutely continuous (see Definition 3.9 for the meaning of $\mathrm{Pa}_\mathcal{G}(\mathsf{X})$). [Heckerman et al., 1995] assumes these parameters are mutually independent with absolutely continuous marginals. While not explicitly Bayesian, Meek [1995] argues that given a hypothesised causal model, unfaithful[2] distributions are unlikely because they violate mutual absolute continuity.

Putting these properties together, we have:

- We may "explain" the independence $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{Z}_i | (\mathsf{G}, \mathrm{id}_C)$ with a hypothesised causal structure $\mathcal{G}$
- If either Z_i is a parent of E_i or E_i and Z_i are parents of X_i in $\mathcal{G}$ then the relevant continuity property for Theorem 5.8 holds

For example, suppose we consider a class of structural hypotheses $\mathcal{G}_i$ where, in all structures, we have $\mathsf{E}_i \rightarrowtriangle \mathsf{X}_i$ and $\mathsf{G}_i \rightarrowtriangle \mathsf{X}_i$. One such structure is illustrated below:

In $\mathcal{G}_1$ (and, indeed, in every $\mathcal{G}_i$) we have $\mathrm{Pa}_{\mathcal{G}_i}(\mathsf{X}_i) = (\mathsf{Z}_i, \mathsf{E}_i)$. Furthermore, G_Z^E is determined by G^E and G_E^Z, which are both parental parameters in $\mathcal{G}_1$ and therefore mutually absolutely continuous with G_{EZ}^X. Finally, $\mathsf{G}_{EXz}^Y = \mathsf{G}_{EX}^Y$ for all $z \in Z$ is also a parental conditional and therefore also mutually absolutely continuous with G_{EZ}^X. Thus, given hypothesis $\mathcal{G}_1$, if we also accept the assumption of precedent, Theorem 5.8 follows. In fact, we can similarly argue that Theorem 5.8 follows for all structures $\mathcal{G}_i$ where:

- $\mathsf{E}_i \rightarrowtriangle \mathsf{X}_i$ and $\mathsf{Z}_i \rightarrowtriangle \mathsf{X}_i$
- $\mathsf{E}_i\mathsf{X}_i$ and $\mathsf{Z}_i \rightarrowtriangle \mathsf{E}_i$

On the other hand, consider

Here we also have $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{Z}_i | (\mathsf{G}, \mathrm{id}_C)$, but absolute continuity is not obviously supported; neither G_{EZ}^X nor G_Z^E are parental parameters. In fact, we can observe that in $\mathcal{G}_2$ we have $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{Z}_i | (\mathsf{X}_i, \mathsf{G}, \mathrm{id}_C)$ given an arbitrary choice of G_{EX}^Y, but not $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{E}_i | (\mathsf{X}_i, \mathsf{G}, \mathrm{id}_C)$, so by the contrapositive of Theorem 5.8 we must not have the relevant absolutely continuous conditionals.

[2] See the referenced paper for a definition.

If, in our original example, instead of conditioning on medical practitioners, we take Z_i to be an individual's clothing size and (for the sake of argument) find the same result: mortality outcomes Y_i are independent of clothing size Z_i conditional on body mass index X_i. There are no doubt some differences between people with the same body mass index who wear differently sized clothes – height, for example – but it is not clear from the given data whether we should conclude that any common actions affecting a person's health do so via their body mass index, or whether there are features relevant to a person's health that fail to be correlated with clothing size after conditioning on body mass index. This example is motivated by structure (??), though it's hard to come up with an example that is inarguably an instance of this structure and also not excessively convoluted.

While we've argued that structure (??) implies $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}} \mathsf{E}_i|(\mathsf{X}_i, \mathcal{G}_1, \mathsf{G}, \mathrm{id}_C)$ via Theorem 5.8, we can note that it also implies this independence via d-separation. On the other hand, (??) does not imply such a conditional independence via Theorem 5.8 or via d-separation. This correspondence is not exact - the structure (??) implies $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}} \mathsf{E}_i|(\mathsf{X}_i, \mathcal{G}, \mathsf{G}, \mathrm{id}_C)$ via d-separation, but not via Theorem 5.8. The existence of such stuctures may not be especially surprising, given that Theorem 5.8 establishes sufficient but not necessary conditions.

The alternative d-separation criterion does uphold the general rule we observed, where $\mathsf{E}_i \rightarrow \mathsf{X}_i$ and $\mathsf{Z}_i \rightarrow \mathsf{X}_i$ or $\mathsf{E}_i\mathsf{X}_i$ and $\mathsf{Z}_i \rightarrow \mathsf{E}_i$ are sufficient to establish $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}} \mathsf{E}_i|(\mathsf{X}_i, \mathcal{G}_i, \mathsf{G}, \mathrm{id}_C)$, while the absence of either of these edges is compatible with structures that do not imply the relevant independence.

If we assume structural models are expressing the assumption that parental conditional probabilities are mutually absolutely continuous, it may be the case that d-separation properties capture all interesting consequences like those of Theorem 5.8 where mutual absolute continuity supports additional conditional independence implications. It is well known that conditional independences relationships not implied by a structure $\mathcal{G}$ can be broken by choosing slightly different parental conditionals (see for example Meek [1995], Zhang and Spirtes [2003]), thus under the assumption that all parental conditionals relative to a structure $\mathcal{G}$ are mutually absolutely continuous we would not find additional conditional independence properties. We may wonder if there are interesting mutual absolute continuity assumptions that are not representable as a directed acyclic graph. In fact, the conditions for Theorem 5.8 are somewhat weaker than those expressed by a structural model – it only assumes mutual absolute continuity for a subset of conditionals, whereas a structural model assigns every variable a parental set and implies the mutual absolute continuity of every parental conditionals. However, as we have argued, this difference does not seem to matter in this case as both approaches yield the same conclusion.

In this discussion, we have suggested that structural causal models may play a role informing judgements of mutual absolute continuity. While this is a standard feature of the interpretation of structural causal models – and Lemeire and Janzing [2013] has proposed that this idea should play a critical role in our understanding of causal structures – for the purposes of assessing consequences of actions, the interpretation of causal models as "interventional oracles" [Pearl, 2009, Section 1.3.1] is usually given prominence. Here, we consider precedent as an alternative to the assumption of structural interventions, and lean on the interpretation of mutual absolute continuity to derive nontrivial conclusions.

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?". Firstly, we examined the assumption of conditionally independent and identical responses, and its equivalent form in IO contractibility, which we argued was often an unreasonable assumption and secondly, we examined an approach based on the principle of precedent, or the idea that the decision maker's options have been taken before, and some of their consequences observed. Our approach allows us to consider the question of what observations and consequences have in common independently from any prior knowledge the decision maker might have about how their choices influence outcomes – neither Theorem 4.18 nor Theorem 5.8 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 4.18 are unlikely to be compelling in many situations. While the diverse precedent assumption may be more broadly plausible, it is at this stage difficult to evaluate. Speculatively, it may be possible to make progress on this question by better understanding when structural assumptions support this conclusion, via for example the causal version of the principle of maximum entropy.

For practical purposes, a generalisation of Theorem 5.8 to approximate independence is in order, and such a generalisation may also bring additional clarity to the diverse precedent assumption.

Despite these challenges, we are encouraged by a number of features of this work. Using decision making as a starting point for constructing models means that, at the outset, we are only making commitments a decision maker is likely to already be making if they want to apply a formal theory of decision making. The informal idea of precedent that underpins Theorem 5.8 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 5.8 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.

More precisely, the assumption that parental conditionals are mutually absolutely continuous is explicit in Bayesian approaches to causal discovery such as Heckerman et al. [1995]. It is arguably implicit in the justification for non-Bayesian approaches such as in Meek [1995], who argues that it is reasonable to neglect sets of parental conditionals that have Lebesgue measure zero (though no explicit prior is placed on the parameters). The general idea that parental conditionals will usually avoid "precise coordination" has also been suggested as an alternative foundation for causal discovery Lemeire and Janzing [2013], Mooij, J.M. et al. [2016], ?.

Concretely, some structural models imply conditional absolute continuity in the sense of Definition ??. If we are able to narrow our structural hypotheses to a subset of these models, then we can conclude that conditional absolute continuity holds and (if we also have precedent), so does Theorem 5.8.

If the investigator is confident a priori that their actions will affect X_i, we might justify the first assumption by arguing that their actions affect the observed X_i via the unobserved E_i and X_i is thus a descendant of E_i. The non-independence of Z_i and X_i may be observed, and the direction of their relationship may be argued on the basis of temporal ordering, or inferred via, for example, one of the methods in [Peters et al., 2017]. The third assumption may be justified by appealing to faithfulness given the observed conditional independence and the first two assumptions, as graphs satisfying 1 and 2 but not 3 will not feature Y_i d-separated from Z_i given E_i.

In any case, the three assumptions imply that the joint distribution of $(\mathsf{G}^{\mathsf{EZ}}, \mathsf{G}^X_{EZ}, \mathsf{G}^Y_{EXZ})$ is absolutely continuous with respect to the Lebesge measure, and therefore G^X_{EZ} is absolutely continuous with respect to the Lebesgue measure conditional on G^E_Z and G^Y_{EXZ}, sufficient for conditional absolute continuity.

The assumption of conditional absolute continuity is disjunctive, and there is correspondingly an alternate set of structural assumptions that yield the assumption as a conclusion.

- $1'$ Z_i is a parent of E_i (and not independent of E_i)
- $2'$ E_i is a parent of X_i (and not independent of X_i)
- $3'$ Z_i is a nondescendant of X_i
- $4'$ $(\mathsf{E}_i, \mathsf{Z}_i)$ is a nondescendant of Y_i

Note that we can motivate these assumptions from alternative premises: the third assumption here follows from the first two and acyclicity, and the fourth from the first two plus the observation Z_i independent of Y_i given X_i and faithfulness.

In this case, we have $(\mathsf{G}^E_Z, \mathsf{G}^{XY}_{EZ})$ mutually absolutely continuous with respect to the Lebesgue measure, which gives G^E_Z absolutely continuous WRT Lebesgue conditional on $(\mathsf{G}^X_{EZ}, \mathsf{G}^Y_{EXZ})$ also.

Under either set of structural assumptions, d-separation and faithfulness yields a very similar result to Theorem 5.8. Both sets of assumptions leave us with an unblocked path from E_i to Z_i after conditioning on

X_i. Thus, if there is an unblocked path from E_i to Y_i after conditioning on X_i then there is also an unblocked path from Z_i to Y_i. Conversely, the absence of an unblocked path from Z_i to Y_i conditional on X_i – i.e. X_i d-separates these nodes – implies also the absence of an unblocked path from Y_i to E_i, i.e. that Y_i is independent of E_i conditional on X_i.

References

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

Erhan Çinlar. Probability and Stochastics. Springer, 2011.

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

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

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

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

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

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

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

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

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

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

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

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

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

David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning, 20(3):197–243, September 1995. ISSN 1573-0565. doi: 10.1007/BF00994016. URL `https://doi.org/10.1007/BF00994016`.

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

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

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

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

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

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

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

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

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

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

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

Yang Liu and Huw Price. Ramsey and Joyce on deliberation and prediction. Synthese, 197(10):4365–4386, October 2020. ISSN 1573-0964. doi: 10.1007/s11229-018-01926-8. URL `https://doi.org/10.1007/s11229-018-01926-8`.

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.

Mooij, J.M., Peters, J., Janzing, D., Zscheischler, J., Schölkopf, B., and Amsterdam Machine Learning lab (IVI, FNWI). Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. Journal of Machine Learning Research, 17, 2016. ISSN 1532-4435. URL `http://dare.uva.nl/personal/pure/en/publications/distinguishing-cause-from-effect-using-observational-data-methods-and-benchmarks(f4e35a2b-cd22-41a4-` `.html`.

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

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

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

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

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

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

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

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

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

Peter Spirtes and Richard Scheines. Causal Inference of Ambiguous Manipulations. Philosophy of Science, 71(5):833–845, December 2004. ISSN 0031-8248, 1539-767X. doi: 10.1086/425058. URL https://www.cambridge.org/core/journals/philosophy-of-science/article/abs/causal-inference-of-ambiguous-manipulations/2A605BCFFC1A879A157966473AC2A6D2. Publisher: Cambridge University Press.

Jiji Zhang and Peter Spirtes. Strong Faithfulness and Uniform Consistency in Causal Inference. In Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence, UAI'03, pages 632–639, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc. ISBN 978-0-12-705664-7. URL http://dl.acm.org/citation.cfm?id=2100584.2100661. event-place: Acapulco, Mexico.

A String Diagrams

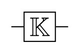
We use a string diagram notation to represent probabilistic functions. This is a notation created for reasoning about abstract Markov categories, and is somewhat different to existing graphical languages. The main difference is that in our notation wires represent variables and boxes (which are like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

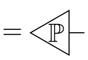
We make use of string diagram notation for probabilistic reasoning. Graphical models are often employed in causal reasoning, and string diagrams are a kind of graphical notation for representing Markov kernels. The notation comes from the study of Markov categories, which are abstract categories that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

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

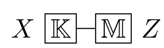
A.1 Elements of string diagrams

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

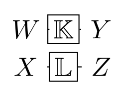
$\mathbb{K} :=$ 

$\mu :=$ 

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

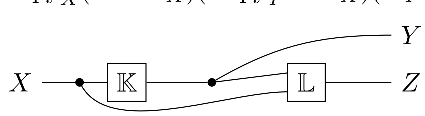
$\mathbb{L}\mathbb{M} :=$ 

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

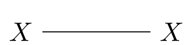
$\mathbb{K} \otimes \mathbb{L} :=$ 

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

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

$=$ 

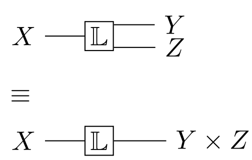
A space X is identified with the identity kernel $\text{id}^X : X \to \Delta(\mathcal{X})$. A bare wire represents the identity kernel:

$\text{Id}^X :=$ 

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 $X \times Y$:

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

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:



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

A.2 Special maps

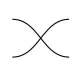
Definition A.1 (Identity map). The identity map $\mathrm{Id}_X : X \rightarrow X$ defined by $(\mathrm{id}_X)(A|x) = \delta_x(A)$ for all $x \in X$, $A \in \mathcal{X}$, is represented by a bare line.

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

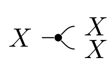
Definition A.2 (Erase map). Given some 1-element set $\{*\}$, the erase map $\mathrm{Del}_X : X \rightarrow \{*\}$ is defined by $(\mathrm{Del}_X)(*|x) = 1$ for all $x \in X$. It "discards the input". It looks like a lit fuse:

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

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

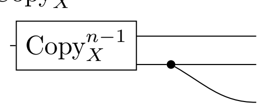
$\mathrm{Swap}_{X,Y} :=$ 

Definition A.4 (Copy map). The copy map $\mathrm{Copy}_X : X \rightarrow X \times X$ is defined by $(\mathrm{Copy}_X)(A \times B|x) = \delta_x(A)\delta_x(B)$ for all $x \in X$, $A, B \in \mathcal{X}$. It makes two identical copies of the input, and is drawn as a fork:

$\mathrm{Copy}_X :=$ 

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

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

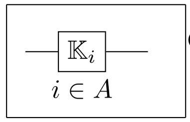
$\mathrm{Copy}^n_X =$ 

$n > 1$

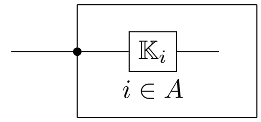
Plates In a string diagram, a plate that is annotated $i \in A$ means the tensor product of the $|A|$ elements that appear inside the plate. A wire crossing from outside a plate boundary to the inside of a plate indicates an $|A|$-fold copy map, which we indicate by placing a dot on the plate boundary.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

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

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

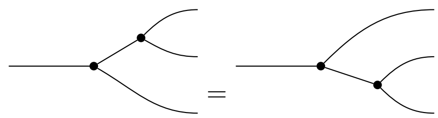


$$\mathrm{Copy}_X^{|A|}(\bigotimes_{i\in A} \mathbb{K}_i) :=$$



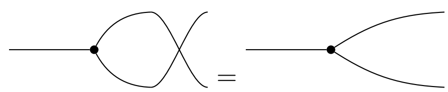
A.3 Commutative comonoid axioms

Diagrams in Markov categories satisfy the commutative comonoid axioms.

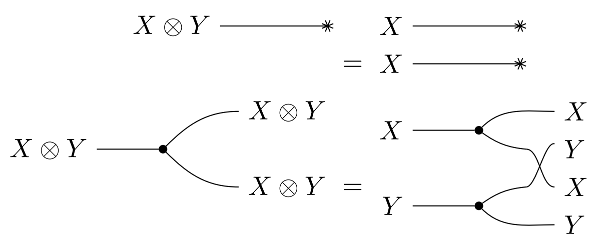


(6)

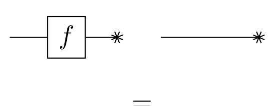
(7)



as well as compatibility with the monoidal structure



and the naturality of Del, which means that

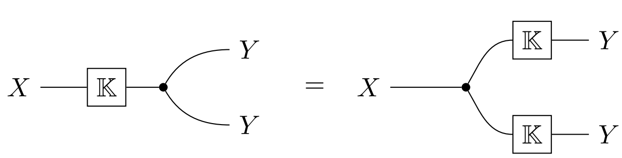


(8)

A.4 Manipulating String Diagrams

Planar deformations along with the applications of Equations (6) through to Equation (8) are almost the only rules we have for transforming one string diagram into an equivalent one. One further rule is given by Theorem A.6.

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

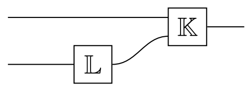


holds iff $\mathbb{K}$ is deterministic.

A.4.1 Examples

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

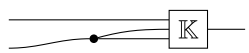
For arbitrary $\mathbb{K} : X \times Y \rightarrowtriangle Z$, $\mathbb{L} : W \rightarrowtriangle Y$



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

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

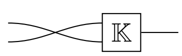
For arbitrary $\mathbb{K} : X \times Y \times Y \rightarrowtriangle Z$:



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

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

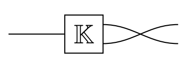
For arbitrary $\mathbb{K} : X \times Y \rightarrowtriangle Z$



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

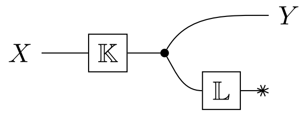
The swap map before a kernel switches the input arguments.

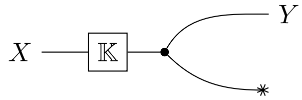
For arbitrary $\mathbb{K} : X \rightarrowtriangle Y \times Z$



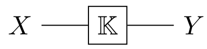
$$\begin{aligned}
&= \mathbb{K}\mathrm{Swap}_{YZ} \\
(\mathbb{K}\mathrm{Swap}_{YZ})(A \times B|x) &= \int_{Y\times Z} \delta_y(B)\delta_z(A)\mathbb{K}(\mathrm{d}y \times \mathrm{d}z|x) \\
&= \int_{B\times A} \mathbb{K}(\mathrm{d}y \times \mathrm{d}z|x) \\
&= \mathbb{K}(B \times A|x)
\end{aligned}$$

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

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

$=$ 

by Eq. (8)

$=$ 

by Eq. (7)

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

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

B Symmetries of conditional probabilities

B.1 Equality of equally sized contractions

This is the proof of Theorem 4.9.

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

Definition B.1 (Finite transposition). Given two equally sized sequences $A, B \in \mathbb{N}^n$ with $A = (a_i)_{i\in[n]}$, $B = (b_i)_{i\in[n]}$, $A \to B : \mathbb{N} \to \mathbb{N}$ is the permutation such that

$$[A \to B](a_i) = b_i$$

that sends the ith element of A to the ith element of B and vise versa. Note that $B \to A$ is the inverse of $A \to B$.

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

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

$$\mathbb{K}_j(\mathrm{id}_{Y^i} \otimes \mathrm{Del}_{Y^{j-i}}) = \mathbb{K}_i \otimes \mathrm{Del}_{X^{j-i}} \tag{9}$$

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

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

Proof. Take any $x \in X^{\mathbb{N}}$ and let $x_{|m} \in X^n$ be the first n elements of x. By Equation (9), for any $A_i \in \mathcal{Y}$, $i \in [m]$

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

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

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

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

$$\mathbb{Q}_x(\bigtimes_{i\in[n]} A_i \times Y^{\mathbb{N}}) = \mathbb{K}_n(\bigtimes_{i\in[n]} A_{\rho(i)}|x_{[n]}) \tag{10}$$

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

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

so by the Monotone convergence theorem, the sequence $\mathbb{Q}_x(\bigtimes_{i\in[n]} A_i)$ converges as $n \to \infty$ to $\mathbb{Q}_x(\bigtimes_{i\in\mathbb{N}} A_i)$. $x \mapsto \mathbb{Q}_x^{\mathsf{Z}_n}(\bigtimes_{i\in[n]} A_i)$ is measurable for all n, $\{A_i \in \mathcal{Y}|i \in \mathbb{N}\}$ by Equation (10), and so $x \mapsto Q_x$ is also measurable.

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

Corollary B.3. Given $(\mathbb{P}_C, \Omega, \mathcal{F})$, $\mathsf{W} : \Omega \to V$ and two pairs of sequences $(\mathsf{V}, \mathsf{X}) := (\mathsf{V}_i, \mathsf{X}_i)_{i\in\mathbb{N}}$ and $(\mathsf{Y}, \mathsf{Z}) := (\mathsf{Y}_i, \mathsf{Z}_i)_{i\in\mathbb{N}}$ with corresponding variables taking values in the same sets $V = Y$ and $X = Z$, if $(\mathbb{P}_C, \mathsf{V}, \mathsf{X})$ and $(\mathbb{P}_C, \mathsf{Y}, \mathsf{Z})$ are both local over W and

$$\mathbb{P}^{\mathsf{X}_{[n]}|\mathsf{W}\mathsf{V}_{[n]}} = \mathbb{P}^{\mathsf{Z}_{[n]}|\mathsf{W}\mathsf{Y}_{[n]}}$$

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

$$\mathbb{P}^{\mathsf{X}|\mathsf{W}\mathsf{V}} = \mathbb{P}^{\mathsf{Z}|\mathsf{W}\mathsf{Y}}$$

Proof. By assumption of locality

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

hence for all $n, m > n$

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

and, in particular, by lemma B.2, $\mathbb{P}^{\mathsf{X}|\mathsf{W}\mathsf{V}}$ and $\mathbb{P}^{\mathsf{Z}|\mathsf{W}\mathsf{Y}}$ are the limits of the same sequence. □

Theorem 4.9. Given a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ and some W, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{W}\mathsf{D}}$ is IO contractible over W if and only if for all subsequences $A, B \subset \mathbb{N}^{|A|}$ and for every α

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

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

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

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

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

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

if A is infinite, then we can take finite subsequences A_m that are the first m elements of A and similarly for B_m. Then by previous reasoning

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

then by Corollary B.3

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

Finally, by locality

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

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

B.2 Examples of symmetries

These are the examples referenced in Section 4.2. Example B.4 shows that neither locality nor exchange commutativity is implied by the other.

Example B.4. We prove the claim by way of presenting counterexamples.

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

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

then

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

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

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

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

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

Example B.5. Consider an experiment where I first flip a coin and record the results of this flip as the outcome Y_1 of "step 1". Subsequently, I can either copy the outcome from step 1 to the result for "step 2" (this is the input $\mathsf{D}_1 = 0$), or flip a second coin use this as the input for step 2 (this is the input $\mathsf{D}_1 = 1$). D_2 is an arbitrary single-valued variable. Then for all d_1, d_2

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

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

B.3 Representation theorem preliminaries

This is the proof of Lemmas 4.16 and B.6 and Theorem 4.17 from Section 4.4. In addition, Lemmas B.6 and B.7 are presented and proved, which will be later used in the proof of Theorem 4.18.

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

Definition 4.10. Given a sequential input-output model $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ on $(\Omega, \mathcal{F})$ with countable D, $\#_j^k$ is the variable

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

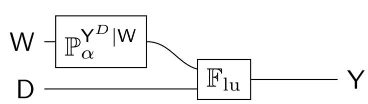
In particular, $\#_j^k$ is equal to the number of times $\mathsf{D}_i = j$ over all $i < k$.

Definition 4.11. Given a sequential input-output model $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ on $(\Omega, \mathcal{F})$, define the tabulated conditional distribution $\mathsf{Y}^D : \Omega \to Y^{\mathbb{N}\times D}$ by

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

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

Lemma 4.16. Suppose a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ is given with D countable and D infinitely supported. Then for some W, α, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}$ is IO contractible if and only if

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

$$\iff$$

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

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

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

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

$$\mathbb{P}_\alpha^{(\mathsf{Y}^D_{ij})_{\mathbb{N}\times D}|\mathsf{W}} = \mathbb{P}_\alpha^{(\mathsf{Y}^D_{\eta(i,j)})_{\mathbb{N}\times D}|\mathsf{W}} \tag{11}$$

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

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

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

$$\begin{aligned}\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}\mathsf{D}_{\mathsf{R}^{-1}}}(A|w,d) &= \mathbb{P}_\alpha^{\mathsf{Y}_{\mathsf{R}^{-1}}|\mathsf{W}\mathsf{D}_{\mathsf{R}^{-1}}}(A|w,d)\\ &\overset{\mathbb{P}_\cdot}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}_{\mathsf{R}^{-1}}|\mathsf{W}\mathsf{D}_{\mathsf{R}^{-1}}}(A|w,e)\\ &= \mathbb{P}_\alpha^{\mathsf{Y}^D}(A|w)\end{aligned}\tag{12}$$

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

Now,

$$\mathbb{P}_\alpha^{\mathsf{Y}_{\mathsf{R}^{-1}}|\mathsf{W}\mathsf{D}_{\mathsf{R}^{-1}}}(A|w,d) = \int_R \mathbb{P}_\alpha^{\mathsf{Y}_\rho|\mathsf{W}\mathsf{D}_\rho}(A|d)\mathbb{P}_\alpha^{\mathsf{R}^{-1}|\mathsf{W}\mathsf{D}_{\mathsf{R}^{-1}}}(\mathrm{d}\rho|w,d)\tag{13}$$

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 $n \in \mathbb{N}$

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

By Corollary B.3, it must therefore be the case that

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

Then from Equation (13)

$$\begin{aligned}\mathbb{P}_\alpha^{\mathsf{Y}_{\mathsf{R}^{-1}}|\mathsf{W}\mathsf{D}_{\mathsf{R}^{-1}}}(A|w,d) &\overset{\mathbb{P}_\cdot}{\cong} \int_R \mathbb{P}_\alpha^{\mathsf{Y}_\rho|\mathsf{W}\mathsf{D}_\rho}(A|d)\mathbb{P}_\alpha^{\mathsf{R}^{-1}|\mathsf{W}\mathsf{D}_{\mathsf{R}^{-1}}}(\mathrm{d}\rho|w,d)\\ &\overset{\mathbb{P}_\cdot}{\cong} \int_R \mathbb{P}_\cdot^{\mathsf{Y}|\mathsf{W}\mathsf{D}}(A|w,d)\mathbb{P}_\alpha^{\mathsf{R}^{-1}|\mathsf{W}\mathsf{D}_{\mathsf{R}^{-1}}}(\mathrm{d}\rho|w,d)\\ &\overset{\mathbb{P}_\cdot}{\cong} \mathbb{P}_\cdot^{\mathsf{Y}|\mathsf{W}\mathsf{D}}(A|w,d)\end{aligned}\tag{14}$$

for all $i, j \in \mathbb{N}$. Then by Equation (12) and Equation (14)

$$\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}(A|w) = \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{W}\mathsf{D}}(A|w,e)\tag{15}$$

Take some $d \in D^{\mathbb{N}}$. From Equation (15) and IO contractibility of $\mathbb{P}_\cdot^{\mathsf{Y}|\mathsf{W}\mathsf{D}}(A|e)$,

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

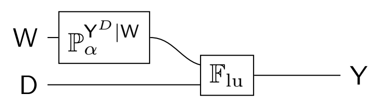
It remains to be shown that Y^D is invariant to finite permutations within rows. Consider some finite permutation within columns $\eta : \mathbb{N} \times D \to \mathbb{N} \times D$, note that $e_{\eta(i,j)} = j$ and hence $(e_{\eta(i,j)})_{i\in\mathbb{N},j\in D} = e$. Thus

$$\begin{aligned}
\mathbb{P}_\alpha^{(\mathsf{Y}^D_{\eta(i,j)})_{\mathbb{N}\times D}|\mathsf{W}}(A|w) &= \mathbb{P}_\alpha^{(\mathsf{Y}^D)_{\mathbb{N}\times D}|\mathsf{W}}\mathrm{Swap}_\eta(A|w) \\
&= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}\mathrm{Swap}_\eta(A|w,e) && \text{from Eq. (15)} \\
&= \mathbb{P}_\alpha^{\mathsf{Y}_\eta|\mathsf{WD}}(A|w,e) \\
&= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}_{\eta^{-1}}}(A|w,e) && \text{by exchange commutativity} \\
&= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}(A|w,(e_{\eta^{-1}(i,j)})_{i\in\mathbb{N},j\in D}) \\
&= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}(A|w,e) \\
&= \mathbb{P}_\alpha^{(\mathsf{Y}^D_{ij})_{\mathbb{N}\times D}|\mathsf{W}}(A|w) && \text{from Eq. (15)}
\end{aligned}$$

If: We construct a conditional probability according to Definition 4.11 and verify that it satisfies IO contractibility.

Suppose

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



where $\mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}$ satisfies Equation (11).

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

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

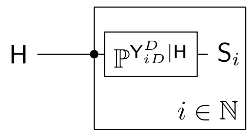
so $\mathbb{K}$ is IO contractible by Theorem 4.9. $\square$

As a consequence of Lemma 4.16 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.

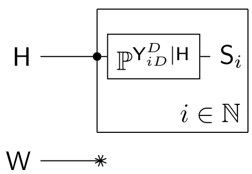
Lemma B.6. Suppose a sequential input-output model $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ is given with D countable, D infinitely supported over some W and $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ IO contractible over the same W. Then, letting H be the directing random conditional of $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ (Definition 4.13) and $\mathsf{Y}^D_{iD} := (\mathsf{Y}^D_{ij})_{j\in D}$, we have for all $i \in \mathbb{N}$, $\mathsf{Y}^D_{iD} \perp\!\!\!\perp^e_{\mathbb{P}_\cdot} (\mathsf{Y}^D_{\mathbb{N}\setminus\{i\}D}, \mathsf{W}, \mathrm{id}_C)|\mathsf{H}$ and

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

Proof. Fix $w \in W$ and consider $\mathbb{P}^{\mathsf{Y}^D}_{\alpha,w} := \mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}(\cdot|w)$. From Lemma 4.16, we have the exchangeability of the sequence $(\mathsf{Y}^D_{1D}, \mathsf{Y}^D_{2D}, ...)$ with respect to $(\mathbb{P}_{\alpha,w}, \Omega, \mathcal{F})$ as a special case of the invariance of $\mathbb{P}_\alpha^{(\mathsf{Y}^D_{ij})_{\mathbb{N}\times D}|\mathsf{W}}$ to permutations of rows. By the column exchangeability of $\mathbb{P}^{\mathsf{Y}^D}_{\alpha,w}$, from Kallenberg [2005, Prop. 1.4] (where H is precisely what Kallenberg calls the directing random measure)

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

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

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

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

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

Again, the right hand side does not depend on 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 4.17. 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

$$\mathsf{Y}^{D,A}_{ij} = \mathsf{Y}^D_{\mathsf{Q}(i)j} \tag{16}$$

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 (16) 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.

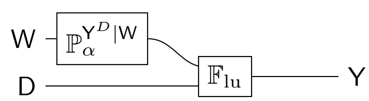
Let $\mathrm{con}_q : Y^{\mathbb{N}\times D} \rightarrow Y^{\mathbb{N}\times D}$ be the Markov kernel associated with the function that sends $(\mathsf{Y}^D_{ij})_{i\in\mathbb{N},j\in D}$ to $(\mathsf{Y}^D_{q(i)j})_{i\in\mathbb{N},j\in D}$. Then for any $B \in \mathcal{Y}^{\mathbb{N}\times D}$, w, q:

$$\begin{aligned}
\mathbb{P}_\alpha^{\mathsf{Y}^{D,A}|\mathsf{WQ}}(B|w,q) &= \mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}\mathrm{con}_q(B|w) \\
&= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}\mathrm{con}_q(B|w,e) && \text{by Eq.(15)} \\
&= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}}(B|w,e) && \text{by Theorem 4.9} \\
&= \mathbb{P}_\alpha^{\mathsf{Y}^D|\mathsf{W}}(B|w) && \text{by Eq.(15)}
\end{aligned} \tag{17}$$

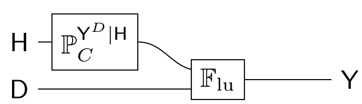
Finally, take H_A the directing random measure of $\mathsf{Y}^{D,A}$. We conclude from the equality Eq. (17) and from the fact that there is a one-to-one map from directing random measures to exchangeable distributions that $\mathsf{H}_A \overset{\mathbb{P}_\alpha}{\cong} \mathsf{H}$. □

The following is a technical lemma that will be used in Theorem 4.18.

Lemma B.7. 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

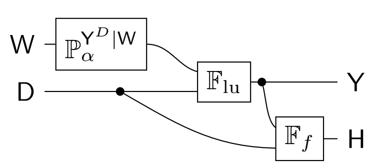
$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}} =$ 

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

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

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 4.16.

Y^D is a function of Y and D (see Definition 4.11) 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 4.12). Because $\mathsf{H} = f(\mathsf{Y}, \mathsf{D})$, we have $\mathsf{H} \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{W}, \mathrm{id}_C)|(\mathsf{Y}, \mathsf{D})$. Thus

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

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 4.16), for each row we can send $\#_{[d=j]_i}$ to i and obtain

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

but by Lemma B.6, the sequence $(\mathsf{Y}^D_{iD})_{i\in\mathbb{N}}$ are mutually independent conditional on 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,

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

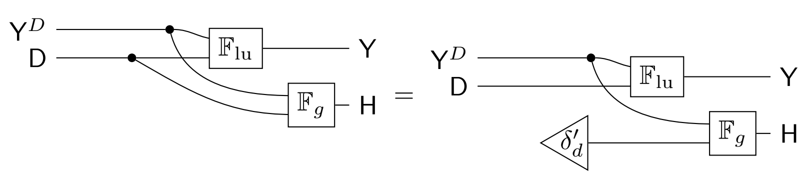
And, as a consequence, defining

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

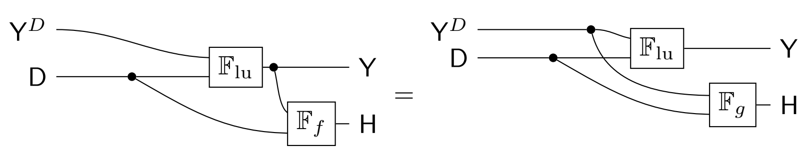
we have

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

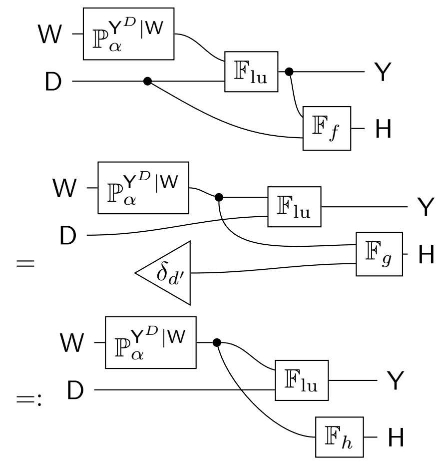
which in turn implies the almost sure equality of the associated Markov kernels:



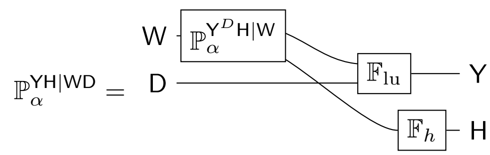
but we also have, by the definitions of f and g



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

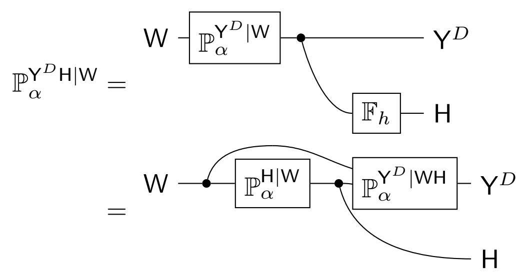


because H is a deterministic function of Y^D, this implies



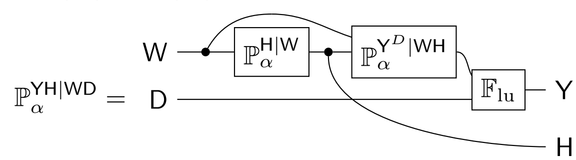
(18)

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

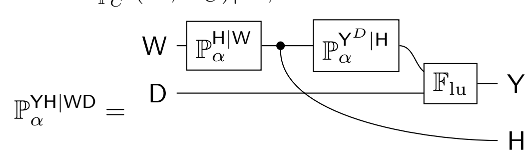


(19)

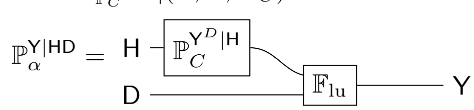
and so by substituting Equation (19) into (18)



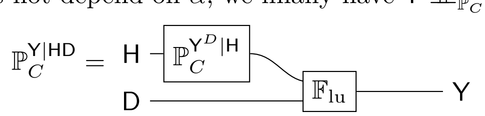
From Lemma 4.16 we also have $\mathsf{Y}^D \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{W}, \mathrm{id}_C)|\mathsf{H}$, so



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

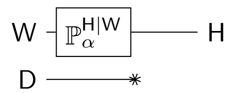


Because the right hand side does not depend on α, we finally have $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{W}, \mathrm{id}_C)|(\mathsf{H}, \mathsf{D})$ and the result



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

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



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

B.4 Representation theorem

This is the proof of the main result from Section 4, Theorem 4.18.

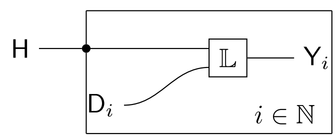
Theorem 4.18. Suppose a sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ with sample space $(\Omega, \mathcal{F})$ is given with D countable and D infinitely supported. Then the following are equivalent:

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

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

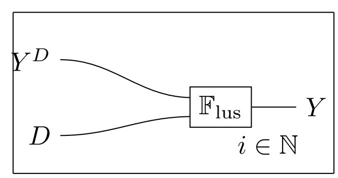
3. There is some $\mathbb{L} : H \times X \rightarrow Y$ such that

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



Proof. As a preliminary, we will show

$$\mathbb{F}_{\mathrm{lu}} = \tag{20}$$



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

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

Recall that lu is the function

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

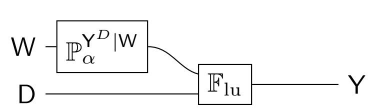
By definition, for any $\{A_i \in \mathcal{Y} | i \in \mathbb{N}\}$

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

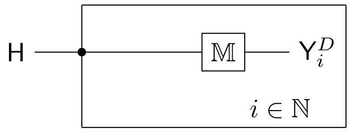
which is what we wanted to show.

(1) $\implies$ (3): From Lemma 4.16, we have some Y^D such that

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

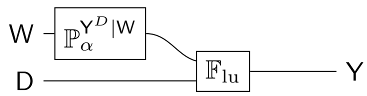


and by Lemma B.6

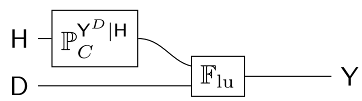
$$\mathbb{P}_C^{\mathsf{Y}^D|\mathsf{H}} =$$ 

(21)

By Lemma 4.16, for each $w \in W$

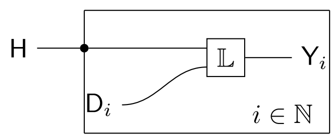
$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WD}} =$$ 

and so by Lemma B.7

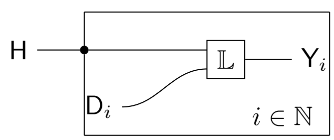
$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{HD}} =$$ 

(22)

We can substitute Equations (21) and (20) into (22) for

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

(3) $\implies$ (2): If

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

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

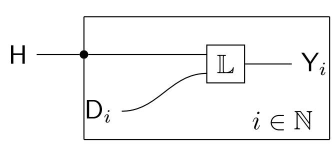
$$\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{HD}_i\mathsf{Y}_{\neq i}\mathsf{D}_{\neq i}} \overset{\mathbb{P}_C}{\cong} \mathbb{L} \otimes \text{Del}_{Y^{\mathbb{N}} \times X^{\mathbb{N}}}$$

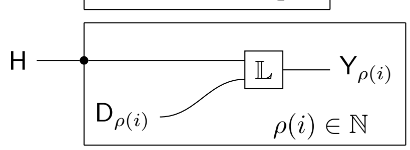
hence $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{\neq i}, \mathsf{D}_{\neq i}, \text{id}_C)|(\mathsf{H}, \mathsf{D}_i)$

(2) $\implies$ (1): Take $\mathsf{W} := \mathsf{H}$. Because we assume $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{[1,i)}, \mathsf{D}_{[1,i),\text{id}_C})|(\mathsf{H}, \mathsf{D}_i)$ we can take $\mathbb{L} := \mathsf{H}_X^Y = \mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{HX}_i}$ for all i, α (existence given by Theorem 4.3) and

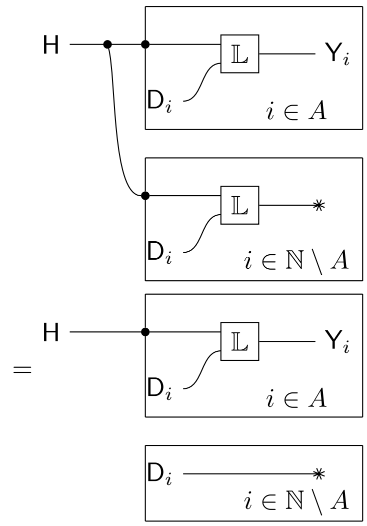
$$\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{HD}_i\mathsf{Y}_{[1,i)}\mathsf{D}_{[1,i)}} \overset{\mathbb{P}_C}{\cong} \mathbb{L} \otimes \text{Del}_{Y^{i-1} \times X^{i-1}}$$

by taking the semidirect product of the conditionals

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

$$=$$ 

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



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

B.5 Consequences of Theorem 4.18

Theorem 4.18 says that a data independent sequential input-output model $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ features conditionally independent and identical response functions $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{HD}_i}$ 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.

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 B.8 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 $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ 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).

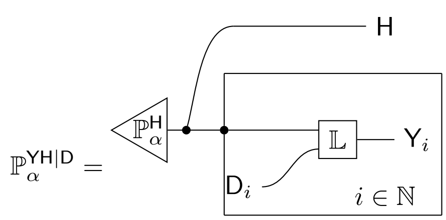
Theorem B.8 (Data-independent IO contractibility). Suppose a sequential input-output model $(\mathbb{P}_\cdot, \mathsf{D}, \mathsf{Y})$ with sample space $(\Omega, \mathcal{F})$ is given with D countable and, letting $E \subset D^{\mathbb{N}}$ be the set of all sequences for which each $j \in D$ occurs infinitely often, $\mathbb{P}_\alpha^{\mathsf{D}}(E) = 1$ for all α. Then the following are equivalent:

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

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

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

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



Proof. See Appendix B.5. □

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

Theorem B.9. 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. $\square$

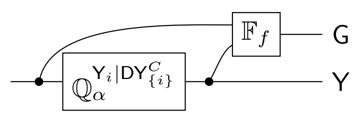
Lemma B.10 (Exchangeably dominated conditionals). Given $(\mathbb{P}_C, \Omega, \mathcal{F})$ and variables D, Y, if for any α there is some $\mathbb{Q}_\alpha$ such that $\mathbb{Q}_\alpha^{\mathsf{DY}}$ is exchangeable with directing random measure G, D is infinitely supported over G with respect to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

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

$$\mathbb{Q}_\alpha^{\mathsf{Y}_i\mathsf{D}_i|\mathsf{G}} = \mathbb{Q}_\alpha^{\mathsf{Y}_j\mathsf{D}_j|\mathsf{G}} \tag{23}$$

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

$$\mathbb{Q}_\alpha^{\mathsf{Y}_i\mathsf{G}|\mathsf{DY}_{\{i\}^\complement}} = $$



$$\begin{aligned} &\overset{P}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}_i\mathsf{G}|\mathsf{DY}_{\{i\}^\complement}} \\ \implies \mathbb{Q}_\alpha^{\mathsf{Y}_i|\mathsf{GDY}_{\{i\}^\complement}} &\overset{P}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{GDY}_{\{i\}^\complement}} \end{aligned} \tag{24}$$

It follows from weak union that

$$\begin{aligned} &\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}_C} (\mathsf{D}_{\{i\}^\complement}\mathsf{Y}_{\{i\}^\complement})|(\mathsf{D}_i, \mathsf{G}, \mathrm{id}_C) \\ \iff &\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{GY}_{\{i\}^\complement}\mathsf{D}_{\{i\}^\complement}}(A|d_i, g, d, y) \overset{P}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{G}}(A|d_i, g) \qquad \forall A, d_i, g, d, y, \alpha \\ \implies &\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{D}_{\{i\}^\complement}\mathsf{Y}_{\{i\}^\complement})|(\mathsf{D}_i, \mathsf{G}, \mathrm{id}_C) \end{aligned} \tag{25}$$

where Eq. (25) follows from Eq. (24).

Finally, from Eq. (23) and Eq. (25)

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

Thus $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ features independent and identical responses conditioned on G, and by Lemma 4.19 it also has independent and identical responses conditioned on H. Finally, the infinite support of D over G with respect to Q_α implies D is also infinitely supported over G with respect to $\mathbb{P}_\alpha$, so by Theorem 4.18 $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{HD}}$ is IO contractible. $\square$

Theorem 4.20. A data-independent sequential input-output model $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ features conditionally independent and identical response functions $\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{G}}$ with D infinitely supported over G only if for any sets $A, B \subset \mathbb{N}$ such that D_A and D_B are also infinitely supported over G and any $i, j \in \mathbb{N}$ such that $i \notin A$, $j \notin B$,

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

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

Proof. Only if: By Theorem 4.18 and Lemma 4.19, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{HD}}$ is IO contractible. By Theorem 4.17, H is almost surely a function of both $(\mathsf{D}_A, \mathsf{Y}_A)$ and $(\mathsf{D}_B, \mathsf{Y}_B)$ and, furthermore, $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{D}_A, \mathsf{Y}_A)|(\mathsf{D}_i, \mathsf{H}, \mathrm{id}_C)$, $\mathsf{Y}_j \perp\!\!\!\perp^e_{\mathbb{P}_C}$

$(\mathsf{D}_B, \mathsf{Y}_B)|(\mathsf{D}_j, \mathsf{H}, \mathrm{id}_C)$. Hence there is some $f : D^{\mathbb{N}} \times Y^{\mathbb{N}} \to H$ such that for all $E \in \mathcal{Y}, d_i \in D, d \in D^{\mathbb{N}}, y \in Y^{\mathbb{N}}$

$$\begin{aligned}
\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{Y}_A,\mathsf{D}_A}(E|d_i, y, d) &= \mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}}(E|d_i, f(y, d)) \\
&= \mathbb{P}_\alpha^{\mathsf{Y}_j|\mathsf{D}_j\mathsf{H}}(E|d_i, f(y, d)) \qquad (26) \\
&= \mathbb{P}_\alpha^{\mathsf{Y}_j|\mathsf{D}_j\mathsf{Y}_B,\mathsf{D}_B}(E|d_i, y, d)
\end{aligned}$$

Where Eq. (26) follows from Theorem 4.9.

If: By construction

$$\mathbb{Q}_\alpha^{\mathsf{Y}_i\mathsf{D}_i\mathsf{Y}_{\{i\mathtt{C}\}}\mathsf{D}_{\{i\mathtt{C}\}}} := \mathbb{Q}_\alpha^{\mathsf{D}_i\mathsf{Y}_{\{i\mathtt{C}\}}\mathsf{D}_{\{i\mathtt{C}\}}} \odot \mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{Y}_{\{i\mathtt{C}\}},\mathsf{D}_{\{i\mathtt{C}\}}}$$

is exchangeable, and by domination $\mathbb{Q}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{Y}_{\{i\mathtt{C}\}},\mathsf{D}_{\{i\mathtt{C}\}}} \overset{\mathbb{P}}{\cong} \mathbb{Q}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{Y}_{\{i\mathtt{C}\}},\mathsf{D}_{\{i\mathtt{C}\}}}$. The result follows from Lemma B.10. □

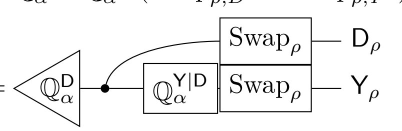
Theorem B.11. Given $(\mathbb{P}_c dot, \mathsf{Y}, \mathsf{D})$, if $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}}$ is exchange commutative for each α, and for each α $\mathbb{P}_\alpha^{\mathsf{D}}$ is absolutely continuous with respect to some exchangeable distribution $\mathbb{Q}_\alpha^{\mathsf{D}}$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

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

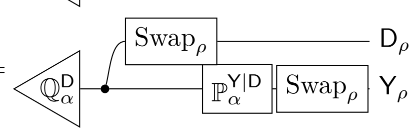
We will show $\mathbb{Q}_\alpha^{\mathsf{DY}}$ is unchanged by finite permutations of $(\mathsf{D}_i, \mathsf{Y}_i)$ pairs. For some finite permutation $\rho : \mathbb{N} \to \mathbb{N}$:

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

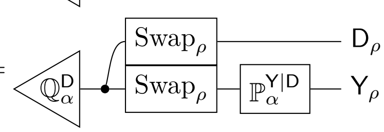
=



=

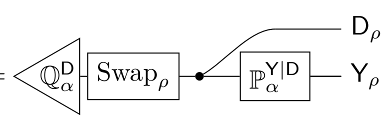


=



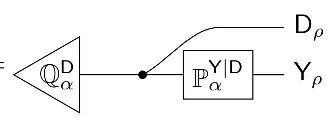
(27)

=



(28)

=



(29)

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

Where line (27) follows from exchange commutativity, (28) follows from Theorem A.6 and the fact that the swap map is deterministic and line (29) comes from the exchangeability of $\mathbb{Q}_\alpha^{\mathsf{D}}$.

Because $\mathbb{P}_\alpha^{\mathsf{D}}$ is dominated by $\mathbb{Q}_\alpha^{\mathsf{D}}$ by assumption, we have $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}} \overset{\mathbb{P}}{\cong} \mathbb{Q}_\alpha^{\mathsf{Y}|\mathsf{D}}$, which implies $\mathbb{Q}_\alpha^{\mathsf{Y}_i|\mathsf{DY}_{\{i\}\mathtt{C}}} \overset{\mathbb{P}}{\cong} \mathbb{Q}_\alpha^{\mathsf{Y}_i|\mathsf{DY}_{\{i\}\mathtt{C}}}$ and from Lemma B.10 we therefore have $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{HD}}$ IO contractible over H, and from Theorem 4.18 we have $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathrm{id}_C|(\mathsf{D}, \mathsf{H})$ and so $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{HD}}$ IO contractible over H also. □

C Precedented options

C.1 IO contractibility from diverse precedent

This is the proof of Theorem 5.8 in Section 5.

Definition 5.7. Given a latent CIIR see-do model $(\mathbb{P}_\cdot, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i, \mathsf{Z}_i)_{i\in\mathbb{N}\cup\{c\}})$ with E, X, Y and Z all discrete, recall G is the directing random conditional of $(\mathbb{P}_\cdot, \mathsf{Z}_\mathbb{N}, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i)_{i\in\mathbb{N}})$.

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

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

as well as the precedent condition:

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

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

Theorem 5.8. Given a see-do model $(\mathbb{P}_\cdot, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i, \mathsf{Z}_i)_{i\in\mathbb{N}\cup\{c\}})$ with E, X, Y and Z all discrete sets, suppose among the observations $i \in \mathbb{N}$ the pairs $(\mathsf{Z}_i, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i))$ share conditionally independent and identical responses and, for all observations and consequences $i \in \mathbb{N}\cup\{c\}$, pairs $(\mathsf{E}_i, (\mathsf{X}_i, \mathsf{Y}_i))$ also share conditionally independent and identical responses. Take G to be the directing random conditional of $(\mathbb{P}_\cdot, \mathsf{Z}_\mathbb{N}, (\mathsf{E}_i, \mathsf{X}_i, \mathsf{Y}_i)_{i\in\mathbb{N}})$.

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

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

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

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

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

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

$$\sum_{e\in E} \mathsf{G}_{exz}^y \frac{\mathsf{G}_{ez}^x \mathsf{G}_z^e}{\sum_{e'\in E} \mathsf{G}_{e'z}^x \mathsf{G}_z^{e'}} \overset{\mathbb{Q}_\alpha}{\cong} \sum_{e\in E} \mathsf{G}_{exz'}^y \frac{\mathsf{G}_{ez'}^x \mathsf{G}_{z'}^e}{\sum_{e'\in E} \mathsf{G}_{e'z'}^x \mathsf{G}_{z'}^{e'}}$$

Conditioning on $\mathsf{G}_{EXZ}^Y = g_{EXZ}^Y$

$$\sum_{e\in E} g_{exz}^y \frac{\mathsf{G}_{ez}^x \mathsf{G}_z^e}{\sum_{e'\in E} \mathsf{G}_{e'z}^x \mathsf{G}_z^{e'}} \overset{\mathbb{P}_C}{\cong} \sum_{e\in E} \mathsf{g}_{exz'}^y \frac{\mathsf{G}_{ez'}^x \mathsf{G}_{z'}^e}{\sum_{e'\in E} \mathsf{G}_{e'z'}^x \mathsf{G}_{z'}^{e'}} \tag{30}$$

Eq. (30) defines a polynomial constraint on $\mathsf{G}_{\{z,z'\}}^{\mathsf{Ex}}$ for each x, z, z'. If $g_{exz}^y = g_{e'xz}^y$ for all e, e' then this constraint is trivial; if $g_{exz}^y = g_{exz'}^y$ also, then it is satisfied for every possible value of $\mathsf{G}_{E\{z,z'\}}^x$, otherwise it is unsatisfiable.

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 for G_{EZ}^x subject to the restriction $g_{exz}^y \neq g_{e'xz}^y$ has Lebesgue measure 0. We will do this by showing that, assuming $g_{exz}^y > g_{e^<xz}^y$ for some $e, e^<$, we can find alternative realisations of G_z^e that lead to unequal values of the left hand side of Eq (30) without affecting the right hand side.

Let g_{ez}^x and g_z^e be a possible realisation of G_{ez}^x and G_z^e. Assuming $g_{exz}^y > g_{e^<xz}^y$, 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 $g_z'^e = 0.5 g_z^e$ and $g_z'^{e^<} = g_z^{e^<} + 0.5 g_z^e$ and equal to

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

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

because by assumption the denominator remains the same. But then

$$\sum_{e \in E} g^y_{exz} \frac{g^x_{ezz}{}^e}{\sum_{e' \in E} g^x_{e'zz}{}^{e'}} > \sum_{e \in E} g^y_{exz'} \frac{g^x_{ezz'}{}'^e}{\sum_{e' \in E} g^x_{e'z'z'}{}'^{e'}} \tag{31}$$

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

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

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

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

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

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

Thus the conclusion in Eq. (31) follows for the same reasons as before. Considering $g^x_{ez'} < g^x_{e<z'}$, analogous reasoning implies Eq. (31) once again.

Thus, unless $g^y_{exz} = g^y_{e'xz}$ for all e, e' and z, Eq. (30) implies a nontrivial constraint on G^x_{Ez} for some z. Thus for some e, e', z, x and y the set of solutions $S := \{g^X_{EZ} | \mathsf{G}^X_{EZ} = g^X_{EZ}$ satisfies Eq. (30) for all $x, z \wedge g^y_{exz} \neq g^y_{e'xz}\}$ has Lebesgue measure 0 [Okamoto, 1973], and so by domination

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

On the other hand, by assumption, the set $T := \{g^E_z | \mathsf{G}^E_z = g^E_z$ satisfies Eq. (30)$\}$ has measure 1. Thus we conclude that with the exception of a $\mathbb{Q}_\alpha$ measure 0 set, $g^y_{exz} = g^y_{e'xz}$. That is, $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{E}_i | (\mathsf{Z}_i, \mathsf{X}_i, \mathsf{G}, \mathrm{id}_C)$. By contraction with $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}} \mathsf{Z}_i | (\mathsf{X}_i, \mathsf{G}, \mathrm{id}_C)$, we have $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{Q}} (\mathsf{Z}_i, \mathsf{E}_i) | (\mathsf{X}_i, \mathsf{G}, \mathrm{id}_C)$.

By CIIR of the $(\mathsf{E}_i | (\mathsf{X}_i, \mathsf{Y}_i))$ pairs, we have for all i, $\mathbb{Q}_\alpha^{\mathsf{Y}_i \mathsf{X}_i | \mathsf{E}_i \mathsf{G}} \overset{\mathbb{Q}_\alpha^{\mathsf{E}_i | \mathsf{G}}}{\cong} \mathbb{Q}_\alpha^{\mathsf{Y}_c \mathsf{X}_c | \mathsf{E}_c \mathsf{G}}$. Because we have a representative version G^{XY}_E of $\mathbb{Q}_\alpha^{\mathsf{Y}_i \mathsf{X}_i | \mathsf{E}_i \mathsf{G}} forall i \in \mathbb{N}$ (Theorem 4.3) and precedent implies that any set of measure 0 with respect to $\mathbb{Q}_\alpha^{\mathsf{E}_i | \mathsf{G}}$ for all $i \in \mathbb{N}$ also has measure 0 with respect to $\mathbb{Q}_\alpha^{\mathsf{E}_c | \mathsf{G}}$, we have

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

and thus

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

completing the proof. $\square$

C.2 Diverse precedent from independent causal mechanisms

Here we prove Theorem ??.

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

If either of the following hold:

$$\mathbb{P}^{\mathsf{G}^E_z|\mathsf{G}^E_{z'}}(\cdot|g^E_{z'}) \ll U_{\Delta(E)} \text{ almost all } g^E_{z'} \qquad \text{and } \mathsf{G}^E_Z \perp\!\!\!\perp^e_{\mathbb{P}} \mathsf{G}^{XY}_{EZ}|\mathrm{Id}_C \tag{32}$$

$$\mathbb{P}^{\mathsf{G}^X_{Ez}|\mathsf{G}^X_{Ez'}}(\cdot|g^X_{Ez'}) \ll U_{\Delta(X)} \text{ almost all } g^X_{Ez'} \qquad \text{and } \mathsf{G}^X_{EZ} \perp\!\!\!\perp^e_{\mathbb{P}} \mathsf{G}^{\mathsf{Y}}_{\mathsf{EXZ}}|\mathrm{Id}_C \tag{33}$$

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

Proof. If the conditions on line (32) hold, then we require $(\mathsf{G}^E_{z'}, \mathsf{G}^E_z) \perp\!\!\!\perp^e_{\mathbb{P}} (\mathsf{G}^{XY}_{EZ}, 1_I)|\mathrm{id}_C$ for diverse precedent. $\mathbb{1}_I = [\![\mathsf{G}^Y_{Xz} = \mathsf{G}^Y_{Xz'}]\!]$ is determined by $\mathsf{G}^{\mathsf{XY}}_{\mathsf{EZ}}$, so by decomposition it is sufficient to show $(\mathsf{G}^E_{z'}, \mathsf{G}^E_z) \perp\!\!\!\perp^e_{\mathbb{P}} \mathsf{G}^{XY}_{EZ}|\mathrm{id}_C$, but this follows directly from the assumption $\mathsf{G}^{\mathsf{E}}_{\mathsf{Z}} \perp\!\!\!\perp^e_{\mathbb{P}} \mathsf{G}^{\mathsf{XY}}_{\mathsf{EZ}}|\mathrm{Id}_C$.

If the conditions on line (33) hold, then we require $(\mathsf{G}^X_{Ez'}, \mathsf{G}^X_{Ez}) \perp\!\!\!\perp^e_{\mathbb{P}} (\mathsf{G}^{XY}_{EZ}, 1_I)|\mathrm{id}_C$ for diverse precedent. As $\mathbb{1}_I$ is also determined by G^Y_{EXZ}, this follows by an argument analogous to the above. □