Functorial Causal Models

Towards an Algebraic Approach to Causal Inference

by
Dichuan (David) Gao

A Dissertation submitted in partial fulfillment of the
requirements for the Degree of Bachelor of Science
in the Department of Mathematics at Brown University

Providence, Rhode Island, USA
May 2022

Acknowledgements

This project would not have been possible without the amazing support of many people. I am grateful to my advisor, Justin Holmer, who both facilitated the beginning of this project, and gave me invaluable guidance throughout the two-year process until its completion. I would also like to thank Melody Chan and Thomas Goodwillie for sharing their expertise in this project.

The idea of combining causal inference with categorical logic was first introduced to me when Owen Lynch joined our research group. So I am indebted to Owen for this crucial idea, which spans most of the original parts of this thesis. I am also grateful to the students of the Mathematics Department at Brown University with whom I had invaluable discussions about the project. I especially thank Kyle Ferendo and Joe Hlavinka for the hours we spent discussing category theory.

I wish to thank my friends Alvin Fan, Ken Kawamura, Tomoki Yamanaka, Hiroaki Cho, and Kento Suzuki for the many discussions ranging from metaphysics to epistemology to public health. These all helped me organize the mathematics in this thesis into a coherent picture. Also, I am grateful to Alvin, Ken, and Tomoki for lending extra pairs of eyes during the editing of this manuscript.

Finally, I wish to express my deep gratitude for my family who have always supported me materially, mentally, and spiritually: my mother, Mei; my father, Yang; and my significant other, Dai Qing. Thank you for your love.

Contents

1 Introduction

The phrase "correlation is not causation" is frequently repeated in statistics classrooms. Statisticians have a solid grasp on what "correlation" is. And, using a bit of intuition, we know that "causation" is definitely not the same as "correlation" (the reading on my barometer is correlated with the chances of rain. But certainly, my barometer's movement does not cause rain). If causation is not correlation, what is it? In particular, when we make a claim such as "the spark caused the explosion", what are we saying?

One thing we are saying is that, if we were to model the world correctly using mathematical formalism, we should take into account a particular kind of connection - a *causal* connection, between the spark and the explosion. So, an adequate understanding of causation should include, at least, a mathematical formalism for modelling causal connections. This thesis is a survey of the kinds of mathematical formalism for modelling causation. It culminates, in the final chapters, with a proposal for a brand new formalism.

We begin, in chapter 2, with a brief overview of the philosophical discourse about causation. The ideas presented in this discourse underpin much of the mathematical models we develop, and they also serve to illustrate the various challenges that mathematical formalism for causal modelling needs to address. In particular, in this chapter we will see that

1. A successful causal model not only needs to describe what we *do* observe in the world, but also what we *would* observe, *if we were* to intervene, as agents, on certain objects in the world.

2. Causal models are not black boxes, but have *internal structure*. In particular, causal models should have compositionality: they should be built out of smaller units, and they in turn should be capable of building up a larger picture.

3. Differing causal models are not stand-alone ideas, but bear *meaningful relations* to each other. A successful formalism for causal modelling should make explicit these meaningful relations between models.

We will see, in subsequent chapters, that these requirements are gradually fulfilled by more and more complex mathematical formalisms.

In chapter 3, we review the formalism known as the potential outcomes framework, or Rubin Causal Models (RCM). We will see that RCMs are very useful in clinical settings, in which there is exactly one cause (a treatment), and one effect (an outcome) to be modelled. They fulfill requirement number 1 above, but not 2 and 3.

In chapter 4, we review the formalism known as Structural Causal Models (SCM). We explain the sense in which SCMs are really generalizations of RCMs, and discuss the major strengths of the SCM approach. In particular, we will see that they fulfill requirements 1 and 2, but they do not fully meet requirement 3.

Then, in chapter 5, we take a brief detour, and survey some core concepts in category theory. These concepts serve as groundwork for chapter 6. Readers

who are comfortable in the topic of higher category theory should probably skip this chapter.

Finally, in chapter 6, we develop our proposal for a new formalism: Functorial Causal Models (FCMs). We make clear that this formalism is a stronger, if more cumbersome, formalism than the SCM approach. In particular, FCMs satisfy all three requirements.

Chapters 2, 3, 4, and 5 are surveys of well-established literature. Chapter 6 takes abstract ideas from Evan Patterson's foundational thesis [15], but the definitions and theorems in this chapter are mostly original.

2 Philosophical Theories of Causality

2.1 Hume's Regularity Theory of Causation

In *A Treatise of Human Nature*, David Hume defines causation as

> an object precedent and contiguous to another, and where all the objects resembling the former are plac'd in like relations of priority and contiguity to those objects, that resemble the latter. [7]

We may summarize the Humean theory of causation as follows:

Definition 2.1. (Humean Regularity Theory) An object A is a cause of another object B if:

1. A precedes B in time; and
2. A is spatiotemporally contiguous to B, that is to say, there is a time-like path connecting A to B; and
3. For any object A' of the same type as (i.e. "resembling") A, there is a corresponding object B' of the same type as B, which follows A' temporally and is spatiotemporally contiguous to A'. (This condition is called the regularity condition).

Notice that, in HRT, there is no mention whatsoever of an **underlying mechanism** through which regularity occurs: if causation exists, there need not be an explanation for why this regularity occurs. Thus, regularity is not "metaphysically thick": it is only a statistical property of what happens to obtain in the actual world, requiring no notion of necessity or explainability to back it up.

There is a very good reason for Hume to insist on this independence from metaphysical necessity. As an empiricist, Hume held that all ideas come strictly from our sensory perception. But there can be no sensory perception of *necessity*: if I observe that a billiard ball runs into another, and the second ball moves, I have only perceived this chain of events, and not the *necessity* that the second ball moves after the first ball runs into it.

The HRT has several advantages. First, it avoids any commitment to a metaphysical account of necessity, or of any sort of underlying causal mechanism. Second, it explains how we, as experimenters, can perceive causation. Indeed, since temporal precedence and spatiotemporal contiguity are both easy to perceive, to observe a case of Humean causation, one needs only to have the ability to observe regularity. This is easier to defend than, say, a nativist account, where we are born with certain notions of necessity or causal mechanisms.

However, the HRT has three serious defects. First, if the regularity condition is to be well-defined, we must partition "objects" into "types", or relations of

"resemblance". It is not immediately clear how we can do this without being overly subjective.[1]

Second, our everyday notion of causation doesn't seem to require regularity. This can be seen in any case where, while there is clearly causation in the usual sense in which we use the word, there cannot be any other event even remotely resembling this particular occurrence. Take, for example, the claim that the big bang caused the cosmic microwave background to be as it is. Many readers will believe this claim. But they certainly will not expect to be able to observe another occurrence resembling the big bang, and thereby verify that in that occurrence, a cosmic microwave background resembling ours to appear.

Third, the HRT criteria don't seem to be sufficient for our everyday notion of causation.

Example 2.1. (**Fork**) Every time the needle on my barometer points to a lower-than-usual number, it begins to rain in the next hour or so. Then the movement of the needle temporally precedes the rain, is spatiotemporally contiguous to the rain, and this relation obtains regularly. Should I conclude that the movement of the needle causes rain?

It would certainly be absurd to say that the movement of the needle causes rain. We know that it is a drop in atmospheric pressure which causes both the barometer needle to drop, and the rain; it just happens to be the case that the barometer reacts faster than the rain does. The HRT fails to exclude such cases of spurious causation.

To sum up, the HRT presents us with three challenges: first, the problem of partitioning particular events into "types" or relations of "resemblance"; second, the problem that regularity is not necessary for causation; and third, the problem that regularity is not sufficient for causation.

2.2 Counterfactual Theories of Causation

Counterfactual theories of causation (CTC) attempt to address some of the problems faced by regularity theories. The main ingredient of these theories is the idea of "counterfactual dependence". In CTC, causation is defined in terms of chains of counterfactual dependences. The bulk of the classical CTC is developed by David Lewis [12].

Consider the following pairs of sentences, taken from the Lewis text [12]:

1. (a) If Oswald didn't kill Kennedy, someone else did.

   (b) If Oswald hadn't killed Kennedy, someone else would have.

[1] John Stuart Mill refines HRT by defining such a partition using the notion of "laws of nature". According to Mill, a scientific proposition p is a law of nature if, for every deductive system S of scientific facts that strikes an appropriate balance between simplicity and strength, p obtains as either an axiom or a theorem of S [14]. Since types come prepackaged in the statement of a law of nature, two objects have "resemblance" if and only if they are both instantiations of the same type in a relevant law of nature. However, this theory itself has been criticized as being overly mind-dependent, in virtue of its reliance on "appropriate balance". See [1, 20, 2]

2. (a) If Kangaroos don't have tails, Kangaroos topple over when they hop.
   (b) If Kangaroos didn't have tails, Kangaroos would topple over when they hop.

The sentence 1a is true as a matter of logic: Kennedy was, in fact, killed. So I can make the logical deduction that someone, either Oswald or someone else, killed Kennedy. But 1b is not merely a matter of logic. It contains some extra information about the situation at the time just before Kennedy was killed, such that one might infer from 1b that Kennedy was rather hated by a sufficiently large group of people. The second pair of sentences extracts this distinction a bit further. Sentence 2a is vacuously true. Because Kangaroos do in fact have tails, the sentence "if Kangaroos don't have tails, then P" is true for all P. But sentence 2b is not vacuous at all. It is a biomechanical claim about the way in which Kangaroos balance themselves when they hop, such that this balancing is necessarily dependent on having a tail. We call sentences of the form 1b and 2b "counterfactual sentences":

Definition 2.2. A counterfactual sentence is a sentence of the form "if it were the case that A, then it would be the case that B", where A and B are events. We denote this relation by

$$A \mathrel{\square\!\!\rightarrow} B$$

Note, the event A need not be false for the counterfactual sentence to be well-formed. That is, a counterfactual sentence need not be counter-factual.

Definition 2.3. An event B is said to **counterfactually depend** on A if

$$\neg A \mathrel{\square\!\!\rightarrow} \neg B.$$

It is important to note here that counterfactual *dependence* is a counterfactual *sentence* involving the *negations* of the events under question.

Example 2.2. The sentence "if kangaroos didn't have tails, kangaroos would topple over when they hop" is equivalent to "kangaroos' ability to balance themselves counterfactually depends on their having tails".

How are we to provide semantics to this binary operation between events? Consider sentence 2b again. One may reasonably think that this sentence means something to the effect of "imagine a world where Kangaroos don't have tails - in that world, they would topple over when they hop". But of course, this need not hold true in every possible world where kangaroos don't have tails. There may well be a possible world where kangaroos use wings to balance themselves. So, what we mean by sentence 2b is this: in a world where, although kangaroos don't have tails, other aspects of the world are suitably similar to our actual world, kangaroos would topple over [12].

But we also cannot limit ourselves to the possible world where kangaroos don't have tails, but all else remain exactly the same. For example, are we to imagine a possible world where kangaroos don't have tails, but the trails left

behind when they move about are the same? If so, then we must imagine some other mechanism through which the trails are left behind, making that world still dissimilar to our actual world. Thus, we want the sentence $A \boxdot\!\!\rightarrow B$ to mean this: among all possible worlds where A is true, the ones where B is also true are the closest to the actual world. We formalize this notion in the following definitions, which are my own representation of Lewis' work:

Definition 2.4. A **similarity-ordered-multiverse** is a partially ordered set $(\Omega, \leq)$ with a unique minimal element ω^*. The minimal element is called the **actual world**, and elements of Ω are called **possible worlds**. If $\omega_1 \leq \omega_2$, we say ω_1 is more similar to the actual world than ω_2 is.

The idea is that a counterfactual sentence $A \boxdot\!\!\rightarrow B$ is true in the actual world if and only if, among all the possible worlds where A is true, the closest ones are the ones where B is also true:

Definition 2.5. (Lewis' Semantics for Counterfactuals). Let $(\Omega, \leq, \omega^*)$ be a similarity-ordered-multiverse. Let A and B be subsets of Ω. Then $A \boxdot\!\!\rightarrow B$ is true if and only if:

1. A is empty; or
2. There exists a subset $D \subset A \cap B$ such that, for any $\omega \in A \cap B^c$, there exists some $\omega' \in D$ such that $\omega' < \omega$.

Example 2.3. If similarity could be measured in terms of real numbers - that is, if Ω is given the structure of a metric space (and the partial order determined by distance from the point ω^*), then $A \boxdot\!\!\rightarrow B$ is true if and only if either A is empty, or there exists some closed disk $D = \mathcal{B}(\omega^*, r)$ around ω^* such that $D \cap A \cap B \neq \emptyset$, but $D \cap A \cap B^c = \emptyset$.

Now, one may be tempted to say that causality is equivalent to counterfactual dependence in the above-defined sense. However, this can't be right. Consider the following example:

Example 2.4. (**Early Pre-emption**) Suppose Sarah throws a rock at a window. Taro, who was about to do the same with a much bigger rock, sees this, and gives up on throwing his rock. Sarah's rock breaks the window. We would certainly say that Sarah's rock throwing caused the window to break. But if Sarah hadn't thrown the rock, Taro would have, and thus the window would still break. So the breaking of the window does not counterfactually depend on Sarah's rock throwing.

So it would seem that counterfactual dependence is not a necessary condition for causation. Lewis proposes that causation should be defined as chains of counterfactual dependence - that is, causation is the transitive closure of counterfactual dependence relations. So, in example 2.4, we may add an intermediate event - the event of Sarah's rock flying in midair toward the window. Now this event counterfactually depends on Sarah's throwing. But also, by the

time the rock is in midair, Taro had already decided not to throw his rock. So the shattering of the window also counterfactually depends on the intermediate event. Thus, there is a chain of counterfactual dependencies from Sarah's throw to the shattering of the window.

The advantage of this account of causation is that it resolves many issues found in the HRT. First, there is no longer any need to partition particular occurrences in the actual world into "types" or "resemblances", since the only comparisons being made in the CTC are those between possible worlds in a similarity-oriented-multiverse. Second, there is no difficulty for the CTC to assert that the big bang caused the cosmic microwave background to be as it is. Again, this is because the only comparisons being made are those between possible worlds; and readers can imagine possible worlds where, say, the big bang occurred but in a slightly different way. Third, the CTC successfully excludes cases of fork from causation. Consider example 2.1. There is a chain of counterfactual dependence from the drop in atmospheric pressure to the movement of the barometer needle, but there is no such chain connecting the movement of the needle to the rain.

However, the CTC in turn also suffers from two important issues. First is the problem of transitivity. Since Lewis defines causation as chains of counterfactual dependence, so causation must be transitive: if A causes B, and B causes C, then A causes C. But many counterexamples to transitivity have been given (Lewis himself gives a catalogue of these in [13]). Consider this classic scenario:

Example 2.5. (**Non-Transitivity of Causation**) A hiker is on her way up the mountain. A boulder rolls down the mountain toward the hiker (call this event B). Seeing the boulder, the hiker ducks to avoid the boulder (call this event D). Having ducked successfully, the hiker is able to continue her hike (call this event C). Further, suppose that the boulder was on a trajectory such that, if the hiker had not ducked, she would have been struck and become incapacitated. Then B causes D, and D causes C, but B certainly does not cause C: the boulder rolling towards the hiker cannot possibly be a cause for the hiker's ability to continue the hike.

The second problem CTC suffers from is the problem of late pre-emption. We have seen that, by defining causation as chains of counterfactual dependence, rather than as counterfactual dependence itself, Lewis was able to resolve the problem of early pre-emption. But consider a similar case of pre-emption, which cannot be resolved by the chain modification:

Example 2.6. (**Late Pre-Emption**) Suppose Sarah and Taro both simultaneously throw a rock at a window. Sarah's rock reaches the window only a split-second before Taro's rock does. Sarah's rock breaks the window, and Taro's rock sails through the broken window only a split-second after. We would certainly say that Sarah's rock throwing caused the window to break. But this time, we cannot construct a chain of counterfactual dependence leading from Sarah's rock throwing to the window breaking, since there is no moment when Sarah's rock is in midair, and Taro's rock has been laid down.

So to sum up, the CTC addresses the main weaknesses of the HRT. However, it suffers from the problem of non-transitivity, and the problem of late pre-emption. As we will see in the concluding section of this thesis (section 7.2), the mathematical viewpoint will provide us with something of an answer to these two problems - but it is arguably an unsatisfactory one.

2.3 Towards a Mathematical Point of View

The most common statistical methods of modelling causation today are, in many ways, direct descendants of the CTC. However, they differ from our philosophical theories in a few key ways. In particular, there are a few key notions central to the mathematical theory of causal models, which are not present in the philosophical account of counterfactual causation. We introduce these notions here, not only as prerequisites for the following chapters, but also as philosophically interesting ideas in their own rights.

First, instead of reasoning about **events**, we will reason about a more general notion, known as **random variables**. An event is an unknown true-or-false value: it either occurs or it doesn't. On the other hand, a random variable is an unknown value of any number of possible outcomes (in particular, a random variable whose range consists of two outcomes is mathematically indistinguishable from an event). I will not discuss the various equivalent definitions of random variables. For a standard treatment, see Sheldon Ross's book [18].

Second, we will be dealing with the class of statistical models known as **generative models**. A generative model consists of a mathematical description of some process which *generates* data. For example, the heliocentric model places the sun at the center of the solar system, thereby describing the process of motion of the sun and its planets, which in turn generates the patterns of ecclesiastical motion that we can actually observe here on earth. Thus, a generative model is concerned with more than description and categorization of the data; it asks for the underlying process.

A generative model can, in principle, be computerized. That is to say, it is in principle possible to write a computer program, such that outputs of this program behave as the generative model specifies. For example, we can write a computer simulation of the solar system, in accordance with the heliocentric model. Outputs of this simulation will consist of time-series of positions for the sun and its planets. From these outputs, we may then retrieve the patterns of motion that would have been observed from earth, if the sun and planets really do move as the outputs specify. If the model is good, we expect these predicted patterns of motion to be similar to the patterns of motion that are actually observed. Note, I say that this is possible *in principle*, because limitations on computational capacity should not prevent a mathematical model from being generative in nature, even though they may prevent the actual programmatization of such a model.

A full mathematical description of such a generative model will be referred to as the **noumenal content** of the model. This will stand in contrast to the **empirical content** of the model. A generative model is empirically falsifiable

when some (or all) of the random variables in the model are things that we can measure in the real world. For example, the positions of planets relative to earth can be measured in observatories, therefore making the heliocentric model falsifiable. Something that makes a model empirically falsifiable is called an empirical content of the model. In general, the empirical content of a model supervenes on the noumenal content. Not all noumenal content is observable, but certainly all observable components of a model are consequences of the mathematical description of that model.

The empirical content of a causal model can be roughly divided into two categories. First, causal models contain **observational content**. These are things which can be measured and falsified just by observing the system that this model is meant to represent. For example, the patterns of motion of planets would be part of the observational content of the heliocentric model. Second, causal models contain **interventional content**: a generative causal model tells us what to expect when we, as experimenters, intervene and disrupt the causal flow of things in a certain way. For example, suppose we had enough explosives to blow Mars into pieces. The heliocentric model would tell us something about how the solar system would behave, if we chose to do so. This is in principle falsifiable: if we do choose to blow up Mars, and the solar system doesn't behave as the heliocentric model tells us it would, then we know the model is wrong.

In the rest of this thesis, we will always describe causal models in the following order: 1) noumenal content, 2) observational content, 3) interventional content.

The concept of counterfactual dependence in the CTC will be replaced by a similar but subtly different notion of **generative dependence**. A random variable B generatively depends on A in a model M if, in order to compute a value for B in accordance with M, we required the value of A. For example, in the heliocentric model, if B is the position of the earth at time $t+1$, then B would generatively depend on the earth's position and velocity at time t, as well as on the position of the sun and every other planet at time t. Whenever a variable B generatively depends on A, we say that A is a **generative parent** (or just **parent** for short) of B. These generative dependence relations can then be composed into chains and networks, echoing the transitivity of causal influence in the CTC. A generative model, then, is a composite structure of these generative dependence relations.

The task of *finding* a generative model that suits some given observational data is quite difficult. Because a generative model is concerned with the underlying process, and not just with a mere description of the data, so it is in general impossible to "read off" a good model from the data itself. Instead, it is necessary to instantiate candidate models, compute the consequences of the models, and see whether the results fit our observational data. It is therefore useful to keep track of the *relations* between all the different candidate models, so that progress can be made. In practice, of course, this can be done by the exchange of prose between researchers who have personal expertise in the subject matter being modelled. However, from a mathematical point of view,

it is beneficial to also keep track of these relations in a *formal* manner, so that definitions, conjectures, and proofs can be made about the relations between models.

Thus, a good mathematical notion of causal models should satisfy three requirements:

1. A causal model should specify, quantitatively, the generative dependence between some cause-effect pairs of variables. As we will see in chapter 3, Markov kernels are good mathematical structures for specifying these generative dependencies.

2. A causal model should specify a composite structure of generative dependencies. It should tell us how, in the phenomenon being modelled, generative dependencies come together to form chains and networks. As we will see in chapter 4, directed acyclic graphs are the tools to use for specifying these composite structures.

3. Differing causal models should admit some formal notion of relations between them. As we will see in chapter 6, functors and natural transformations are good tools for thinking about these relations between causal models.

Let us now dive into the mathematics of causal models, and see to what extent these requirements can be met.

3 Potential Outcome Models

Potential Outcome Causal Models, also known as Rubin Causal Models (RCM), is in a sense the direct mathematical descendant of the counterfactual theory of causation. This framework makes computationally rigorous the meaning of a counterfactual/generative dependence of events, and provides fertile ground for statistical estimation of causal effects. In this section, we will roughly follow Imben's treatment of this model [8], and Rubin's own treatment [9].

Before we begin discussing this framework, let us look at the archetypal situation in which the RCM would be useful. This example will continue to run throughout this thesis, so it is very important that we get a good grasp of it here.

Example 3.1. (**Accupill**) Suppose Amy, Bob, Carlos, and Dylon are sick from COVID-19. Pfizer is testing a new pill, the Accupill, which is intended to help patients recover from COVID-19. Amy and Bob are given the Accupill, while Carlos and Dylon are given placebos. Amy, Bob, and Carlos recover within 5 days, while Dylon does not. However, we do not know whether Amy and Bob would have recovered had they not been given the Accupill, nor do we know whether Carlos and Dylon would have recovered had they been given the pill. The situation is summarized in table 1.

Name	Treatment	Outcome if Accupill	Outcome if Placebo
Amy	Accupill	Recovered	?
Bob	Accupill	Recovered	?
Carlos	Placebo	?	Recovered
Dylon	Placebo	?	Non Recovered

Table 1: RCM Table for the Accupill Example

Did Accupill help Amy recover? To answer this question, we must find out what the outcome would have been, if Amy were given the Placebo instead. So, the question of finding the causal effect of Accupill is essentially a missing data problem.

3.1 Noumenal Content of Rubin Causal Models

We formalize the Accupill example (example 3.1) as follows:

Definition 3.1. (**RCM**) A Rubin Causal Model consists of the following data:

- A fixed integer N, known as the **sample size**;
- A fixed set $\mathscr{T}$, known as the **set of available treatments**;
- For each $i = 1, \ldots, N$, a random variable X_i, known as the **pre-treatment features of the individual** i;

- A random variable $T = (T_1, \ldots, T_N)$ whose range is $\mathscr{T}^N$, known as the **factual**[2] **treatment regime**, with each $T_i \in \mathscr{T}$ being the **factual treatment on individual** i;
- For each $t \in \mathscr{T}^N$, a random variable $Y(t) = (Y_1(t), \ldots, Y_N(t))$ whose range is $\mathbb{R}^N$, known as the **potential outcome if regime** t **were given**, with each $Y_i(t) \in \mathbb{R}$ being the **potential outcome for individual** i.

In addition, we define the composite random variable $Y := Y(T) = (Y_1(T), \ldots, Y_N(T))$ to be the **factual outcome**. In other words, the factual outcome is the same as "the potential outcome, if the factual treatment were given".

Example 3.2. In the Accupill example (example 3.1), we would have:

- $N = 4$, since there are 4 individuals;
- $\mathscr{T} = \{0, 1\}$ with 0 representing placebo, and 1 representing Accupill;
- The actualized value x_1 of the random variable X_1 represents Amy's clinical features; and likewise for $x_2, ..., x_4$.
- The actualized value t of T is $(1, 1, 0, 0)$;
- The actualized value $y(1, 1, 0, 0)$ of $Y(1, 1, 0, 0)$ is $y(1, 1, 0, 0) = (1, 1, 1, 0)$ (where now 1 represents recovery and 0 represents non-recovery). This is also the actualized value of the factual outcome $Y(T)$. For all other $t' \neq (1, 1, 0, 0)$, the random variable $Y(t')$ is not actualized.

In practice, quite a few additional key assumptions need to be made before we can do statistics on an RCM. We now describe these assumptions.

Definition 3.2. (**Single World Assumption**) An RCM is said to satisfy the single world assumption if the following sentence is true: if T is actualized to $t \in \mathscr{T}^N$, then for every $t' \neq t$, the random variable $Y(t')$ is not actualized. In other words, an outcome that is actualized must be the factual outcome.

Definition 3.3. (**Non-Interference Assumption**) An RCM is said to satisfy the non-interference assumption if the potential outcome for an individual i depends only on the treatment received by individual i, and not those received by anyone else. Symbolically, this is equivalent to the following requirement:

$$\text{For each } 1 \leq i \leq N \text{ and for each } t, t' \in \mathscr{T}^N, \text{ if } t_i = t'_i, \text{ then } Y_i(t) = Y_i(t').$$

If the non-interference assumption is satisfied, then it makes sense to speak of the potential outcome for individual i if treatment t_i were given to her, without regard to the treatment given to other individuals. So we will use $Y_i(t_i)$ to denote $Y_i(t)$. So in example 3.2, we can say $Y_1(1) = 1$.

[2]It is crucial to distinguish between our use of the words "factual" and "actual". Actualization will always refer to the actualization of a random variable, and is opposed to "random". But "factual" refers only to the precise sense shown in this definition, opposed to "counterfactual".

Definition 3.4. (**Homogeneity**) Given that an RCM satisfies the non-interference assumption, it is further said to be homogeneous if each individual is drawn from the same population. Symbolically, this means that for each fixed treatment t, the joint variable $(X_i, T_i, Y_i(t))$ for $i = 1, \ldots, N$ are i.i.d.

If homogeneity is satisfied, then for each t, we define $(X, T, Y(t))$ to be a random variable drawn from the same distribution as each $(X_i, T_i, Y_i(t))$. It then makes sense to ask for the value

$$\mathbb{E}[Y(t) \mid X = x]$$

for any pre-treatment feature x. This will be the crucial piece of information needed for decision making: given this patient has characteristic x, which treatment gives the best expected outcome?

Definition 3.5. (**Weak Unconfoundedness**) Given an RCM satisfying non-interference, the RCM is furthered called weakly unconfounded if the information contained in X_i "covers" all the confounding factors between the treatment and the outcome on i. Symbolically, for all $i = 1, \ldots, N$ and all $t_i \in \mathscr{T}$,

$$D_i(t_i) \amalg Y_i(t_i) \mid X_i$$

where $D_i(t_i)$ is the indicator variable

$$D_i(t_i) = \begin{cases} 1 & T_i = t_i \\ 0 & \text{otherwise.} \end{cases}$$

These assumptions are common, although in empirical studies it will be important for the researcher to actually check whether these assumptions are reasonable ones.

3.2 Observational Content of Rubin Causal Models

Which variables in an RCM are observed? In general settings, the feature vectors X_i, the factual treatment regime T, and the factual outcome Y are observed. But the counterfactual outcomes $Y(t)$ for $t \neq T$ are not observed. Thus, the observational content of RCMs at least include X, T, and Y. However, it includes a little more than that.

If an RCM satisfies non-interference, homogeneity, and weak unconfoundedness, then expected values of counterfactual outcomes can be observationally approximated by stratifying a large enough population, and observing the actualized treatments and outcomes there. Conversely, if observed treatments and outcomes within a properly stratified population do not approximate the causal effects indicated by an RCM, then the RCM is wrong. This is formalized in the following theorem:

Theorem 3.1. *If an RCM satisfies non-interference, homogeneity, and weak unconfoundedness, then for any treatment $t^* \in \mathscr{T}$ and any feature vector x such that*

$\mathbb{P}\{T = t^*, X = x\} > 0$*:*

$$\frac{1}{|\{i : T_i = t^* \text{ and } X_i = x\}|} \sum_{i=1}^{N} Y_i \cdot \mathbb{I}_{T_i = t^* \text{ and } X_i = x} \xrightarrow{a.s.} \mathbb{E}[Y(t^*) \mid X = x]$$

as $N \to \infty$*. Here* X *and* $Y(t^*)$ *are random variables instantiated to be i.i.d. with each of* $Y_i(t^*)$ *and* X_i*.*

Proof. This will be a direct consequence of the following two lemmas. □

Lemma 3.1. *If an RCM satisfies non-interference and weak unconfoundedness, then for each* i*, each treatment* $t^* \in \mathscr{T}$*, and each possible feature vector* x*,*

$$\mathbb{E}[Y_i(t^*) \mid X_i = x] = \mathbb{E}[Y_i \mid T_i = t^*, X_i = x].$$

where $Y_i := Y_i(T_i)$*.*

Proof.

$$\begin{aligned}
\mathbb{E}[Y_i(t^*) \mid X_i = x] &= \mathbb{E}[Y_i(t^*) \mid T_i = t^*, X_i = x] \qquad \text{(weak unconfoundedness)} \\
&= \mathbb{E}[Y_i(T_i) \mid T_i = t^*, X_i = x] \\
&= \mathbb{E}[Y_i \mid T_i = t^*, X_i = x].
\end{aligned}$$

□

Lemma 3.2. *If an RCM satisfies non-interference and homogeneity, then for each* i*, each treatment* $t^* \in \mathscr{T}$*, and each possible feature vector* x*, if* $\mathbb{P}\{T = t^*, X = x\} > 0$*, then*

$$\frac{1}{|\{i : T_i = t^* \text{ and } X_i = x\}|} \sum_{i=1}^{N} Y_i \cdot \mathbb{I}_{T_i = t^* \text{ and } X_i = x} \xrightarrow{a.s.} \mathbb{E}[Y \mid T = t^*, X = x]$$

as $N \to \infty$*.*

Proof. By homogeneity, for every t^*, $(X_i, T_i, Y_i(t^*)) \overset{i.i.d.}{\sim} (X, T, Y(t^*))$. Since the factual outcome is defined as $Y_i = Y_i(T_i)$, so the observed factual variables will satisfy $(X_i, T_i, Y_i) \overset{i.i.d.}{\sim} (X, T, Y)$.

A direct application of the strong law of large numbers then yields the statement of the lemma. □

3.3 Interventional Content of Rubin Causal Models

In some sense, the interventional content of RCMs is as simple as can be. The variables $Y_i(t_i)$ directly describe the outcome that we would observe if we, as experimenters, intervened and gave the ith individual treatment t_i. This is empirically obtainable using randomized trials:

Theorem 3.2. *If an RCM satisfies non-interference and homogeneity, then for each treatment t^*,*

$$\frac{1}{N}\sum_{i=1}^{N} Y_i(t^*) \xrightarrow{a.s.} \mathbb{E}[Y(t^*)]$$

as $N \to \infty$. Furthermore, for each possible feature vector x such that $\mathbb{P}\{X = x\} > 0$,

$$\frac{1}{|\{i : X_i = x\}|}\sum_{i=1}^{N} Y_i(t^*) \cdot \mathbb{I}_{X_i = x} \xrightarrow{a.s.} \mathbb{E}[Y(t^*) \mid X = x]$$

as $N \to \infty$.

Proof. Both statements are consequences of the strong law of large number, given that $(X_i, T_i, Y_i(t^*)) \overset{i.i.d.}{\sim} (X, T, Y(t^*))$. □

Notice that the convergence in theorem 3.2 occurs faster with respect to N than does the convergence in theorem 3.1. This is because, when we perform stratification in an observational study, the size of the sample within each strata $|\{i : X_i = x \text{ and } T_i = t\}|$ becomes small, and thus the total sample size needs to be large. On the other hand, when we perform a randomized trial, the size of the sample is either N or $|\{i : X_i = x\}|$, which is larger. Hence the capacity to perform randomized experiments is extremely useful.

3.4 Potential Outcomes as Markov Kernels

As part of our chapter on RCMs, I want to take a brief detour, in order to develop the idea of a Markov kernel, which will become very useful in the next chapters. Although this idea is not usually associated with RCMs, I believe that the notation used to describe the noumenal content of RCMs is prototypical of Markov Kernels.

In an RCM, for each treatment $t \in \mathscr{T}$, the potential outcome is a random variable $Y(t)$. This evokes the idea that $Y(-)$ is some kind of morphism: a function which takes input in $\mathscr{T}$, and which outputs random variables. Let's make this a little more concrete.

If, in a programming language like Python, we want to write a function that takes a value $t \in T$ and spits out a value $y \in Y$, where T and Y are sets, it is possible that the computation of the function can involve some calls to random number generators. In this case, the Python function would not, in the mathematical sense, be a function: for each input t, we do not always get the same output y. However, for each input t, we do always obtain output y according to some probability distribution. The probability distribution is completely determined by t. This randomness-involving function plays the role of $Y(-)$ in RCMs (although it has a few extra assumptions that RCMs in general do not require). Such randomness-involving functions are known as Markov Kernels, and are formalized in the following:

Definition 3.6. (Markov Kernel) A **Markov kernel** M **from** T **to** Y, where $(T, \mathscr{A}_T)$ and $(Y, \mathscr{A}_Y)$ are measurable spaces, is a measurable function $M : T \to Prob(Y)$, where $Prob(Y)$ is the space of all possible probability measures on Y.

Alternatively, if you prefer a more bare-bones definition, the above can be equivalently formulated as

Definition 3.7. (Markov Kernel, alternative definition) Given T, Y measurable spaces equipped with σ-algebras $\mathscr{A}_T, \mathscr{A}_Y$, a **Markov kernel** M **from** T **to** Y is a function $M : T \times \mathscr{A}_Y \to [0, 1]$, such that:

1. For each $t \in T$, the map $M(t, -) : \mathscr{A}_Y \to [0, 1]$ is a probability measure;
2. For each event $C \in \mathscr{A}_Y$, the map $M(-, C) : T \to [0, 1]$ is measurable with respect to $\mathscr{A}_T$ and the Borel σ-algebra on $[0, 1]$.

That the above two definitions are equivalent is verified by the definition of $Prob(Y)$ as a measurable space.

Example 3.3. The family of normal distributions $\mathcal{N}$ is a Markov kernel

$$\mathcal{N} : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}.$$

Given a mean $\mu \in \mathbb{R}$ and a variance $\sigma^2 \in \mathbb{R}_{>0}$, we obtain a probability measure $\mathcal{N}(\mu, \sigma^2)$ on $\mathbb{R}$.

Example 3.4. Let I be the measurable space of a single point, and let X be any measurable space. Then a Markov kernel $I \to X$ is nothing more or less than a probability measure on X.

Example 3.5. In an RCM satisfying non-interference, if, for each $t \in \mathscr{T}$, the random variable $Y_i(t)$ has the same range $\mathscr{Y}$, then $Y_i(-)$ is a Markov Kernel from $\mathscr{T}$ to $\mathscr{Y}$, where $\mathscr{T}$ is given the discrete σ-algebra. This is because a random variable is nothing but a measurable space (its range) together with a probability measure.

Going forward, we will usually denote a Markov kernel from a space T to a space Y by the symbol $\mathbb{P}_{Y|T}$, to emphasize that it is something which allows you to retrieve a probability distribution on Y, given a value in T. The evaluation of this Markov kernel at an event $C \in \mathscr{A}_Y$ and at a point $t \in T$ will be denoted $\mathbb{P}_{Y|T}(C \mid t)$.

Now, it is frequently convenient to work with probability density functions rather than the probability measures themselves. We know that a probability density function is defined by a probability measure with respect to some fixed "intrinsic" measure on the underlying space (say, the Lebesgue measure on Euclidean space). We develop the equivalent notion for Markov kernels:

Definition 3.8. Let $\mathbb{P}_{Y|T} : T \to Y$ be a Markov kernel, and let μ be a measure on Y. A μ**-conditional density function** of $\mathbb{P}_{Y|T}$ is a function $f_{Y|T} : Y \times T \to [0, 1]$ such that, for all $t \in T$ and all event $E \in \mathscr{A}_Y$,

$$\int_E f_{Y|T}(y \mid t)\, \mu(dy) = \mathbb{P}_{Y|T}(E \mid t).$$

The Lebesgue-Radon-Nikodym Theorem tells us that, when a μ-conditional density function exists for $\mathbb{P}_{Y|T}$, it is μ-almost-everywhere unique. So it makes sense to talk about *the* μ-conditional density function of $\mathbb{P}_{Y|T}$. In almost all the cases in the subsequent chapters, μ will be either a counting measure on a discrete space, or the Lebesgue measure on a Euclidean space. However, example 3.6 shows that not all Markov kernels are absolutely continuous. To obtain the capacity to work with Markov kernels that do not admit conditional density functions, we will require the infrastructure developed in chapters 5 and 6.

Example 3.6. (A Markov Kernel without Density Function) Suppose $P_{S|W}$ is a Markov kernel that models the number of walking steps a patient takes in any given minute in a day. The domain of the kernel is $W = \{0, 1\}$, where 0 denotes that the patient is asleep, and 1 denotes that the patient is awake. When the patient is awake, the number of walking steps she takes is just given by an exponential distribution E. So $P_{S|W}(- \mid 1) = E$. On the other hand, when the patient is asleep, she is either completely motionless, or she is sleep walking. So $P_{S|W}(- \mid 0) = (1 - p)\delta_0 + pE$ where $0 \leq p \leq 1$ is the probability that the patient is sleepwalking, and δ_0 is the Dirac distribution at 0. So if $p \neq 1$, then the kernel $P_{S|W}$ does not have a density function with respect to the Lebesgue measure on $\mathbb{R}_{\geq 0}$.

This concludes our brief detour on the topic of Markov kernels. This topic will become central in the subsequent chapters of this thesis, so readers may find it helpful to refer back to this section as you read the following chapters.

3.5 Concluding Remarks

As our discussion in this chapter has shown, RCMs are extremely versatile in cases where one causal pair (one treatment, one effect) is being examined. Plenty of statistics can be done because of the relatively small number of variables in the model, and because of the well-developed set of assumptions that can frequently be made. However, recall from section 2.2 that causal effects should, more often than not, be composable. A main philosophical difficulty with RCMs is that they don't teach us how to deal with chains and networks of causal effects. Complex systems such as the ecosystem and the stock market have many causes and many effects linked to each other, and the study of systems like these will require what is known as structural causal models, which we will discuss in chapter 4.

4 Structural Causal Models

Structural Causal Models address the core weakness of Rubin Causal Models: with SCMs, we will be able to model chains and networks causal effects. In other words, SCMs will incorporate the element of Lewis's Counterfactual Theory of Causation that RCMs fail to incorporate: the compositionality of causation.

In this chapter, we begin in section 4.1 by laying out the necessary graph-theoretical knowledge about directed acyclic graphs. Then, in section 4.2, we define the noumenal content of structural causal models, in accordance with Judea Pearl's treatment in [6]. In section 4.3, we continue to follow Pearl's theory and describe the observational content of SCMs, and in particular, we describe d-separation as the purely structural component of that observational content. Finally, in section 4.4, we describe the interventional content of SCMs via the operation known as the Single World Intervention Graph (SWIG), following Richardson et. al. [16]. Pearl has an equally powerful formulation of the interventional content of SCMs; however, the SWIG formulation lends easier into the category theoretic formulation that we will develop in chapter 6, and therefore we have chosen to introduce it over Pearl's formulation.

4.1 Directed Acyclic Graphs

Directed Acyclic Graphs, often referred to as DAGs, are used to represent all kinds of computational relations, whether these relations are causal or otherwise. The idea is simple: a directed acyclic graph is a graph (i.e. a set of vertices and edges connecting those vertices), where the edges have directions (they point from one vertex to another), and where, if one were to walk along the edges, one could never go in a circle.

But to reason with such structures, we need to develop some formal language. Let's begin by defining a directed graph.

Definition 4.1. A **directed graph** G consists of sets $V(G)$ and $E(G)$, together with functions

$$E(G) \underset{t}{\overset{s}{\rightrightarrows}} V(G)$$

The set $V(G)$ is called the set of vertices (thought of as dots), and the set $E(G)$ is called the set of edges. The function s picks out the source of an edge, and the function t picks out the target. So an edge $e \in E(G)$ is thought of as an arrow pointing from its source $s(e)$ to its target $t(e)$. Indeed, if $s(e) = v$ and $t(e) = w$, we will use the phrase $e : v \to w$ as a shorthand for the assertion that "$s(e) = v$ and $t(e) = w$".

Definition 4.2. Given a graph G, a **(directed) path** γ in G is a sequence of edges $\{e_i\}_{i=1}^n$ such that each e_{i+1} begins where e_i ends:

$$s(e_{i+1}) = t(e_i)$$

for each $i = 1, \ldots, n-1$. The number n is called the length of the path γ, denoted $|\gamma|$.

We say the **source of** γ is $s(\gamma) := s(e_1)$, and the **target of** γ is $t(\gamma) := t(e_n)$. We can also say γ is a path **from** $s(\gamma)$ **to** $t(\gamma)$, and denote $\gamma : s(\gamma) \to t(\gamma)$.

For any pair of vertices $v, w \in V(G)$, we denote the set of directed paths from v to w as $\Gamma(v, w)$, and we denote the set of all directed paths in the graph G as

$$\Gamma(G) := \bigcup_{v,w \in V(G)} \Gamma(v, w).$$

Definition 4.3. Given a graph G, an **(undirected) path** λ in G is a sequence of $n+1$ vertices $\{v_i\}_{i=0}^n$ together with a sequence of n edges $\{e_i\}_{i=1}^n$, such that for each $i = 1, ..., n$,

$$\begin{cases} s(e_i) = v_{i-1} \\ t(e_i) = v_i \end{cases} \quad \text{or} \quad \begin{cases} s(e_i) = v_i \\ t(e_i) = v_{i-1} \end{cases}$$

The number n is called the length of the path λ, denoted $|\lambda|$.

We say λ is an undirected path **between** v_0 and v_n. For any pair of vertices $v, w \in V(G)$, we denote the set of undirected paths between v and w as $\Lambda(v, w)$. Note that $\Lambda(v, w) = \Lambda(w, v)$. As before, we overload the notation, denoting the set of all undirected paths in the graph G as

$$\Lambda(G) := \bigcup_{v,w \in V(G)} \Lambda(v, w).$$

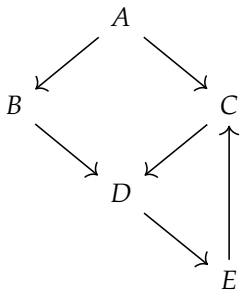
Definition 4.4. A **(directed) cycle** is a directed path γ whose source and target coincide: $s(\gamma) = t(\gamma)$. The set of cycles starting and ending at vertex v is denoted

$$\kappa(v) := \Gamma(v, v)$$

and the set of all cycles in the graph G is denoted

$$\kappa(G) := \bigcup_{v \in V(G)} \kappa(v).$$

Example 4.1. Consider the directed graph G:



This graph has five vertices, and six edges. Consider the vertices $A, D \in V(G)$. The set of directed paths $\Gamma(A, D)$ consists of two paths: one through B and one through C. The path going $C \to D \to E \to C$ is a directed cycle in G. The undirected path A, B, D, C, A is an "undirected cycle", but not a directed cycle.

Definition 4.5. A **directed acyclic graph (DAG)** is a directed graph G that has no cycles: $\kappa(G) = \emptyset$.

Definition 4.6. Let G be a DAG, and let v be a vertex in G. We say the set of **parents** of v is

$$\mathrm{pa}(v) := \{w \in V : \exists e : w \to v \in E\},$$

the set of **children** of v is

$$\mathrm{ch}(v) := \{w \in V : \exists e : v \to w \in E\},$$

the set of **ancestors** of v is

$$\mathrm{anc}(v) := \{w \in V : \Gamma(w, v) \neq \emptyset\}$$

and the set of **descendants** of v is

$$\mathrm{des}(v) := \{w \in V : \Gamma(v, w) \neq \emptyset\}.$$

A vertex v is called a **root vertex** if $\mathrm{pa}(v) = \emptyset$, and a **leaf vertex** if $\mathrm{ch}(v) = \emptyset$.

At this point, we can answer the question: why is DAG an appropriate tool for thinking about computational relations? The answer is this: by having no directed cycles, it becomes possible to "compute" (whatever that means in any given context) each relevant variable, such that no variable is ever computed before any of its parents. One can clearly see the importance of this condition in any computational scenario. This fact is formalized as follows:

Theorem 4.1. *(Topological Sorting of DAGs) Let G be a DAG where $V(G)$ is finite with size n. Then there exists an ordering $\{v_i\}_{i=1}^n$ of all the vertices in G, such that for each i,*

$$\mathrm{pa}(v_i) \subset \{v_1, \ldots, v_{i-1}\}.$$

Proof. This fact is well known and is proven by Kahn in [10] □

4.2 Noumenal Content of Structural Causal Models

The rough idea of SCMs is that each generative dependence relation, i.e. each node in relation to its parents, behaves "like an RCM", in that its core causal content is expressed by a Markov kernel. This idea is formalized in the following definition:

Definition 4.7. A **Structural Causal Model (SCM)** $\mathcal{G} = (G, X_*, \mathbb{P}_{*|pa(*)})$ consists of the following data:

1. A directed acyclic graph G;
2. For each vertex v in G, an assigned measurable space X_v, equipped with σ-algebra $\mathscr{A}_v$;

3. For each vertex v in G, an assigned Markov kernel

$$\mathbb{P}_{v|pa(v)} : \prod_{w \in \text{pa}(v)} X_w \to X_v.$$

The Markov kernel $\mathbb{P}_{v|pa(v)}$ is called the "structural kernel" for X_v.

We take the empty product of measurable spaces to be I, the singleton space. So for an exogenous variable, its structural kernel will have domain I, and so is simply a probability measure.

Example 4.2. Consider again examples 3.1 and 3.2. There is a treatment, Accupill, represented by the random variable T. This treatment is hypothesized to have a causal effect on the outcome Y. There is also a variable X, representing the pre-treatment features of a patient, which causally influences both the likelihood of this patient obtaining treatment (T), and the likelihood of recovery (Y). This situation is represented by the following SCM displayed in figure 1.

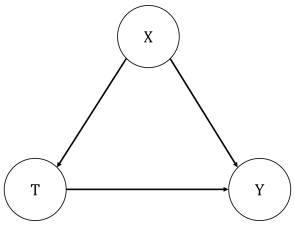


Figure 1: Structural Causal Model for the Accupill Example

There will be three structural kernels in this SCM. First, a kernel $I \to X$ gives a probability distribution over X. Second, a kernel $X \to T$ gives a probability distribution on T for every value of x. Finally, a kernel $X \times T \to Y$

Example 4.3. Suppose that Pfizer now develops a new type of treatment for COVID, and its clinical trial proceeds in the following steps:

1. The patient is given either a placebo or Accupill (T).
2. A viral test is performed on the patient, and the result of the test is recorded (Y).
3. A second treatment, whose content depends on the result of the viral test, is given to the patient (S).
4. The patient either recovers or does not recover (Z).

Suppose there is also a confounding factor H between the viral test result and the content of the second treatment. This situation is represented by the following SCM:

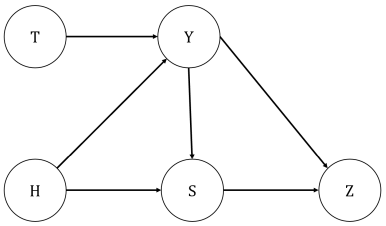


Figure 2: Structural Causal Model for Two-Stage Treatment

There will be five structural kernels in this SCM:

1. The kernel $\mathbb{P}_T : I \to T$ gives a probability distribution over T;
2. The kernel $\mathbb{P}_H : I \to H$ gives a probability distribution over H;
3. The kernel $\mathbb{P}_{Y|T,H} : T \times H \to Y$ gives a distribution over Y for each pair of values $t \in T, h \in H$;
4. The kernel $\mathbb{P}_{S|Y,H} : Y \times H \to S$ gives a distribution over S for each pair of values $y \in Y, h \in H$;
5. The kernel $\mathbb{P}_{Z|Y,S} : Y \times S \to Z$ gives a distribution over Z for each pair of values $y \in Y, s \in S$.

4.3 Observational Content of Structural Causal Models

What is the correspondence between SCMs and observational data? We can take observational data to be presented in terms of empirical distributions on random variables. Let $\mathcal{G} = (G, X_*, \mathbb{P}_{*|pa(*)})$ be an SCM. Then a probability distribution $\mathbb{P}$ over the product measure space $X = \prod_v X_v$ is said to **satisfy** $\mathcal{G}$ if it is "built out of" the structural kernels $\mathbb{P}_{v|pa(v)}$. The precise meaning of this can only be developed in chapter 6, after the development of more infrastructure around Markov categories.

However, in the garden-variety case where the measurable spaces X_v come with their own intrinsic measures μ_v, and where the probability measure $\mathbb{P}$ on the product space X has a density function f with respect to the product measure $\mu = \bigotimes_v \mu_v$, then we can give a simple definition for satisfaction:

Definition 4.8. Let $\mathcal{G} = (G, X_*, \mathbb{P}_{*|pa(*)})$ be an SCM. For each X_v, let μ_v be a measure. Let $\mathbb{P}$ be a probability measure on the product space $X = \prod_v X_v$ with a density function f with respect to the product measure $\mu = \bigotimes_v \mu_v$. Then $\mathbb{P}$ is said to satisfy $\mathcal{G}$ if each structural kernel $\mathbb{P}_{v|pa(v)}$ has a μ_v-conditional density function[3] $f_{v|pa(v)}$, and

$$f(x) = \prod_{v \in V(G)} f_{v|pa(v)}(x_v \mid x_{pa(v)}).$$

[3] in the sense of definition 3.8

The above definition is also known as the "product decomposition rule". We will see in section 6.2 that the product decomposition rule is equivalent to the more general definition of satisfaction, which will work even when density functions don't exist. For the rest of this chapter, we will always work with the case where density functions exist.

Like any graphical model of probabilistic processes, the crucial observational content of SCMs consists of independence relations between variables. Importantly, there is a set of independence relations that arise purely as a consequence of the *topology* of the DAGs underlying the SCMs, so that any two SCMs with the same underlying DAG will both have these independence relations. In this sense, such independence relations form the purely *structural* component of the observational content of SCMs. We begin by examining three elementary topological structures: the chain, the fork, and the collider. These elementary structures will allow us to retrieve all the purely structural independence relations.

Definition 4.9. (**chain**) A chain is a graph isomorphic to the following:

$$v \to w \to u$$

Lemma 4.1. *If $\mathcal{G} = (G, X_*, \mathbb{P}_{*|pa(*)})$ is a structural causal model, $\mathbb{P}$ satisfies $\mathcal{G}$, and G is a chain $v \to w \to u$, then $X_u \amalg X_v \,|\, X_w$, where X_u, X_v, X_w are given the marginal distributions according to $\mathbb{P}$.*

Proof. Since $\mathbb{P}$ satisfies $\mathcal{G}$, so if f is the density function associated with $\mathbb{P}$, then for any $x_v \in X_v$, $x_w \in X_w$, and $x_u \in X_u$:

$$f(x_v, x_w, x_u) = f_v(x_v) f_{w|v}(x_w \mid x_v) f_{u|w}(x_u \mid x_w).$$

Therefore

$$\begin{aligned} f_{v,u|w}(x_v, x_u \mid x_w) &= \frac{f(x_v, x_w, x_u)}{f_w(x_w)} \\ &= \frac{f_v(x_v) f_{w|v}(x_w \mid x_v)}{f_w(x_w)} f_{u|w}(x_u \mid x_w) \\ &= f_{v|w}(x_v \mid x_w) f_{u|w}(x_u \mid x_w). \end{aligned}$$

□

Definition 4.10. (**fork**) A fork is a graph isomorphic to the following:

$$v \leftarrow w \to u$$

Lemma 4.2. *If $\mathcal{G} = (G, X_*, \mathbb{P}_{*|pa(*)})$ is a structural causal model, $\mathbb{P}$ satisfies $\mathcal{G}$, and G is a fork $v \leftarrow w \to u$, then $X_v \amalg X_u \mid X_w$, where X_u, X_v, X_w are given the marginal distributions according to $\mathbb{P}$.*

Proof. Since $\mathbb{P}$ satisfies $\mathcal{G}$, so if f is the density function associated with $\mathbb{P}$, then for any $x_v \in X_v$, $x_w \in X_w$, and $x_u \in X_u$:

$$f(x_v, x_w, x_u) = f_w(x_w) f_{v|w}(x_v \mid x_w) f_{u|w}(x_u \mid x_w).$$

Therefore

$$\begin{aligned} f_{v,u|w}(x_v, x_u \mid x_w) &= \frac{f(x_v, x_w, x_u)}{f_w(x_w)} \\ &= f_{v|w}(x_v \mid x_w) f_{u|w}(x_u \mid x_w). \end{aligned}$$

□

Definition 4.11. (**collider**) A collider is a graph isomorphic to the following:

$$v \to w \leftarrow u$$

Lemma 4.3. *If $\mathcal{G} = (G, X_*, \mathbb{P}_{*|pa(*)})$ is a structural causal model, $\mathbb{P}$ satisfies $\mathcal{G}$, and G is a collider $v \to w \leftarrow u$, then $X_v \amalg X_u$ unconditionally; but in general, it is not the case that $X_v \amalg X_u \mid X_w$, where X_u, X_v, X_w are given the marginal distributions according to $\mathbb{P}$.*

Proof. Since $\mathbb{P}$ satisfies $\mathcal{G}$, so if f is the density function associated with $\mathbb{P}$, then for any $x_v \in X_v$, $x_w \in X_w$, and $x_u \in X_u$:

$$f(x_v, x_w, x_u) = f_v(x_v) f_u(x_u) f_{w|v,u}(x_w \mid x_v, x_u).$$

Marginalizing over X_w, we obtain

$$\begin{aligned} f_{v,u}(x_v, x_u) &= \int f_v(x_v) f_u(x_u) f_{w|v,u}(x_w \mid x_v, x_u)\, dx_w \\ &= f_v(x_v) f_u(x_u) \int f_{w|v,u}(x_w \mid x_v, x_u)\, dx_w \\ &= f_v(x_v) f_u(x_u). \end{aligned}$$

So the unconditional independence obtains. Now, to see that the conditional independence does not hold in general, consider the case where X_v and X_u are both standard normal, and $X_w = X_v + X_u$. In this case, conditional on $X_v = x_v$ and $X_w = x_w$, the random variable X_u has only one possible value, namely $x_w - x_v$. So the conditional independence $X_u \amalg X_v \mid X_w$ does not hold. □

We now extend these three independence lemmas to full generality.

Definition 4.12. (**blocking**) Suppose λ is an undirected path in G. Let Z be a set of vertices in G. We say that λ is blocked by Z if and only if

1. λ contains a chain $v \to w \to u$ or a fork $v \leftarrow w \to u$ such that the middle node w is contained in Z; or

2. λ contains a collider $v \to w \leftarrow u$ such that $w \notin Z$, and no descendent of w is in Z either.

Definition 4.13. (**d-separation**) Let u, v be vertices in a DAG G, and let Z be a set of vertices in G. We say that u and v are d-separated by Z if every undirected path between u and v are blocked by Z. If A, B are sets of vertices in G, we say that A and B are d-separated by Z if for every pair of vertices $a \in A$ and $b \in B$, a and b are d-separated by Z.

Theorem 4.2. *If $\mathcal{G} = (G, X_*, \mathbb{P}_{*|pa(*)})$ is a structural causal model, $\mathbb{P}$ satisfies $\mathcal{G}$, and $u, v \in V(G)$ are d-separated by $Z \subset V(G)$, then $X_u \amalg X_v \mid X_Z$, where X_u, X_v, X_Z are given the marginal distributions according to $\mathbb{P}$.*

Theorem 4.3. *If G is a DAG and $u, v \in V(G)$ are not d-separated by $Z \subset V(G)$, then there exists an SCM $\mathcal{G} = (G, X_*, \mathbb{P}_{*|pa(*)})$ with G as the underlying graph, and a distribution $\mathbb{P}$ on X, such that X_u is not independent on X_v conditional on X_Z.*

Proof. Both theorems 4.2 and 4.3 are proven in [5]. □

In this sense, we see that d-separation is a sound and complete test for conditional independence in SCMs. The topological feature of d-separation in a DAG picks out the full set of conditional independence relations satisfied by every SCM over this DAG. However, two different DAGs might have the same set of d-separations, and therefore imply the same set of independence relations. This is captured in the following definition:

Definition 4.14. (**Markov equivalence**) Let G and G' be DAGs. They are said to be Markov equivalence if:

1. There exists a bijection $i : V(G) \to V(G')$; and
2. For all $u, v \in V(G)$ and $Z \subset V(G)$, Z d-separates u, v in G if and only if $i(Z)$ d-separates $i(u), i(v)$ in G'.

This relation between DAGs is an equivalence relation. The class of all DAGs that are Markov equivalent to G is called the Markov equivalence class of G.

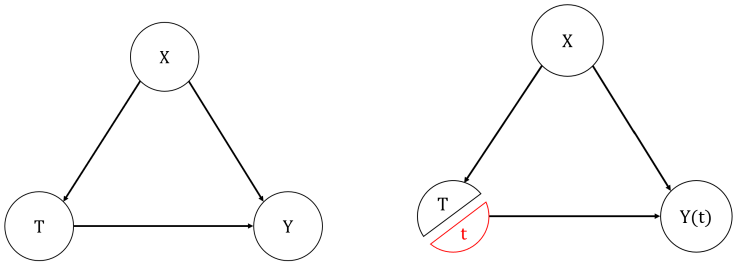
Example 4.4. Let G be the graph $u \to v$, and G' be the graph $u \leftarrow v$. Then G and G' are Markov equivalent (both having no d-separation relations at all), and $\{G, G'\}$ forms a Markov equivalence class.

4.4 Interventional Content of Structural Causal Models

The interventional content of SCMs is defined by a specific operation on SCMs, known as the Single World Intervention Graph (SWIG) [16]. The crucial idea of this section is that there is an operation SWIG, which takes in an SCM $\mathcal{G}$, and returns a different SCM $\mathcal{G}'$, such that the *interventional* content of $\mathcal{G}$ is nothing more or less than the *observational* content of $\mathcal{G}'$. We begin by developing some intuition via the Accupill example (example 4.2) once again.

Example 4.5. Suppose Pfizer is testing their new pill, Accupill, intended to help patients recover from COVID. Henry is a patient. Under normal circumstances, if the pill was just available over the counter, Henry would not have chosen to take the pill (that is to say, his pre-treatment features X are such that he would choose not to take the pill). However, today Henry is participating in a randomized trial, and the experimenter simply makes it the case that Henry will take the pill, regardless of Henry's own preferences.

The intervention by the experimenters can be seen essentially as the decoupling of two things: the treatment Henry *would* have decided to undergo under normal circumstances, and the treatment Henry *actually* takes under the intervention. The situation can be modelled by the operation taking the SCM in figure 3a to the SCM in figure 3b.



(a) Before Intervention $\rightsquigarrow$ (b) After Intervention

Figure 3: SWIG operation on Accupill SCM.

In the new SCM in figure 3b, the node T no longer represents the treatment that Henry receives. Instead, it represents the treatment Henry *would* have received, if there were no experimental intervention. The constant node t (colored red to indicate that it is a constant node) represents the treatment experimentally assigned to Henry. Note that there is no path between T and t. Thus, $Y(t)$ and T are independent conditional on X, reflecting the weak unconfoundedness assumption.

Many reader might find the previous example sufficient for understanding what the SWIG operation does. Nevertheless, for the sake of rigour, we provide a procedural definition of the operation:

Definition 4.15. (**The SWIG Operation**) Let $\mathcal{G} = (G, X_*, \mathbb{P}_{*|pa(*)})$ be an SCM. Let $T \subset V(G)$ (think of this as the set of variables on which we will intervene as experimenters). Let $t = (t_v)_{v \in T}$ be a value in the product measurable space X_T (recall that for $T \subset V(G)$, X_T denotes the product measurable space $\prod_{v \in T} X_v$). The SWIG operation then constructs an SCM $(G', X'_*, \mathbb{P}'_{*|pa(*)})$, which we shall denote as $\mathcal{G}(\text{do}\, T = t)$, as follows:

- (Node Splitting) Each $v \in T \subset V(G)$ is split into two nodes $v'_1, v'_2 \in V(G')$, such that v'_1 inherits the edges pointing into v, and v'_2 inherits the edges

pointing out of v. All other parts of G' are the same as G. In other words, G' is a graph (unique up to isomorphism) admitting a surjective graph homomorphism

$$\phi : G' \to G$$

such that ϕ is bijective on edges, and locally isomorphic at every vertex $v \notin T$, and the preimage $\phi^{-1}(v)$ for each $v \in T$ consists of exactly two vertices $v'_1, v'_2 \in G'$, such that v'_1 has only edges flowing into it, and v'_2 has only edges flowing out of it. The set of edges flowing into v'_1 is precisely $\phi^{-1}\{e : w \to v \mid w \in G\}$, and the set of edges flowing out of v'_2 is $\phi^{-1}\{e : v \to w \mid w \in G\}$.

- (Assigning Measurable Spaces) For $v \in V(G)$, and for each $v' \in \phi^{-1}(v)$ (as said above, there is either one such v' or two such v's), the measurable space $X'_{v'}$ is set to the same space as X_v.

- (Assigning Structural Kernels)

  - For $v \notin T$, let v' be the unique node in $V(G')$ corresponding to v. Then $pa(v) \cong pa(v')$ as a set, and $X_{pa(v)} \cong X_{pa(v')}$ as measurable spaces, by the foregoing constructions. We declare

  $$\mathbb{P}'_{v|pa(v')} := \mathbb{P}_{v|pa(v)}.$$

  - For $v \in T$, let v'_1, v'_2 be the two corresponding nodes in $V(G')$. Then $pa(v'_1) \cong pa(v)$ as sets and $X_{pa(v'_1)} \cong X_{pa(v)}$ as measurable spaces; but $pa(v'_2) = \emptyset$ because v'_2 only has outgoing edges. We declare

  $$\mathbb{P}'_{v'_1|pa(v'_1)} = \mathbb{P}_{v|pa(v)}$$
  $$\mathbb{P}'_{v'_2} = \delta_{t_v}$$

  where δ_{t_v} is the Dirac measure at the point $t_v \in X_v \cong X'_{v'_2}$.

Note here: the above construction fully defines the SWIG operation as an operation on SCMs: for each SCM $\mathcal{G}$, and each possible treatment $t \in X_T$, the above constructions gives us the unique SCM $\mathcal{G}(T = t)$. However, it is convenient, when we visualize $\mathcal{G}(T = t)$ as a graph, to know what to name each vertex in the resulting graph. For example, in example 4.5, we had named the vertices in the resulting graph X, T, t and $Y(t)$. This is not only convenient notationally, but underlines the intimate relation between SCMs and RCMs, where, you will recall, the notation $Y(t)$ was used to denote the potential outcome variable under a given treatment t. We specify here what exactly this notational convention is:

- For $v \notin T$, let v' be the unique corresponding node in G'. The measurable space $X'_{v'}$ is labelled $X_v(X_{u_1} = t_{u_1}, X_{u_2} = t_{u_2}, \dots)$, where $\{u_1, u_2, \dots\}$ is the intersection $T \cap anc(v)$. In other words, we label a node with brackets containing all the treatments that are ancestors of the node.

- For $v \in T$, let v'_1, v'_2 be the two corresponding nodes in G'. $X'_{v'_1}$ is labelled exactly the same as in the previous case as $X_v(X_{u_1} = t_{u_1}, X_{u_2} = t_{u_2}, \dots)$, where $\{u_1, u_2, \dots\} = T \cap anc(v)$. On the other hand, $X'_{v'_2}$ is simply labelled t_v, in lower case, to emphasize that this is a *fixed* node, with the intervention $X_v = t_v$.

Example 4.6. Consider again the two-stage treatment described in example 4.3. The SCM of the situation is given by figure 4.

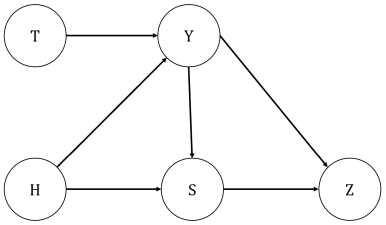


Figure 4: Structural Causal Model for Two-Stage Treatment

Suppose we now perform a clinical trial on a patient, such that the stage-one treatment T is determined by an experimenter, whereas the stage-two treatment S is determined, as before, by Y and the confounding factor H. Then the SWIG operation on the above SCM yields an SCM given by figure 5.

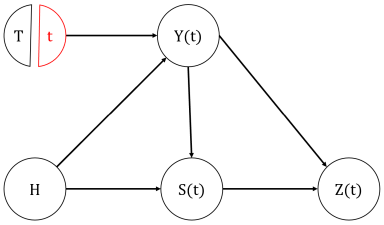


Figure 5: SWIG for Two-Stage Treatment with Intervention on Stage-One

On the other hand, suppose now we perform a clinical trial, in which both the stage-one and the stage-two treatments T, S are determined by an experimenter. Then the SWIG operation on the above SCM yields an SCM given by figure 6.

With the SWIG operation defined, let us retrieve what we said earlier about the interventional content of SCMs. Let $\mathcal{G} = (G, X_*, \mathbb{P}_{*|pa(*)})$ be an SCM. The interventional content of $\mathcal{G}$ says precisely this:

Let $T \subset V(G)$, and $t \in X_T$. Let $\mathcal{G}(\text{do}\, T = t) = (G', X'_*, \mathbb{P}'_{*|pa(*)})$ be the SWIG on $\mathcal{G}$ with respect to T. If we were to intervene as experimenters on the aspects of this world corresponding to T, by forcing them to have values t, then the

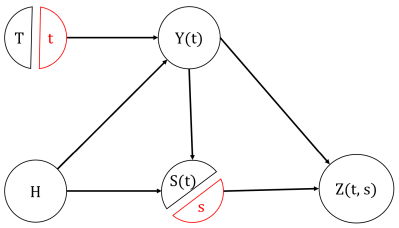


Figure 6: SWIG for Two-Stage Treatment with Intervention on Both Stages

observed outcomes on $X' = \prod_{v \in V(G')} X'_v$ should have a distribution $\mathbb{P}'$ satisfying $\mathcal{G}(\text{do } T = t)$. In this sense, the *interventional* content of $\mathcal{G}$ is nothing more or less than the *observational* content of $\mathcal{G}(\text{do } T = t)$.

4.5 Concluding Remarks

Our discussion of SCMs in this chapter has shown that SCMs are more powerful than RCMs: they can model any finite network of causal relations, and each parent-child relation in an SCM has the same mathematical structure as an RCM. The price, of course, is that SCMs are harder to work with in empirical settings. It is harder to determine, in general, whether an SCM is accurate, and harder still to *come up with* an accurate SCM for some observed data. This is easy to see: the task of determining the accuracy of an SCM involves, as a subtask, that of determining the accuracy of its parent-child relations, which is the same as determining the accuracy of some RCMs.

A second difficulty is that we do not yet have an adequate understanding of SCMs where the Markov kernels do not admit conditional density functions. This will be resolved by the formalism introduced in chapters 5 and 6.

Finally, a philosophical worry is that, since it is very difficult to come up with good SCMs in empirical settings, there is not enough formalism to keep track of the meaningful relations *between* differing SCMs. If one researcher proposes a model M meant to model some phenomenon, and a second researcher improves upon this model to make M' for the same phenomenon, then there is some meaningful relation between M and M' that needs to be recorded. But we do not yet have an adequate notion of *morphisms* between SCMs, and so these relations can only be shared among researchers in the form of prose and expertise. We contend that our work in chapter 6 also makes improvements along these lines.

5 Category Theory for Causal Modelling

We now take a detour, and study the aspects of category theory which will become necessary in developing the last type of causal models in this thesis. Readers that are comfortable with higher category theory should probably skip this chapter. We begin, in section 5.1, by defining categories, functors, and natural transformations. These will provide us with the infrastructure to understand causal models, not just as stand-alone constructs, but as constructs that bear significant relations to each other. We then go on, in section 5.2, to study symmetric monoidal categories. This will give us a kind of "tensor product" with which we can combine random variables in a causal model. In both sections 5.1 and 5.2, we follow the presentation given by Riehl in [17].

In sections 5.3 and 5.4, we study monads and their corresponding Kleisli categories, which is the most natural way to interact with Markov kernels in a purely synthetic way. Finally, in section 5.5, we study the special class of categories called Markov categories. These will allow us to make sense of the fact that, in any generative model (causal or otherwise), data can be both copied and deleted. In these three sections, we follow the presentation given by Fritz in [4], supplemented with the statistical viewpoint given by Patterson in [15].

5.1 Categories, Functors, Natural Transformations

Definition 5.1. (**Category**) A category $\mathscr{C}$ is the following data:

- A collection $\mathrm{Ob}\,\mathscr{C}$ of objects in $\mathscr{C}$ (this need not be a set);
- For each pair of objects X, Y, a set $\mathrm{Hom}(X, Y)$ called the morphisms from X to Y; where, for each $f \in \mathrm{Hom}(X, Y)$, we say X is the domain of f and Y is the codomain of f; denoted $f : X \to Y$;
- For each object X, a specified morphism $1_X \in \mathrm{Hom}(X, X)$ called the identity morphism at X;
- For each triplet of objects X, Y, Z, a function

$$\mathrm{Hom}(X, Y) \times \mathrm{Hom}(Y, Z) \to \mathrm{Hom}(X, Z)$$

  called the composition; where the composition of $f \in \mathrm{Hom}(X, Y)$ with $g \in \mathrm{Hom}(Y, Z)$ is denoted $g \circ f$ or simply gf;

Such that the following axioms are satisfied:

1. (Unitality) For any pair of objects X, Y, and any morphism $f \in \mathrm{Hom}(X, Y)$, we have that $f \circ 1_X = 1_Y \circ f = f$; and
2. (Associativity) For any quadruplet of objects X, Y, Z, W, and any triplet of morphisms

$$X \xrightarrow{f} Y \xrightarrow{g} Z \xrightarrow{h} W$$

   we have that $(h \circ g) \circ f = h \circ (g \circ f)$.

Example 5.1. (Some Categories Relevant To Causal Modelling)

- The category *Set* consists of all sets as objects, and functions as morphisms.
- The category *Top* consists of all topological spaces as objects, and continuous functions as morphisms.
- The category *Meas* consists of measurable spaces as objects, and measurable functions as morphisms.
- The category *DirGraph* consists of directed graphs as objects, and directed graph homomorphisms as morphisms.
- The category *Poset* consists of partially ordered sets as objects, and monotone maps as morphisms.
- Let R be a ring. The category Mat_R consists of all positive integers as objects. A morphism from n to m in this category is any $m \times n$ matrix whose entries are in R. Composition of morphisms is given by matrix multiplication, and identity morphisms are identity matrices.
- The category *EssMeas* consists of measure spaces as objects, and equivalence classes of measurable functions as morphisms. Two measurable functions $f, g : (X, \mathscr{M}, \mu) \to (Y, \mathscr{N}, \nu)$ are equivalent if $f - g = 0$ μ-almost everywhere.

Definition 5.2. (**Isomorphism**) In a category $\mathscr{C}$, a morphism $f : X \to Y$ is called an isomorphism if there exists a morphism $g : Y \to X$ that acts as a two-sided inverse for f: $g \circ f = 1_X$ and $f \circ g = 1_Y$.

Example 5.2. (Isomorphisms in Useful Categories)

- In *Set*, an isomorphism is a bijection.
- In *Top*, an isomorphism is a homeomorphism.
- In *Meas*, an isomorphism is a measurable bijection whose set-theoretic inverse is also measurable.
- In *DirGraph*, an isomorphism is a graph isomorphism.
- In *Poset*, an isomorphism is an order-preserving bijection.
- In Mat_R, an isomorphism is an invertible matrix.
- In *EssMeas*, an isomorphism is the equivalence class of a measurable bijections whose set-theoretic inverse is also measurable.

Readers may have noticed that some of the above categories are *huge* - in the sense that the collection of objects is too large to form a set. For example, there is a category consisting of all sets as objects, but there is no set containing all sets. Indeed, it is important to distinguish these huge categories from small ones:

Definition 5.3. (**Small Category**) A category $\mathscr{C}$ is called small if its collection of objects forms a set.

Now that we know what categories are, we must talk about the relations between them. After all, the spirit of category theory is to never talk about a mathematical construct without talking about the relations between them!

Definition 5.4. (**Functor**) A functor F from a category $\mathscr{C}$ to a category $\mathscr{D}$ consists of the following data:

- For each object X in $\mathscr{C}$, a specified object $F(X)$ in $\mathscr{D}$;
- For each morphism $f : X \to Y$ in $\mathscr{C}$, a specified morphism $F(f) : F(X) \to F(Y)$ in $\mathscr{D}$;

Satisfying the following functoriality axioms:

1. For any object X in $\mathscr{C}$, we have that $F(1_X) = 1_{F(X)}$;
2. For any triplet of objects X, Y, Z in $\mathscr{C}$, and morphisms

$$X \xrightarrow{f} Y \xrightarrow{g} Z$$

in $\mathscr{C}$, we have that $F(g \circ f) = F(g) \circ F(f)$ in $\mathscr{D}$.

Example 5.3. (Identity Functors) For any category $\mathscr{C}$, there is an identity functor $1_{\mathscr{C}} : \mathscr{C} \to \mathscr{C}$ which sends every object X to X itself, and every morphism f to f itself.

Example 5.4. (Forgetful Functors) Many categories are defined with its objects being some kind of structured sets, and its morphisms being the structure-preserving functions (the categories *Set*, *Top*, *Meas*, *DirGraph*, *Poset* satisfy this description). In such a situation, there is a forgetful functor $U : \mathscr{C} \to Set$ sending each object in $\mathscr{C}$ to its underlying set, and each morphism in $\mathscr{C}$ to its underlying set-theoretic function. A category that admits such a forgetful functor is called a **concrete category**.

Example 5.5. There is a functor $Poset \to DirGraph$ sending a poset $(X, \leq)$ to the directed graph whose vertices are the elements of X, and an edge runs from x to y if and only if $x \leq y$. Monotone maps then get mapped to the corresponding graph homomorphisms.

Example 5.6. There is a functor $Meas \to EssMeas$ sending a measurable space to itself, and a measurable function to its equivalence class.

Definition 5.5. (**Natural Transformation**) Let F, G be functors $\mathscr{C} \to \mathscr{D}$. A natural transformation $\alpha : F \to G$ consists of the following data:

- For each object X in $\mathscr{C}$, a morphism $\alpha_X : F(X) \to G(X)$ in $\mathscr{D}$, called the **component of** α **at** X.

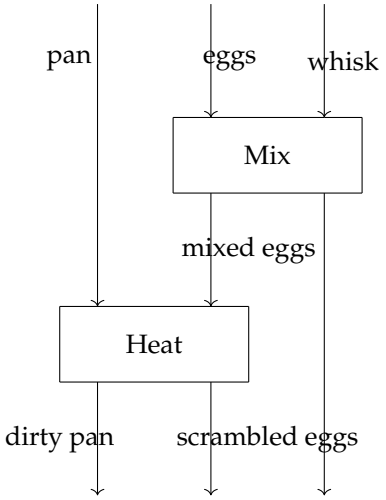
such that for each morphism $f : X \to Y$ in $\mathscr{C}$, the following diagram in $\mathscr{D}$ commutes:

$$\begin{array}{ccc} F(X) & \xrightarrow{\alpha_X} & G(X) \\ {\scriptstyle F(f)}\downarrow & & \downarrow{\scriptstyle G(f)} \\ F(Y) & \xrightarrow[\alpha_Y]{} & G(Y) \end{array}$$

A natural transformation α is called a **natural isomorphism** if each component α_X for each object X in $\mathscr{C}$ is an isomorphism in $\mathscr{D}$.

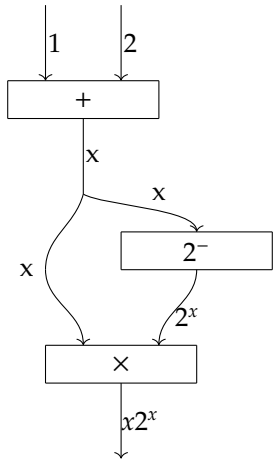
5.2 Symmetric Monoidal Categories

There are so many things in life that can be recorded in the language of string diagrams. Consider the following (perhaps, over-simplified?) recipe for scrambled eggs:



the strings in the diagram represent items in the kitchen, and the nodes (which I have drawn as boxes) represent actions. We begin with a pan, some eggs, and a whisk. We end with a dirty pan, some scrambled eggs, and a whisk.

Let's look at a more mathematical example. Suppose I told you the following: let $x = 1 + 2$; compute $x2^x$. What do you actually do? I'd bet it's something like this:



Diagrams like these are called **string diagrams**, and they naturally live in a special kind of categories called **Monoidal Categories**.

Consider the tensor products of rings and modules. They enabled us to rigorously talk about maps of modules with multiple inputs, without being too restrictive. Since, in the everyday string diagrams displayed above, most "morphisms" have multiple inputs and outputs, we would like a general notion of "tensor products". Loosely speaking, we are after a categorical construction that would allow us to juxtapose both objects and morphisms "in parallel" via some associative and functorial binary operation. Such a binary operation will be called a *monoidal product*.

Definition 5.6. (**Monoidal Category**). A *monoidal category* $(C, \otimes, I)$ is a category C together with a functor $\otimes : C \times C \to C$, called the *monoidal product*, and a fixed object $I \in C$ called the *monoidal unit*, subject to the interchange laws

1. For all objects $x, y \in C$,
$$1_{x \otimes y} = 1_x \otimes 1_y$$
where 1 denotes the identity morphism;

2. For all morphisms $u \xrightarrow{f} v \xrightarrow{h} w$ and $x \xrightarrow{g} y \xrightarrow{k} z$ in C,
$$(h \otimes k) \circ (f \otimes g) = (h \circ f) \otimes (k \circ g).$$

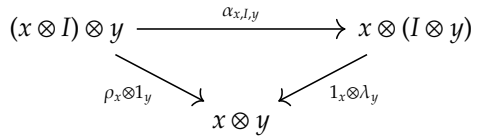
Moreover, there needs to be natural isomorphisms

1. (Unitors) For any object $x \in C$, natural isomorphisms $\lambda_x : I \otimes x \to x$ and $\rho_x : x \otimes I \to x$;

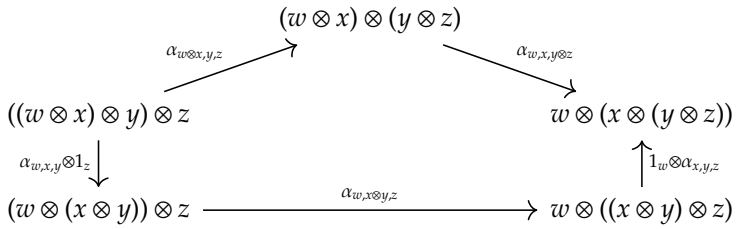
2. (Associators) for any objects $x, y, z \in C$, a natural isomorphism
$$\alpha_{x,y,z} : (x \otimes y) \otimes z \to x \otimes (y \otimes z)$$

subject to the following coherence axioms:

1. (The Triangle Equation) The following diagram commutes:



2. (The Pentagon Equation) The following diagram commutes:



If all the unitor and associator isomorphisms are identities, then the monoidal category is said to be *strict*.

Example 5.7. Here are some examples of monoidal categories where the monoidal product resembles a set-theoretic product or coproduct.

1. $(Set, \times, 1)$ is a monoidal category, where the monoidal product is the cartesian product, and the monoidal unit is the singleton set $1 = \{*\}$.

2. $(Set, +, 0)$ is a monoidal category, where the monoidal product is disjoint union, and the unit 0 is the empty set $\emptyset$.

3. For any field k, $(Vect_k, \otimes, k)$ is a monoidal category, where the monoidal product is the tensor product, and the unit is the field itself viewed as a vector space.

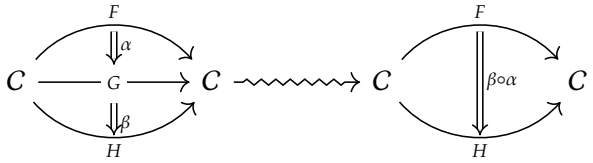
4. For any ring R, $(Mod_R, \otimes, R)$ is a monoidal category, where the monoidal product is the tensor product of modules, and the unit is the ring itself viewed as a module.

5. The category $(Meas, \times, \{*\})$ is a monoidal category, where the monoidal product is the usual product of measurable spaces and measurable functions' and the monoidal unit is the singleton measurable space.

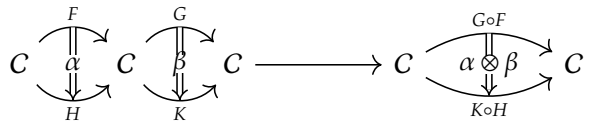
6. Let $(\mathbb{Z}_{\geq 0}, +, 0)$ be the monoidal category whose objects are nonnegative integers. There is exactly one morphism $n \to m$ if $n \leq m$ as integers; and no morphisms $n \to m$ otherwise. The monoidal product is summation, and the monoidal unit is 0. This is a *strict* monoidal category.

Example 5.8. However, note that a monoidal product might not resemble a set-theoretic product or coproduct at all! Fix a category C, there is a category $End(C)$ of endofunctors $C \to C$. The objects in this category are functors $F : C \to C$, and the morphisms in this category are natural transformations $\alpha : F \Rightarrow G$.

In the category $End(C)$, arrows $\alpha : F \Rightarrow G$ and $\beta : G \Rightarrow H$ can be composed by

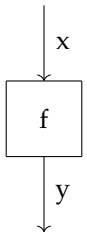


While objects $F : C \to C$ and $G : C \to C$ can be juxtaposed via functor composition $G \otimes F = G \circ F$. Now if there was an arrow $\alpha : F \to H$ and another arrow $\beta : G \to K$, we can find an arrow $\beta \otimes \alpha : G \otimes F \to K \otimes H$ by "horizontal" composition of natural transformations:

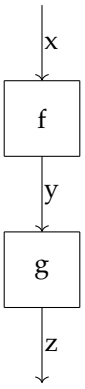


So the category $End(C)$ is a monoidal category, with monoidal product defined by composition on objects and horizontal composition on arrows, and with monoidal unit the identity functor 1_C.

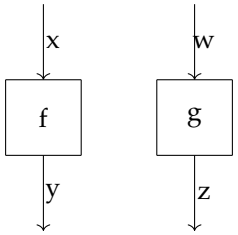
Now we can build the graphical language of string diagrams. Let $(C, \otimes, I)$ be a monoidal category. Then a morphism $f : x \to y$ can be represented by a box (or if you prefer, "node") labelled "f", with an incoming wire labeled "x", and an outgoing wire labeled "y".



The composition $g \circ f$ of morphisms $f : x \to y$ and $g : y \to z$ is represented by juxtaposition in series:



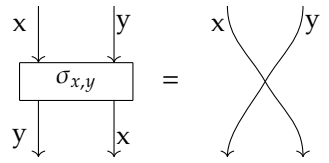
The monoidal product $f \otimes g : x \otimes w \to y \otimes z$ of morphisms $f : x \to y$ and $g : w \to z$ is represented by juxtaposition in parallel:



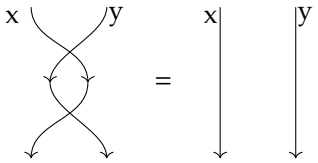
Identity morphisms $1_x : x \to x$ should be understood as the wires themselves, and the monoidal unit I is not drawn at all - it simple "permeates" the whole surface. Notice that the associativity laws and unitality laws (of both arrow composition and monoidal products), as well as the interchange laws, are implicit in the graphical syntax. If these laws did not hold, the graphical language would not be well-defined.

Often in a monoidal category, the monoidal products $x \otimes y$ and $y \otimes x$ are not "literally" equal, but contain the same information. For example, in *Sets*, the set $A \times B$ is not literally the same as the set $B \times A$, but they are isomorphic as sets. This is captured in the following definition:

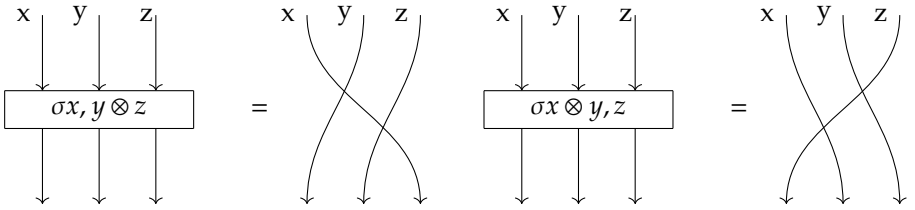
Definition 5.7. (Symmetric Monoidal Category) A monoidal category $(C, \otimes, I)$ is said to be *symmetric* if there is a natural isomorphism $\sigma_{x,y} : x \otimes y \to y \otimes x$, which we call *braidings* or *symmetries*, and depicted as crossed wires:



satisfying the involutivity axiom $\sigma^{-1}_{x,y} = \sigma_{y,x}$:



and two coherence axioms:



Example 5.9. The monoidal categories $(Sets, \times, 1)$, $(Sets, +, 0)$, $(Vect_k, \otimes, k)$, $(Mod_R, \otimes, R)$, and $(Meas, \times, \{*\})$ are all symmetric.

Example 5.10. However, the category $End(C)$ is not symmetric. In general, function composition is not commutative up to natural isomorphism.

Now that we have the structure of symmetric monoidal categories, and a good graphical language of them, we need to examine the proper relations between such categories. In particular, what are the functors that respect the symmetric monoidal structure?

Definition 5.8. (Lax Symmetric Monoidal Functors) Let $\mathscr{C}, \mathscr{D}$ be symmetric monoidal categories. A lax symmetric monoidal functor $F : \mathscr{C} \to \mathscr{D}$ consists of the following data:

1. A functor $F : \mathscr{C} \to \mathscr{D}$;
2. A morphism $\epsilon : I_{\mathscr{D}} \to F(I_{\mathscr{C}})$;
3. A natural transformation $\mu_{x,y} : F(x) \otimes_{\mathscr{D}} F(y) \to F(x \otimes_{\mathscr{C}} y)$ for $x, y \in \mathscr{C}$;

such that the following conditions hold:

1. (Associativity) For all objects $x, y, z \in \mathscr{C}$, the following diagram commutes:

$$\begin{array}{ccc}
(F(x) \otimes_{\mathscr{D}} F(y)) \otimes_{\mathscr{D}} F(z) & \xrightarrow{\alpha^{\mathscr{D}}_{F(x),F(y),F(z)}} & F(x) \otimes_{\mathscr{D}} (F(y) \otimes_{\mathscr{D}} F(z)) \\
\downarrow{\scriptstyle \mu_{x,y} \otimes id_{F(z)}} & & \downarrow{\scriptstyle id_{F(x)} \otimes \mu_{y,z}} \\
F(x \otimes_{\mathscr{C}} y) \otimes_{\mathscr{D}} F(z) & & F(x) \otimes_{\mathscr{D}} F(y \otimes_{\mathscr{C}} z) \\
\downarrow{\scriptstyle \mu_{x \otimes_{\mathscr{C}} y, z}} & & \downarrow{\scriptstyle \mu_{x, y \otimes_{\mathscr{C}} z}} \\
F((x \otimes_{\mathscr{C}} y) \otimes_{\mathscr{C}} z) & \xrightarrow[F(\alpha^{\mathscr{C}}_{x,y,z})]{} & F(x \otimes_{\mathscr{C}} (y \otimes_{\mathscr{C}} z))
\end{array}$$

where $\alpha^{\mathscr{C}}, \alpha^{\mathscr{D}}$ denote the associators of the categories $\mathscr{C}, \mathscr{D}$ respectively, and

2. (Unitality) For all $x \in \mathscr{C}$ the following diagrams commute:

$$\begin{array}{ccc} I_{\mathscr{D}} \otimes_{\mathscr{D}} F(x) & \xrightarrow{\epsilon \otimes id_{F(x)}} & F(I_{\mathscr{C}}) \otimes_{\mathscr{D}} F(x) \\ {\scriptstyle \lambda^{\mathscr{D}}_{F(x)}}\downarrow & & \downarrow{\scriptstyle \mu_{I_{\mathscr{C}},x}} \\ F(x) & \xleftarrow{F(\lambda^{\mathscr{C}}_x)} & F(I_{\mathscr{C}} \otimes_{\mathscr{C}} x) \end{array} \qquad \begin{array}{ccc} F(x) \otimes_{\mathscr{D}} I_{\mathscr{D}} & \xrightarrow{id_{F(x)} \otimes \epsilon} & F(x) \otimes_{\mathscr{D}} F(I_{\mathscr{C}}) \\ {\scriptstyle \rho^{\mathscr{D}}_{F(x)}}\downarrow & & \downarrow{\scriptstyle \mu_{x,I_{\mathscr{C}}}} \\ F(x) & \xleftarrow[F(\rho^{\mathscr{C}}_x)]{} & F(x \otimes_{\mathscr{C}} I_{\mathscr{C}}) \end{array}$$

3. (Symmetry) For all $x, y \in \mathscr{C}$ the following diagram commutes:

$$\begin{array}{ccc} F(x) \otimes_{\mathscr{D}} F(y) & \xrightarrow{\sigma^{\mathscr{D}}_{F(x),F(y)}} & F(y) \otimes_{\mathscr{D}} F(x) \\ {\scriptstyle \mu_{x,y}}\downarrow & & \downarrow{\scriptstyle \mu_{y,x}} \\ F(x \otimes_{\mathscr{C}} y) & \xrightarrow[F(\sigma^{\mathscr{C}}_{x,y})]{} & F(y \otimes_{\mathscr{C}} x) \end{array}$$

If the morphisms ϵ and $\mu_{x,y}$ are isomorphisms, then F is called a **strong symmetric monoidal functor**. If, further, ϵ and $\mu_{x,y}$ are identity morphisms, then F is called a **strict symmetric monoidal functor**.

Note: In the rest of this paper, unless otherwise specified, all symmetric monoidal functors are assumed to be strict.

Example 5.11. Here are some examples of strict symmetric monoidal functors:

1. Identity functors on symmetric monoidal categories are strict symmetric monoidal functors.
2. The forgetful functor $(Meas, \times, \{*\}) \to (Set, \times, \{*\})$ is a strict symmetric monoidal functor.
3. The quotient functor $(Meas, \times, \{*\}) \to (EssMeas, \times, \{*\})$, sending a measurable space to itself, and a measurable function to its equivalence class, is a strict symmetric monoidal functor.

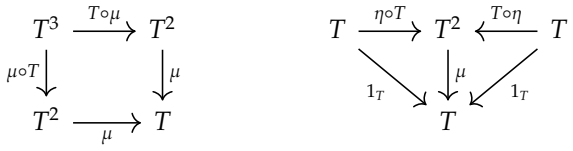
5.3 Monads and Kleisli Categories

We now need to develop the formalism that will allow us to understand how the notion of markov kernels naturally arises from the category of measurable spaces. This is the formalism of monads, which we now define. Intuitively, a monad is an endofunctor, which behaves like an algebraic monoid. Here is the rigorous definition:

Definition 5.9. (**Monad**) Let $\mathscr{C}$ be a category. A monad on $\mathscr{C}$ consists of the following data:

- A functor $T : \mathscr{C} \to \mathscr{C}$;
- A **unit** natural transformation $\eta : id_{\mathscr{C}} \to T$ where $id_{\mathscr{C}}$ is the identity functor;
- A **multiplication** natural transformation $\mu : T^2 \to T$ where T^2 is the functor $T \circ T$,

satisfying the following commutative diagrams:



Example 5.12. (**The Probability Monad**, otherwise known as **the Giry monad**) Let *Prob* be the functor *Meas* $\to$ *Meas* defined as follows:

- If X is a measurable space, $Prob(X)$ is the measurable space consisting of all probability measures P on X, equipped with the coarsest σ-algebra such that for all measurable set $A \subset X$, the evaluation map

  $$P \mapsto P(A)$$

  is a measurable function $Prob(X) \to [0, 1]$;
- If $f : X \to Y$ is a measurable function, $Prob(f) : Prob(X) \to Prob(Y)$ is the push-forward map

  $$P \mapsto P \circ f^{-1}$$

  which is well-defined because for every measurable $B \subset Y$, the inverse image $f^{-1}(B)$ is measurable in X.

Then *Prob* can be made into a monad by specifying:

- The unit transformation $\eta : id_{Meas} \to Prob$ is defined by setting each component $\eta_X : X \to Prob(X)$ as the map $x \mapsto \delta_x$, where δ_x is the point-mass distribution at x;
- The multiplication transformation $\mu : Prob^2 \to Prob$ is defined by the disintegration operation: for each measurable space X, $\mu_X : Prob^2(X) \to Prob(X)$ sends $Q \in Prob^2(X)$ to the measure $\mu_X(Q)$ defined by

  $$\mu_X(Q)(A) = \int_{P \in Prob(X)} P(A)\, dQ(P)$$

  for measurable $A \subset X$.

A monad allows us to make brand-new categories out of existing ones. This is done via the construction of the Kleisli category, which we now define:

Definition 5.10. (Kleisli Category) Let $\mathscr{C}$ be a category, and T be a monad on $\mathscr{C}$. The Kleisly Category $\mathscr{C}_T$ is defined as follows:

- An object in $\mathscr{C}_T$ is the same as an object in $\mathscr{C}$;
- A morphism $X \to Y$ in $\mathscr{C}_T$ is a morphism $X \to T(Y)$ in $\mathscr{C}$;
- For each object X in $\mathscr{C}_T$, the identity $1_X \in \mathscr{C}_T(X, X)$ is the component of the unit transformation $\eta_X : X \to T(X)$ in $\mathscr{C}$;
- For each sequence of morphisms

$$X \xrightarrow{f} Y \xrightarrow{g} Z,$$

in $\mathscr{C}_T$, which corresponds to morphisms $X \xrightarrow{f} T(Y)$ and $Y \xrightarrow{g} T(Z)$ in $\mathscr{C}$, the composition $g \circ f$ in $\mathscr{C}_T$ is defined as the morphism $\mu_Z \circ T(g) \circ f$ in $\mathscr{C}$:

$$X \xrightarrow{f} T(Y) \xrightarrow{T(g)} T^2(Z) \xrightarrow{\mu_Z} T(Z)$$

viewed as a morphism $X \to Z$ in $\mathscr{C}_T$.

Example 5.13. (Category of Markov Kernels) The Kleisli category $Meas_{Prob}$, which we shall denote as *Markov*, has:

- Measurable spaces as objects;
- Measurable functions $X \to Prob(Y)$ (that is to say, markov kernels $X \to Y$), as morphisms.

However, from this construction alone, the structure of this category still remains somewhat opaque to us. We will explicitly work out the structure of this category in section 5.4. For now, a bit more background.

Definition 5.11. (Symmetric Monoidal Monad) Let $\mathscr{C}$ be a symmetric monoidal category. A symmetric monoidal monad on $\mathscr{C}$ is a monad T, which, as a functor, is lax symmetric monoidal.

Example 5.14. The probability monad *Prob* is a symmetric monoidal monad on *Meas*. The product transformation $\nabla_{X,Y} : Prob(X) \times Prob(Y) \to Prob(X \times Y)$ is given by $(P, Q) \mapsto P \otimes Q$, where P is any probability measure on X, Q is any probability measure on Y, and $P \otimes Q$ is the independent product of the two measures.

Explicitly, this means that for "rectangles" $A \times B \subset X \times Y$, where A measurable in X and B measurable in Y, we define

$$P \otimes Q(A \times B) = P(A) \cdot Q(B).$$

Since these "rectangles" form an elementary family generating the σ-algebra on the product space $X \times Y$, so the above equation fully specifies our measure $P \otimes Q = \nabla_{X,Y}(P, Q)$.

Lemma 5.1. *If T is a symmetric monoidal monad on $\mathscr{C}$, then the Kleisli category $\mathscr{C}_T$ inherits the symmetric monoidal structure on $\mathscr{C}$, with monoidal products on morphisms defined as follows: if $f : X \to Y$ and $g : Z \to W$ are morphisms in $\mathscr{C}_T$, then $f \otimes_{\mathscr{C}_T} g$ corresponds to the following morphism in $\mathscr{C}$:*

$$X \otimes Z \xrightarrow{f \otimes_{\mathscr{C}} g} T(Y) \otimes T(W) \xrightarrow{\nabla_{Y,W}} T(Y \otimes W)$$

Proof. This is a standard result. See proposition 1.2.2 in [19]. □

Example 5.15. Thus, the category $Markov = Meas_{Prob}$ is a symmetric monoidal category. In the following section, we will make explicit the structure of this category.

5.4 The Category of Markov Kernels

In section 5.3, the most important running example was that of *Markov*, obtained by taking the Kleisli category of the monad *Prob* over the symmetric monoidal category *Meas*. In this section, we will explicitly describe the structure of *Markov*.

We have encountered markov kernels twice already: once in sections 3.4 and 4.2, where they were defined explicitly in terms of "conditional" probability distributions corresponding to randomness-involving functions; and once in section 5.3, where they were defined as morphisms in the Kleisli category $Meas_{Prob}$. That these markov kernels form a category was not surprising: we would expect that Markov Kernels can be composed. After all, if I write a randomness-involving function in Python that takes a value in X and spits out a value in Y, and then write another randomness-involving function that takes value in Y and spits out a value in Z, then I ought to be able to call these two functions one after another. But what exactly is the resulting composite markov kernel?

Lemma 5.2. *If $M : X \to Y$ and $N : Y \to Z$ are markov kernels, then their composition $N \circ M$ in Markov is defined by*

$$(N \circ M)(C \,|\, x) = \int_Y N(C \,|\, y)\, dM(y \,|\, x)$$

for any measurable $C \subset Z$, and any point $x \in X$.

Proof. By definition of the Kleisli category, $N \circ M$ is the measurable function defined by

$$\mu_Z \circ Prob(N) \circ M : X \to Prob(Z)$$

Now, by the push-forward definition,

$$(Prob(N) \circ M)(B \,|\, x) = M(N^{-1}(B) \,|\, x).$$

for any measurable $B \subset Prob(Z)$. Let Q be the measure $M(N^{-1}(-) \,|\, x)$ on $Prob(Z)$. Then by definition of μ_Z, we have that

$$(\mu_Z \circ Prob(N) \circ M)(C \,|\, x) = \int_{Prob(Z)} P(C)\, dQ(P).$$

But since $Q(B) = M(N^{-1}(B) \,|\, x)$, for any B disjoint from the image of $N : Y \to Prob(Z)$, we have $Q(B) = 0$. So we may restrict our domain of integration from $Prob(Z)$ to $\operatorname{im} N$:

$$(\mu_Z \circ Prob(N) \circ M)(C \,|\, x) = \int_{\operatorname{im} N} P(C)\, dQ(P).$$

Now we perform a change of variable by $P = N(- \,|\, y)$, so that $dQ(P) = dM(y \,|\, x)$, to obtain

$$(\mu_Z \circ Prob(N) \circ M)(C \,|\, x) = \int_Y N(C \,|\, y)\, dM(y \,|\, x).$$

□

In other words, we retrieve the Chapman-Kolmogorov equations for the composition of stochastic maps. If we consider the above in terms of concrete Python functions, it is not mysterious at all. How can it happen that the final output z is in some subset $C \subset Z$? Well, the intermediate value y, which is the output by the first function M and also the input for the second function N, could have been any value in Y. For a given y, we know that the probability that the final output is in C is $N(C \mid y)$. So, overall, the probability that the final output is in C becomes the integral of that value over the values in Y, with respect to the measure $M(- \mid x)$.

Now that we understand the composition of markov kernels, let's discuss the monoidal structure on *Markov*. As discussed in lemma 5.1, the monoidal product on objects in *Markov* will be the same as the monoidal product on objects in *Meas*: the product of measure spaces.

On the other hand, pick markov kernels $M : X \to Y$ and $N : Z \to W$. Since lemma 5.1 gave the monoidal product on morphisms in Kleisli categories as

$$X \otimes Z \xrightarrow{\substack{M \otimes N \\ Meas}} Prob(Y) \otimes Prob(W) \xrightarrow{\nabla_{Y,W}} Prob(Y \otimes W)$$

and since the transformation $\nabla_{Y,W}$ sends a pair of probability measures to their independent product, so the monoidal product of markov kernels $M \otimes N$ is the "conditional version" of independent product: for any measurable sets $A \subset Y$ and $B \subset W$, and for any input points $x \in X, z \in Z$:

$$M \otimes N(A \times B \mid x, z) = M(A \mid x) \cdot N(B \mid y),$$

which fully specifies the markov kernel $M \otimes N$, since the set of rectangles of the form $A \times B$ generates the σ-algebra on $Y \times W$.

We emphasize the fact that this monoidal product reflects an independence condition on the stochastic maps $M : X \to Y$ and $N : Z \to W$. In our python analogy, if M and N are both randomness-involving functions, then $M \otimes N$ corresponds to the computation of calling M and N on separate computers simultaneously, and taking their outputs together.

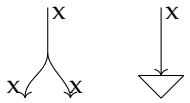
5.5 Markov Categories

It is nice that the category *Markov* exists, and has the structure of a symmetric monoidal category. But in fact, *Markov* is even nicer! It is a category that comes with morphisms corresponding to the operations of "copying" and "deleting" information. In this section, we define precisely what it means for a category to have these nice properties. In fact, symmetric monoidal categories that come with "copying" and "deleting" morphisms are called "Markov categories", because these are the essential features of the category *Markov*.

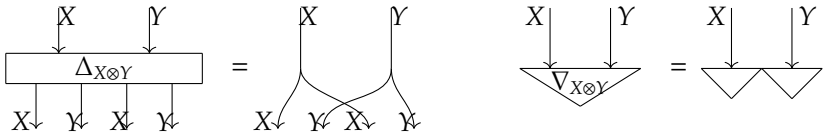
Definition 5.12. (Supply of Commutative Comonoids) For a symmetric monoidal category $(C, \otimes, I)$, a supply of commutative comonoids on C is the following data: for each object $X \in C$,

- A **copying** morphism $\Delta_X : X \to X \otimes X$;
- A **deleting** morphism $\nabla_X : X \to I$;

which are respectively depicted in the string diagram language as



and which must satisfy the following equations:



for every pair of objects $X, Y \in C$.

Example 5.16. In the symmetric monoidal category $(Sets, \otimes, 1)$, there is a supply of commutative comonoids where the copying morphisms are the functions $X \to X \otimes X$ defined by $x \mapsto (x, x)$; and the deleting morphisms are the functions $X \to 1$ where everything in X is sent to the single point in 1.

Example 5.17. In the symmetric monoidal category *Markov*, there is a supply of commutative comonoids given as follows:

- The copying morphisms are the kernels $\Delta_{\mathcal{X}} : \mathcal{X} \to \mathcal{X} \times \mathcal{X}$, where, for each input $x \in \mathcal{X}$, the resulting distribution $\Delta_{\mathcal{X}}(- \mid x)$ is the distribution assigning all the weight to the single point (x, x), i.e.

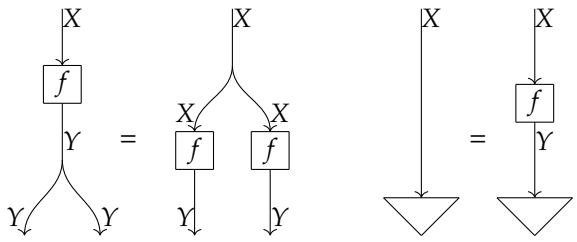
$$\Delta_{\mathcal{X}}(C \mid x) = \begin{cases} 1 & (x,x) \in C \\ 0 & (x,x) \notin C \end{cases}$$

  for all $C \in \Sigma_{\mathcal{X}} \otimes \Sigma_{\mathcal{X}}$;

- The deleting morphisms are the kernels $\nabla_{\mathcal{X}} : \mathcal{X} \to I$ sending everything to the only possible probability distribution on the one-point space I.

As you can see, the copying and deleting morphisms are not automatically required to be "natural". That is to say, they are not required to commute with the other morphisms in the symmetric monoidal category. They may commute with some morphisms, and this is captured in the following definition:

Definition 5.13. (Supply Homomorphism) If a symmetric monoidal category $(\mathcal{C}, \otimes, I)$ has a supply of commutative comonoids, then a morphism $f : X \to Y$ in $\mathcal{C}$ is a supply homomorphism if $f \circ \Delta_X = \Delta_Y \circ f$ and $\nabla_X = \nabla_Y \circ f$, i.e. the following equations hold:



Example 5.18. In the symmetric monoidal category $(Sets, \times, 1)$, every morphism is a homomorphism with respect to the supply of commutative comonoids.

Example 5.19. In the category *Markov*, every morphism $M : \mathcal{X} \to \mathcal{Y}$ commutes with the deleting morphism $\nabla_{\mathcal{Y}} \circ M = \nabla_{\mathcal{X}}$. However, the only morphisms that commute with the copying morphism are precisely those that are deterministic; i.e. those Markov kernels $M : \mathcal{X} \to \mathcal{Y}$ such that, for every $x \in \mathcal{X}$, the distribution $M(- \mid x)$ has all the density concentrated at one point in $\mathcal{Y}$.

We are now ready to define Markov Categories in general:

Definition 5.14. (Markov Categories) A Markov category is a symmetric monoidal category that supplies commutative comonoids.

Definition 5.15. (Functor of Markov Categories) A functor of Markov categories $F : \mathcal{C} \to \mathcal{D}$ is a functor that preserves the symmetric monoidal structure

and the supply of commutative comonoids of $\mathcal{C}$ and $\mathcal{D}$. Concretely, this is a symmetric monoidal functor $F : \mathcal{C} \to \mathcal{D}$ such that

$$F\left(\Delta_X^{\mathcal{C}}\right) = \Delta_{F(X)}^{\mathcal{D}}$$
$$F\left(\nabla_X^{\mathcal{C}}\right) = \nabla_{F(X)}^{\mathcal{D}}$$

for every object $X \in \mathcal{C}$.

Definition 5.16. (Cartesian Categories) If a Markov category has the additional property that every morphism is a supply homomorphism, then it is called a Cartesian category.

6 Functorial Causal Models

We now develop a new formalism for causal modelling, which is in many ways a natural extension of the Structural Causal Models developed in chapter 4. This new formalism, which we call Functorial Causal Models (FCM), is a combination of the ideas in SCM theory, and the ideas in Categorical Logic. The field of Categorical Logic was first launched by Lawvere in his thesis *Functorial Semantics of Algebraic Theories* [11], and was first applied to statistical problems by Patterson in his thesis *The Algebra and Machine Representation of Statistical Models* [15].

I build upon these works to develop the notion of a functorial causal model, which amounts to an account of functorial semantics for causal theories. My account will be self-contained (given the previous chapters in this thesis), but may seem out-of-the-blue at points, because I do not think it is appropriate to cover a full review of the field of Categorical Logic. Interested readers may find it helpful to review the works by Lawvere and Patterson [11, 15].

It will take two section to develop the noumenal content of functorial causal models. In section 6.1, we develop the notion of causal theories corresponding to DAGs. In section 6.2, we develop the functorial semantics of these theories, which will allow us to build FCMs. Then, in section 6.3, we describe the observational content of FCMs; and in section 6.4, we describe their interventional content.

6.1 Causal Theories Generated by DAGs

We first make the observation that the noumenal content of an SCM $M = (G, X_*, P_{*|pa(*)})$ can be viewed as consisting of two parts. On the one hand, the graph G is purely algebraic (or structural, or formal, however you prefer to think of it). It encodes structures without telling us what these structures have to do with the real world. On the other hand, the measurable spaces X_* and the structural kernels $P_{*|pa(*)}$ actually have something to do with the real world: they can refer to possible outcomes of real experiments and the statistical relations between these outcomes.

So, we will define causal *theories* in terms of directed acyclic graphs alone. In the sections following this one, we will then encode the data of $X_*, P_{*|pa(*)}$ in the form of functorial *semantics*.

Definition 6.1. (Causal Theory Generated by a DAG) Let G be a directed acyclic graph. The causal theory $\mathcal{T}_G$ generated by G is the small Markov category freely generated by G. More specifically:

1. The set of objects of $\mathcal{T}_G$ is the commutative monoid freely generated by the set of vertices $V(G)$. (Note, the monoidal product in $\mathcal{T}_G$ will, as is usual in category theory, be denoted by the tensor product $\otimes$).

2. For each vertex $v \in V(G)$, there is a morphism

$$p_{v|pa(v)} : \bigotimes_{u \in pa(v)} u \to v$$

in the category $\mathcal{T}_G$. This morphism is known as the **structural morphism** at v.

3. When combined with the structure of Markov categories (namely: composition, monoidal product, braidings morphisms, copying morphisms, and deleting morphisms), the structural morphisms specified in the above items freely generate the set of all morphisms in T_G.

So, we have the notion of a causal theory generated by a DAG: it is a small Markov category, which comes with special morphisms, called the structural morphisms, at each of its generating objects. With this understanding, we can now define causal theories in their full generality, without referring to a particular DAG.

Definition 6.2. (Causal Theory) A causal theory is a small Markov category $\mathcal{T}$ such that:

1. The monoid of objects $Ob(\mathcal{T})$ is a free finite-ranked commutative monoid. Its unique set of free generators is called the set of **variables** of the theory, denoted $V(\mathcal{T})$. The set of variables has a specified ordering $V(\mathcal{T}) = \{v_1, v_2, \dots\}$, called the **generative order** of the variables of theory $\mathcal{T}$.

2. For each variable $v_i \in V(\mathcal{T})$, there is a specified set of variables $pa(v_i) \subset \{v_1, \dots, v_{i-1}\}$ called the **parent variables** of v_i, and a specified morphism '

$$p_{v_i|pa(v_i)} : \bigotimes_{u \in pa(v_i)} u \to v_i$$

called the **structural morphism** at v_i.

Notation. For a subset of variables $U \subset V(\mathcal{T})$, it can get cumbersome to rewrite the expression

$$\bigotimes_{u \in U} u$$

every time we want to talk about the object in $\mathcal{T}_G$ containing exactly one copy of each vertex in U. So, we will abuse notation a little, and talk about the *object*

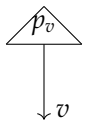
$$U := \bigotimes_{u \in U} u$$

in the theory $\mathcal{T}_G$.

Notation. Let $\mathcal{T}$ be a causal theory, and suppose $v \in V(\mathcal{T})$, and $pa(v) = \emptyset$. Then we say v is an **exogenous variable** of $\mathcal{T}$, and the structural morphism at v is a morphism

$$p_v : I \to v.$$

In order to emphasize that this morphism has the unit I as its domain, we will conventionally depict it as a triangle, rather than a rectangle, as follows:



Now, if G is a DAG, then the causal theory $\mathcal{T}_G$ generated by G has the vertices $V(G)$ as its variables, a topological ordering of $V(G)$ as its generative order, and the parent variables of each variable v_i corresponds to the graph-theoretical parents of v_i in the DAG G. So we know how to get from DAGs to causal theories. Can we go the other way? That is, given a causal theory, is it always generated by a DAG? The answer is no: not every causal theory is generated by a DAG, because not every causal theory is free. However, every causal theory does come *associated* with a DAG:

Definition 6.3. (DAG Associated to a Causal Theory) Let $\mathcal{T}$ be a causal theory. The DAG associated to it is the DAG G where the vertices are the variables $V(\mathcal{T})$, and where there is an edge $u \to v$ if and only if $u \in pa(v)$ in the theory $\mathcal{T}$.

Warning Many different causal theories $\mathcal{T}$ can be associated with the same DAG. In general, if $\mathcal{T}$ is a causal theory, and G is the DAG associated with it, $\mathcal{T} \neq \mathcal{T}_G$ where $\mathcal{T}_G$ is the theory generated by G.

The most important thing about causal theories is that they provide a definite structure with which a "generating" or "sampling" of value can take place. We now define that structure:

Definition 6.4. (Sampling Morphism) Let $\mathcal{T}$ be a causal theory, and let G be the DAG associated to it. The sampling morphism of $\mathcal{T}$ is the morphism

$$p : I \to \bigotimes_{v \in V(\mathcal{T})} v$$

whose string diagram "dualizes" G. Precisely: it is the morphism represented by the string diagram defined by the following procedure (this procedure was developed by Fong in [3]). Although this procedure, as presented here, is a bit obtuse on first reading, we believe that example 6.1 will serve to clarify it.

1. Let $v_1, \ldots, v_n$ be the generative ordering of the variables $V(\mathcal{T})$.
2. Initialize the string diagram D to be empty. That is, initialize it to the string diagram representing the identity morphism of the unit object I.

3. Let $codom(d)$ denote the collection of outgoing strings in the diagram d. At the beginning of this procedure, $codom(d) = \emptyset$. At the end of this procedure, $codom(d)$ will hold the value $V(\mathcal{T})$. Notice: $codom(d)$ is a collection, because it can have repeated elements.

4. For $i = 1, \ldots, n$:

   (a) The set $codom(d)$ will already contain at least one copy of each of $w \in pa(v_i)$. Take one copy of each of $w \in pa(v_i)$. Onto these, attach the morphism $p_{v_i|pa(v_i)}$. By doing so, we have added exactly one copy of v_i to $codom(d)$; while decreasing the number of copies in $codom(d)$ of each $w \in pa(v_i)$ by one.

   (b) Take the single copy of v_i in $codom(d)$. Onto it, attach the morphism $(\Delta_{v_i})^k : v_i \to (v_i)^{\otimes k+1}$, where k is the number of edges in G that have v_i as its source. By doing so, we have made it so that $codom(d)$ contains $k + 1$ copies of v_i.

5. Once the loop terminates, it is easily verified that $codom(d) = V(\mathcal{T})$. Further, we have not added any ingoing strings to d. So d now represents a morphism $p : I \to V(\mathcal{T})$ in the category $\mathcal{T}$. This morphism is the sampling morphism of $\mathcal{T}$.

Example 6.1. Suppose that G is the DAG

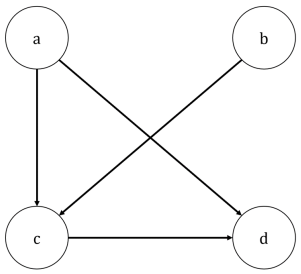
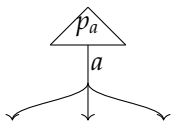
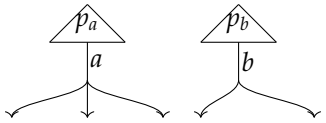


Figure 7: A DAG consisting of four nodes.

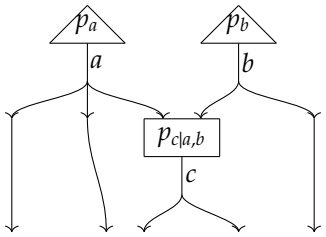
To find the sampling morphism of the causal theory $\mathcal{T}$ generated by G, we start by fixing a topological sorting of the vertices. Take a, b, c, d. The loop will run four times. On the first run, the structural morphism for a is appended to the diagram, together with a copying of a into three copies (since two other vertices requires a as a parent). At the end of the first run, D becomes:



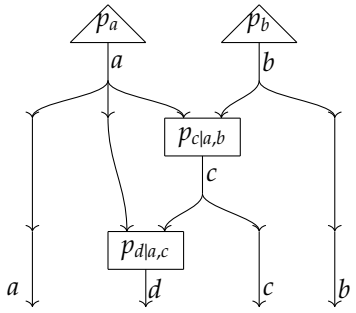
On the second run, the structural morphism for b is appended, together with a copying of b into two copies (since one other vertex requires b as a parent). At the end of the second run, D becomes:



On the third run, the structural morphism for c is appended, together with a copying of c into two copies (since one other vertex requires c as a parent). At the end of the third run, D becomes:



Finally, on the fourth run, the structural morphism for d is appended. There is no need to copy d, since no other vertex requires d as a parent. At the end of the fourth run, D becomes:



The morphism represented by the above diagram is then the sampling morphism p for the causal theory $\mathcal{T}$.

Sometimes, among all the variables $V(\mathcal{T})$ in a causal theory $\mathcal{T}$, only a subset $W \subset V(\mathcal{T})$ are of interest to an observer. In these cases, we require also a morphism $I \to W$, rather than the sampling morphism $I \to V(\mathcal{T})$. This is constructed quite easily with our formalism, and bears an evocative name:

Definition 6.5. (Marginal Sampling Morphism) Let $\mathcal{T}$ be a causal theory, with sampling morphism p. Let $W \subset V(\mathcal{T})$ be any subset of variables, and denote $W^c := V(\mathcal{T}) \setminus W$. The marginal sampling morphism p_W is defined by the composition

$$I \xrightarrow{p} V(\mathcal{T}) \cong W \otimes W^c \xrightarrow{id_W \otimes \nabla_{W^c}} W$$

where the middle isomorphism is a braiding isomorphism, and the right morphism is built out of the deletion morphism $\nabla_{W^c} : W^c \to I$.

One of the most important strengths of the category-theoretical view of causal models is that we can make explicit the meaningful relations between different causal models. Some of these relations are made explicit in terms of morphisms between causal theories. Therefore, we introduce here the notion of a morphism between causal theories. Other relations are made explicit in terms of morphisms between causal models. These will be discussed in section 6.2.

Definition 6.6. (Lax and Colax Causal Theory Morphism) Roughly, lax and colax causal theory morphisms are Markov functors from a theory $\mathcal{T}_1$ to another theory $\mathcal{T}_2$, such that the structural morphisms $p_{v|pa(v)}$ are preserved up to morphisms. Precisely: A lax causal theory morphism from theory $\mathcal{T}_1$ to theory $\mathcal{T}_2$ consists of the following data:

1. A Markov functor $F : \mathcal{T}_1 \to \mathcal{T}_2$;
2. For each $v \in V(\mathcal{T}_1)$, morphisms f_v and g_v such that

$$\begin{array}{ccccc} F(pa(v)) & \xrightarrow{F(p_{v|pa(v)})} & F(v) & = & \bigotimes_{w \in V(\mathcal{T}_2)} w^{\otimes m_w} \\ \downarrow{\scriptstyle f_v} & & & & \downarrow{\scriptstyle g_v} \\ \bigotimes_{w \in V(\mathcal{T}_2)} pa(w)^{\otimes m_w} & & \xrightarrow[\bigotimes_{w \in V(\mathcal{T}_2)} p_{w|pa(w)}^{\otimes m_w}]{} & & \bigotimes_{w \in V(\mathcal{T}_2)} w^{\otimes m_w} \end{array}$$

where

$$F(v) = \bigotimes_{w \in V(\mathcal{T}_2)} w^{\otimes m_w}$$

is the unique decomposition of $F(v)$ into variables w in $\mathcal{T}_2$.

For a colax causal theory morphism, the directions of the morphisms f_v and g_v are reversed. A colax causal theory morphism consists of

1. A Markov functor $F : \mathcal{T}_1 \to \mathcal{T}_2$;
2. For each $v \in V(\mathcal{T}_1)$, morphisms f_v and g_v such that

$$\begin{array}{ccccc} F(pa(v)) & \xrightarrow{F(p_{v|pa(v)})} & F(v) & = & \bigotimes_{w \in V(\mathcal{T}_2)} w^{\otimes m_w} \\ \uparrow{\scriptstyle f_v} & & & & \uparrow{\scriptstyle g_v} \\ \bigotimes_{w \in V(\mathcal{T}_2)} pa(w)^{\otimes m_w} & & \xrightarrow[\bigotimes_{w \in V(\mathcal{T}_2)} p_{w|pa(w)}^{\otimes m_w}]{} & & \bigotimes_{w \in V(\mathcal{T}_2)} w^{\otimes m_w} \end{array}$$

where

$$F(v) = \bigotimes_{w \in V(\mathcal{T}_2)} w^{\otimes m_w}$$

is the unique decomposition of $F(v)$ into variables w in $\mathcal{T}_2$.

Moreover, if f_v and g_v are isomorphisms for every v, then F is called a **strong** causal theory morphism, and if f_v and g_v are identity morphisms for every $v \in V(\mathcal{T}_1)$, then F is called a **strict** causal theory morphism.

Definition 6.7. (The Category of Causal Theories) The category *CauseTh* consists of all causal theories as objects, and all lax and colax causal theory morphisms as morphisms.

Here are two basic but important examples of causal theory morphisms:

Theorem 6.1. *Let $\mathcal{T}$ be a causal theory. Let G be the DAG associated with $\mathcal{T}$. Let $\mathcal{T}_G$ be the causal theory generated by G. Then there is a strict causal theory morphism $\mathcal{T}_G \to \mathcal{T}$.*

Proof. The monoids of objects $Ob(\mathcal{T})$ and $Ob(\mathcal{T}_G)$ are the same: they are both generated by $V(\mathcal{T}) = V(G)$. So let $F : \mathcal{T}_G \to \mathcal{T}$ act as identity on objects.

Then, for all $v \in V(G)$, there is an edge $u \to v$ in G if and only if $u \in pa(F(v))$. So $pa(F(v)) = F(pa(v))$. So we can define F to also act as identity on structural morphisms. Extend F to all morphisms in $\mathcal{T}_G$ so that it preserves the symmetric monoidal structure and the supply of commutative comonoids. Then F is a strict causal theory morphism. □

Notice: this is an extra-nice case of a strict causal theory morphism, because if p is the sampling morphism in $\mathcal{T}_G$, then $F(p)$ is the sampling morphisms in $\mathcal{T}$.

Theorem 6.2. *Let G_1, G_2 be DAGs, and let $\phi : G_1 \to G_2$ be a graph homomorphism. Let $\mathcal{T}_1, \mathcal{T}_2$ be the causal theories generated by G_1, G_2 respectively. Then there is a lax causal theory morphism $F : \mathcal{T}_1 \to \mathcal{T}_2$ induced by ϕ.*

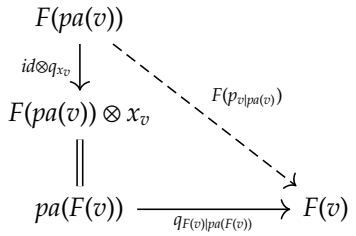
Proof. we can construct a lax causal theory morphism $F : \mathcal{T}_1 \to \mathcal{T}_2$ as follows.

On objects, F just acts as ϕ, so that $F(v) = \phi(v)$ for every variable $v \in V(\mathcal{T}_1)$. Because ϕ is a graph homomorphism, so every parent of v is sent by ϕ to some parent of $\phi(v)$ in G_2. That is, $\phi(pa(v)) \subset pa(\phi(v))$ for all $v \in V(G_1)$. Define x_v to be the object

$$x_v = pa(\phi(v)) \setminus \phi(pa(v))$$

This way, in the theory $\mathcal{T}_2$, we have $pa(F(v)) = F(pa(v)) \otimes x_v$.

Let q_{x_v} be the marginal sampling morphism for x_v in $\mathcal{T}_2$, and let $q_{F(v)|pa(F(v))}$ be the structural morphism at the variable $F(v)$ in $\mathcal{T}_2$. We then declare F to act on morphisms by sending the structural morphism $p_{v|pa(v)}$ in $\mathcal{T}_1$ to the following morphism:



This makes F a lax causal theory morphism $\mathcal{T}_1 \to \mathcal{T}_2$. □

Corollary 6.2.1. *The mapping $G \mapsto \mathcal{T}_G$ sending a graph G to the causal theory generated by it is therefore a functor DirGraph $\to$ CauseTh.*

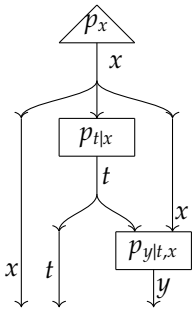
6.2 Functorial Semantics for Causal Theories

We now develop the functorial semantics of causal theories. Roughly, we will take the category *Markov* as the semantics for our causal theories. When we take theories and semantics together as a whole, we obtain the notion of a functorial causal model.

Definition 6.8. (FCM) A functorial causal model is a causal theory $\mathcal{T}$, together with a functor of Markov categories $M : \mathcal{T} \to \mathit{Markov}$.

Notice here that we require M to be a functor of Markov categories. In other words, it must (strictly) preserve the symmetric monoidal structure of $\mathcal{T}$, and it also must take copying and deleting morphisms to copying and deleting morphisms.

Example 6.2. (Accupill FCM) We are back again with our old friend, the Accupill example (see examples 3.2 and 4.2 for previous mentions of this example as RCM and SCM). There is a treatment, Accupill, which has been developed for curing COVID. It is hypothesized that T - whether or not a patient is administered accupill - has a causal effect on Y - whether or not she recovers. There is also a variable X, representing the pre-treatment features of a patient, which causally influences both T and Y. The causal theory $\mathcal{T}$ representing this situation has the following morphism as its sampling morphism:



The causal model representing this situation is a functor $M : \mathcal{T} \to \mathit{Markov}$:

- The functor M acts on the objects in $\mathcal{T}$ as follows:
  - $M(I) = I$ is the monoidal unit in the category *Markov*.
  - $M(x) = X$ is the measurable space of all possible pre-treatment features of all possible patients.
  - $M(t) = T$ is the discrete measurable space $T = \{0, 1\}$, where 0 is interpreted as "no treatment", and 1 is interpreted as "yes treatment".
  - $M(y) = Y$ is the discrete measurable space $Y = \{0, 1\}$, where 0 is interpreted as "does not recover", and 1 is interpreted as "does recover".
- The functor M acts on morphisms as follows:
  - $M(p_x) = P_X$ is the population distribution of pre-treatment features.
  - $M(p_{t|x}) = P_{T|X}$ is the Markov kernel, wherein each pre-treatment feature determines a probability that the patient will take accupill.
  - $M(p_{y|t,x}) = P_{Y|T,X}$ is the Markov kernel, wherein for each treatment and each pre-treatment feature, there is a determined probability that the patient will recover.
  - Copying, deleting, and braiding morphisms are taken to the respective copying, deleting, and braiding morphisms. Composition and monoidal product of morphisms are preserved.

Thus, we see that an FCM has a close correspondence with an SCM. Roughly, the action of the functor M on objects corresponds to the assignment of measurable spaces to vertices in an *SCM*; and the action of M on structural morphisms corresponds to the assignment of structural kernels in an *SCM*. Indeed, this is formalized in the following definitions:

Definition 6.9. (FCM generated by an SCM) Let $(G, X_*, P_{*|pa(*)})$ be an SCM. The FCM generated by it is the Markov functor $M : \mathcal{T} \to \mathit{Markov}$ such that

- $\mathcal{T}$ is the causal theory generated by G;
- For each $v \in V(\mathcal{T})$, $M(v) = X_v$;
- For each $v \in V(\mathcal{T})$, $M(p_{v|pa(v)}) = P_{v|pa(v)}$.

Definition 6.10. (SCM underlying an FCM) Let $M : \mathcal{T} \to \mathit{Markov}$ be an FCM. The SCM underlying it is $(G, X_*, P_{*|pa(*)})$, where

- G is the DAG underlying the causal theory $\mathcal{T}$;
- For each $v \in V(\mathcal{T})$, $X_v = M(v)$;
- For each $v \in V(\mathcal{T})$, $P_{v|pa(v)} = M(p_{v|pa(v)})$.

Note, every SCM generates a unique FCM, but in general, there are many FCMs whose underlying SCM is the same.

Now, although the noumenal content of the functor M is determined by what it does on generating objects and structural morphisms, we would like to keep track also of what M does to the *sampling* morphism p of the theory $\mathcal{T}$. Indeed, this will become crucial to our understanding of the observational content of FCMs. So we adopt the following definition, with a suggestive name:

Definition 6.11. (Sampling Distribution) Suppose $M : \mathcal{T} \to \mathit{Markov}$ is a causal model, and p is the sampling morphism of $\mathcal{T}$. Then $M(p)$ is called the **sampling distribution** of M, and is denoted P.

Notice, this is indeed a distribution, since $P = M(p)$ is a Markov kernel $I \to M(V)$: in other words, a probability distribution over the product space

$$M(V) = \prod_{v \in V} M(v)$$

where V is the set of variables of $\mathcal{T}$

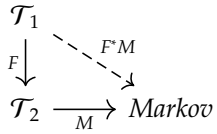
We also note here that, because the sampling distribution depends only on the structural kernels of a model, so two FCMs with the same underlying SCM will share the same sampling distribution.

Multiple models can exist for the same causal theory $\mathcal{T}$. This is not a mere collection of different functors, but in fact it forms a category. Here is what I mean:

Definition 6.12. (Morphism of Causal Models) Let $M_1 : \mathcal{T} \to \mathit{Markov}$ and $M_2 : \mathcal{T} \to \mathit{Markov}$ be FCMs with the same underlying causal theory. A morphism $M_1 \to M_2$ is a monoidal natural transformation $\alpha : M_1 \to M_2$ of the functors.

Thus, the collection of all possible causal models for a theory $\mathcal{T}$ forms a category itself, and we denote this by $\mathit{Mod}(\mathcal{T})$. This operation $\mathcal{T} \mapsto \mathit{Mod}(\mathcal{T})$ is a "contravariant" operation, in the following sense:

Definition 6.13. (Pullback of Causal Models) Let $\mathcal{T}_1, \mathcal{T}_2$ be causal theories, and let $F : \mathcal{T}_1 \to \mathcal{T}_2$ be a morphism of causal theories. Then there is a functor $F^* : \mathit{Mod}(\mathcal{T}_2) \to \mathit{Mod}(\mathcal{T}_1)$, given by pullback along F. That is, if $M : \mathcal{T}_2 \to \mathit{Markov}$ is a model of $\mathcal{T}_2$, then $F^*M : \mathcal{T}_1 \to \mathit{Markov}$ is the model making the following diagram commute:



Thus, making use of both the concept of causal theory morphisms and the concept of model morphisms, we can make sense of meaningful relations between causal models in a purely formal way, instead of relying on prosaic and expert descriptions of these relations.

6.3 Observational Content of Functorial Causal Models

The observational content of an FCM is completely described by its sampling distribution. Let $M : \mathcal{T} \to \mathit{Markov}$ be an FCM. Let p be the sampling morphism of $\mathcal{T}$, and let V be the set of variables. The sampling distribution $P = M(p)$ is a Markov kernel

$$P : I \to M(V).$$

In other words, it is a joint distribution on the measurable spaces $M(v)$ for $v \in V$. This joint distribution fully specifies the observational content of M.

Definition 6.14. (Satisfaction Condition of FCM) Let $M : \mathcal{T} \to \mathit{Markov}$ be an FCM, and let V be its set of variables, and let $P = M(p)$ be its sampling distribution. Let X be a random variable whose range is the measurable space $M(V)$. Then X is said to **satisfy** the model M if its distribution is P.

In particular, consider the string diagram that we built in definition 6.4. Take its image under M in *Markov*. This is a string diagram that represents the sampling distribution P. We think of this string diagram as describing a data-generating *process*, through which the variables $M(V)$ end up having the joint distribution P.

The picture to have in mind is this. Consider again the Accupill example. Refer to diagrams in example 6.2. We imagine building a computer simulation. This simulation first samples a value x in the range of X according to the distribution $P_X = M(p_x)$. Then, it samples a value t in the range of T according to the distribution $P_{T|X}(- \mid x)$. Then, it samples a value y in the range of Y according to the distribution $P_{Y|T,X}(- \mid t, x)$. Finally, it returns the triplet (x, t, y). Thus, when we run this simulation many times, the values returned over the many iterations will have an empirical distribution approximately equal to the sampling distribution P.

The above definition gives the following theorem via the strong law of large numbers:

Corollary 6.2.2. *Let $M : \mathcal{T} \to \mathit{Markov}$ be an FCM, and let X be a random variable satisfying M. Let $f : X \to \mathbb{R}$ be any real-valued function. Let $X_1, \ldots, X_N$ be i.i.d. samples of X. Then*

$$\frac{1}{N} \sum_{k=1}^{N} f(X_k) \overset{a.s.}{\to} \mathbb{E}_P(f)$$

as $N \to \infty$, where P is the sampling distribution of the FCM M.

This then has obvious empirical implications. If a model $M : \mathcal{T} \to \mathit{Markov}$ is intended to represent some real-world process, and f is an observable property of this process, then over many iterations, f should average to approximately its expected value under the sampling distribution P. If it does not, then we have good reason to doubt the correctness of the model M.

Notice, in this definition of satisfaction of an FCM, we did not rely on the existence of any kind of density functions. This represents a major advantage

over definition 4.8. Indeed, we can recover the product-decomposition rule in definition 4.8 via the following theorem:

Theorem 6.3. *Let $M : \mathcal{T} \to$ Markov be an FCM with variables $v_1, \ldots, v_n$. For each j, let μ_j be a measure on the space $M(v_j)$. If each structural kernel $P_{v_j|pa(v_j)}$ in M has a conditional density function $f_{v_j|pa(v_j)}$ with respect to the measure μ_j, then the sampling distribution P of M has a density function f with respect to the product measure $\mu = \bigotimes_j \mu_j$, and*

$$f(x_1, \ldots, x_n) = \prod_{j=1}^{n} f_{v_j|pa(v_j)}(x_j \mid pa(x_j))$$

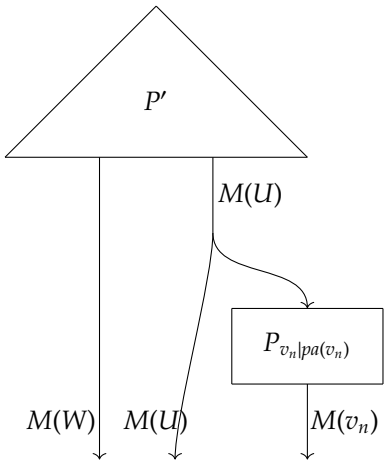
Proof. Since the sampling distribution of M depends only on the structural morphisms of M, which in turn only depends on the underlying SCM, so we may assume without loss of generality that M is an FCM generated by an SCM.

We prove the theorem by induction on the number of variables n. In the base case, suppose that M has only one variable v_1. Then v_1 is necessarily exogenous, and its structural kernel $P_{v_1} : I \to M(v_1)$ has a density function f_{v_1}. The sampling distribution, then, is just P_{v_1} itself. So the desired equality is trivially true:

$$f(x_1) = f_{v_1}(x_1).$$

Now, for the inductive step, let $M' : V' \to$ *Markov* be the submodel of M generated by the SCM with all the same structural kernels and variables, except v_n is removed. Then M' is a model with $n-1$ variables. Suppose, for induction, that the theorem is true on M'.

Then, partition $V(M') = v_1 \otimes \cdots \otimes v_{n-1}$ into $U \otimes W$, where $U = pa(v_n)$, and $W = V(M') \backslash pa(v_n)$. Thus, $U \otimes W = v_1 \otimes \cdots \otimes v_{n-1}$. Then the sampling distribution P for M can be represented as



where P' is the sampling distribution for M'. Now, let $Q : M(U) \to M(U) \otimes M(v_n)$ be the Markov kernel corresponding to the "right branch" of the above diagram.

In other words,

$$Q = (id_{M(U)} \otimes P_{v_n|pa(v_n)}) \circ \Delta_{M(U)}.$$

Then the above string diagram says that

$$P = (id_{M(W)} \otimes Q) \circ P'.$$

In order to compute a density function for P, let's pick measurable sets E_W, E_U, E_n in the spaces $M(W), M(U), M(v_n)$ respectively. Then we can easily compute

$$\begin{aligned} Q(E_U \times E_n \mid x_U) &= \int_U id_{M(U)}\left(E_U \mid x_U^{(1)}\right) P_{v_n|pa(v_n)}\left(E_n \mid x_u^{(2)}\right) d\Delta_n\left(x_U^{(1)}, x_U^{(2)} \mid x_U\right) \\ &= id_{M(U)}(E_U \mid x_U) P_{v_n|pa(v_n)}(E_n \mid x_U) \\ &= \mathbb{I}_{x_U \in E_U} \int_{E_n} f_{v_n|pa(v_n)}(x_n \mid x_U)\, dx_n \end{aligned}$$

and therefore

$$\begin{aligned} P(E_W \times E_U \times E_n) &= (id_W \otimes Q) \circ P'(E_W \times E_U \times E_n) \\ &= \int_W \int_U id_W(E_W \mid x_W) Q(E_U \times E_n \mid x_U, x_n)\, dP'(x_W, x_U) \\ &= \int_W \int_U \mathbb{I}_{x_W \in E_W} \left[\mathbb{I}_{x_U \in E_U} \int_{E_n} f_{v_n|pa(v_n)}(x_n \mid x_U)\, dx_n \right] dP'(x_W, x_U) \\ &= \int_{E_W} \int_{E_U} \int_{E_n} f_{v_n|pa(v_n)}(x_n \mid x_U)\, dx_n\, dP'(x_W, x_U) \end{aligned}$$

Now, by the inductive hypothesis, we know that

$$dP'(x_W, x_U) = \prod_{j=1}^{n-1} f_{v_j|pa(v_j)}(x_j \mid pa(x_j))\, dx_1 \ldots dx_{n-1}$$

where $x_1, \ldots, x_{n-1}$ are the corresponding components of x_W and x_U under the correspondence $U \otimes W = v_1 \otimes \cdots \otimes v_{n-1}$. Thus, we obtain that

$$\begin{aligned} P(E_W \times E_U \times E_n) &= \int_{E_W} \int_{E_U} \left[\int_{E_n} f_{v_n|pa(v_n)}(x_n \mid x_U)\, dx_n \right] \prod_{j=1}^{n-1} f_{v_j|pa(v_j)}(x_j \mid pa(x_j))\, dx_1 \ldots dx_{n-1} \\ &= \int_{E_W} \int_{E_U} \int_{E_n} \prod_{j=1}^{n} f_{v_j|pa(v_j)}(x_j \mid pa(x_j))\, dx_1 \ldots dx_n \end{aligned}$$

Thus, the conditional density function of P is

$$f(x_1, \ldots, x_n) = \prod_{j=1}^{n} f_{v_j|pa(v_j)}(x_j \mid pa(x_j)),$$

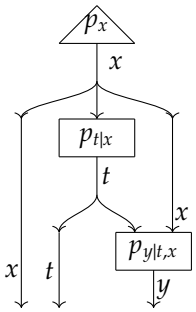
as required. □

6.4 Interventional Content of Functorial Causal Models

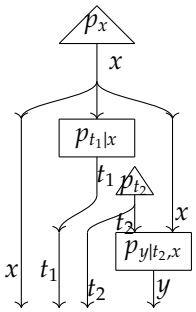
Recall, in section 4.4, we defined the Single World Intervention Graph operation on SCMs. For each SCM $M = (G, X_*, P_{*|pa(*)})$, and each possible intervention $t \in X_T$, the SWIG operation yields another SCM $M(\text{do}\, T = t)$. The *interventional* content of the SCM M is then nothing more or less than the statement that the world would behave according to the *observational* content of the SCM $M(\text{do}\, T = t)$, if the intervention $T = t$ were applied. We will now develop the interventional content of FCMs in a parallel way.

Definition 6.15. (SWIT for Theories Generated by DAGs) Let $\mathcal{T}$ be a causal theory generated by a DAG G, and let $T \subset V(G)$ be the set of variables that we will intervene upon. Let G' be the DAG obtain via the SWIG operation on T. The single-world intervention theory $\mathcal{T}(\text{do}\, T)$ is the causal theory generated by G', with an additional isomorphism $v_1 \cong v_2$ for each $v \in T$.

Example 6.3. (Accupill SWIT) Recall from example 6.2 that the causal theory $\mathcal{T}$ representing the accupill scenario has sampling morphism as follows:



The SWIT operation on $\mathcal{T}$ on the set $T = \{t\}$ will yield a causal theory $\mathcal{T}(\text{do}\, t)$, which contains four variables x, t_1, t_2, y, and the sampling morphism will be



The theory $\mathcal{T}(\text{do}\, t)$ will also have an isomorphism $t_1 \cong t_2$, reflecting the fact that the range of treatment options for the patient is the same as the range of treatment options that the experimenter can select from.

The theories $\mathcal{T}$ and $\mathcal{T}(\text{do}\, T)$ are obviously intimately connected. There exist natural morphisms going in both direction, which we shall describe now. We

will first show their existence, and then delve into the intuitive meaning behind these functors.

Theorem 6.4. *Let G be a DAG, and let $\mathcal{T}$ be the causal theory generated by G. Let $T \subset V(G)$ be any set of variables. Then there is a lax causal theory morphism*

$$F : \mathcal{T}(\text{do}\, T) \to \mathcal{T}$$

such that for $v \in T^c$, F sends the single variable corresponding to v back to v, and for $t \in T$, the variables t_1, t_2 are both sent to t.

Proof. Let G' be the DAG obtained by the SWIG operation on G. So there is a graph homomorphism $\phi : G' \to G$ which is locally isomophic at every vertex outside of T, bijective on the set of edges, and which satisfies $\phi^{-1}(t) = \{t_1, t_2\}$ for $t \in T$.

Now, G' is the DAG associated to $\mathcal{T}(\text{do}\, T)$. However, $\mathcal{T}(\text{do}\, T)$ is not the theory generated by G', because it contains an extra isomorphism $t_1 \cong t_2$ for each $t \in T$.

Let $\mathcal{T}_{G'}$ be the causal theory generated by G'. Then there is a lax causal theory morphism

$$F_\phi : \mathcal{T}_{G'} \to \mathcal{T}$$

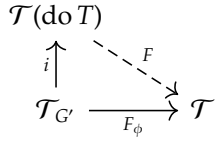
induced by the graph homomorphism ϕ, as described in theorem 6.2.

Because G' is the DAG associated to $\mathcal{T}(\text{do}\, T)$, there is also a strict causal theory morphism

$$i : \mathcal{T}_{G'} \to \mathcal{T}(\text{do}\, T)$$

as described in theorem 6.1.

Now, because for every $t \in T$, F_ϕ sends t_1, t_2 to t, so F_ϕ has an extension along i, defined by sending the extra isomorphisms $t_1 \cong t_2$ to the identity maps $t = t$. This extension is our desired morphism F.



□

Theorem 6.5. *Let G be a DAG, and $\mathcal{T}$ be the causal theory generated by G. Let $T \subset V(G)$ be any set of variables. Then there is a lax causal theory morphism*

$$H : \mathcal{T} \to \mathcal{T}(\text{do}\, T)$$

such that for $t \in T$, the variable t gets sent to t_2 (that is, the corresponding variable in $\mathcal{T}(\text{do}\, T)$ with no parents), and for $v \in T^c$, v just gets sent to the single variable corresponding to it in $\mathcal{T}(\text{do}\, T)$.

Proof. The action of H on objects has been given. We just need to specify the action of H on morphisms. Since $\mathcal{T}$ is the causal theory generated by G, so it suffices to specify the action of H on structural morphisms.

Pick variable v. The set of parents $pa(v)$ may contain variables from both T and T^c. Let $U = pa(v) \cap T$, and $W = pa(v) \cap T^c$. Then as objects,

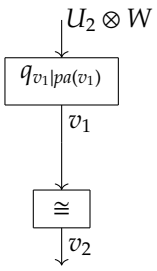
$$pa(v) = U \otimes W.$$

Let $U_2 = \{t_2 : t \in U\}$ be the set of variables in $\mathcal{T}(\text{do } T)$ corresponding to U, with no parents. Then,

$$H(pa(v)) = U_2 \otimes W$$

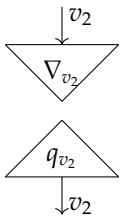
Suppose $v \notin T$. Then $H(v) = v'$, where v' is the single variable corresponding to v in $\mathcal{T}(\text{do } T)$. So $pa(H(v)) = U_2 \otimes W = H(pa(v))$ in $\mathcal{T}(\text{do } T)$. So we may declare $H(p_{v|pa(v)})$ to be the structural morphism $q_{v'|pa(v')}$ at v' in $\mathcal{T}(\text{do } T)$. So at non-treated variables $v \notin T$, H is "locally" strict.

On the other hand, suppose $v \in T$. Then $H(v) = v_2$, where v_2 is the corresponding variable in $\mathcal{T}(\text{do } T)$ with no parents. So declare $H(p_{v|pa(v)})$ to be the morphism

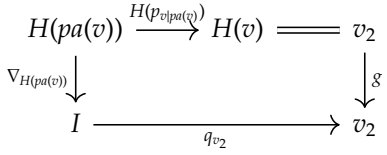


where v_1 is the corresponding variable in $\mathcal{T}(\text{do } T)$ with no children, and $q_{v_1|pa(v_1)}$ is the structural morphism of $\mathcal{T}(\text{do } T)$ at v_1, and the isomorphism $v_1 \cong v_2$ is the one given in the construction of the theory $\mathcal{T}(\text{do } T)$.

Then, if $g : v_2 \to v_2$ is the "forget-and-regenerate" morphism



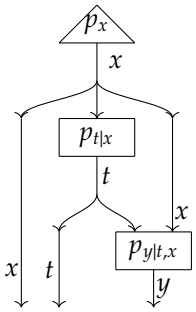
where $q_{v_2} : I \to v_2$ is the structural morphism of $\mathcal{T}(\text{do } T)$ at v_2, then the following diagram commutes:



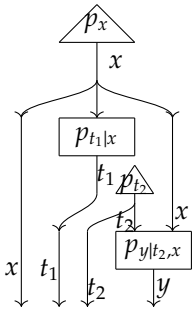
Thus, $H : \mathcal{T} \to \mathcal{T}(\mathrm{do}\, T)$ is a lax causal theory morphism. □

What is the intuitive meaning behind the two causal theory morphisms described in the above theorems? We use the Accupill example once again.

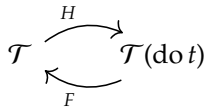
Example 6.4. Recall that, in the accupill example, we originally gave a causal theory $\mathcal{T}$ whose sampling morphism was



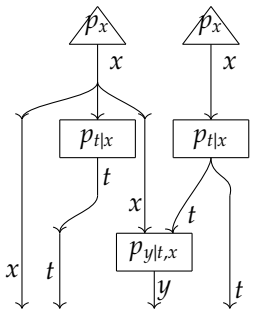
We then performed the SWIT operation on $\mathcal{T}$, and obtained a causal theory $\mathcal{T}(\mathrm{do}\, t)$ whose sampling morphism was



Now, consider the morphisms

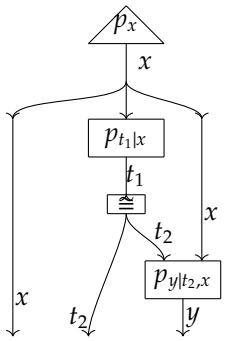


Then F sends the sampling morphism p in $\mathcal{T}(\mathrm{do}\, t)$ to the following morphism in $\mathcal{T}$:



In other words, F gives the experimenter a behavior that is exactly like any patient. The experimenter will have features x drawn from the same population as the patient, and the experimenter will choose the treatment t for the patient as if she were choosing the treatment for herself.

On the other hand, H sends the sampling morphism q in $\mathcal{T}$ to the following morphism in $\mathcal{T}(\text{do}\, t)$:



In other words, H interprets the patient as *both* the patient *and* the experimenter: the experimenter just happens to act precisely according to the patient's will in every case.

If we consider the pullback functors associated with the theory morphisms F and H, the situation becomes even clearer. There are functors between models of $\mathcal{T}$ and $\mathcal{T}(\text{do}\, t)$ via pullback along F and H:

$$Mod(\mathcal{T}) \underset{H^*}{\overset{F^*}{\rightleftarrows}} Mod(\mathcal{T}(\text{do}\, t))$$

If M is a model of $\mathcal{T}$, then F^*M is the model of $\mathcal{T}(\text{do}\, t)$ defined by

$$\mathcal{T}(\text{do}\, t) \xrightarrow{F} \mathcal{T} \xrightarrow{M} Markov$$

So in this model, the structural distribution P_{T_2} is just the same distribution as the marginal sampling distribution for T_1. So F^*M models the experimenter as

behaving exactly like any patient, with features drawn from the same population, and with decision making process just like the patient.

On the other hand, if M' is any model of $\mathcal{T}(\text{do}\, t)$, then H^*M' is the model of $\mathcal{T}$ defined by

$$\mathcal{T} \xrightarrow{H} \mathcal{T}(\text{do}\, t) \xrightarrow{M'} \mathit{Markov}$$

So in this model, the structural kernel $P_{T|X}$ is the kernel $X \to T_2$ that can be constructed in the image of M'. So H^*M' is simply the model of what *would* have happened if, contrary to what M' states, the experimenter simply reproduced the patient's will exactly.

Now that we have constructed the SWIT operation on causal theories generated by DAGs, we are still missing a piece of data: the SWIT operation $\mathcal{T} \mapsto \mathcal{T}(\text{do}\, T)$ does not record the information of *how* the experimenters will pick a treatment for the patient. It only records *which* variables the experimenters will intervene on. This further piece of information, of course, is not part of the syntax of causal theories, but rather is part of the semantics of causal models. So we define:

Definition 6.16. (SWIFT for Causal Models) Let $M : \mathcal{T} \to \mathit{Markov}$ be an FCM, where $\mathcal{T}$ is a causal theory generated by a DAG. Let $T = \{T_1, T_2, \dots\}$ be the set of variables we will intervene upon. Let $\tau = (\tau_1, \tau_2, \dots)$ be a point in the space $M(T)$. Then the SWIFT operation defined on the model M gives a new FCM

$$M(\text{do}\, T = \tau) : \mathcal{T}(\text{do}\, T) \to \mathit{Markov}$$

where:

1. For variables $v \in V(\mathcal{T}(\text{do}\, T))$, let v' be the unique variable in $V(\mathcal{T})$ corresponding to v. Then

$$M(\text{do}\, T = \tau)(v) := M(v')$$

2. For structural morphisms $p_{v|pa(v)}$ in $\mathcal{T}(\text{do}\, T)$:

   (a) If v corresponds to a variable $v' \notin T$, then

   $$M(\text{do}\, T = \tau)(p_{v|pa(v)}) := M(q_{v'|pa(v')})$$

   where $q_{v'|pa(v')}$ is the structural morphism at v' in $\mathcal{T}$. Note that this is well defined, because by the previous item, we have $M(pa(v)) \cong M(pa(v'))$.

   (b) Pick any variable $t \in T$. Suppose v is the variable in $\mathcal{T}(\text{do}\, T)$ corresponding to t with no children. Then by the previous item, $M(pa(v)) \cong M(pa(t))$, and so we let

   $$M(\text{do}\, T = \tau)(p_{v|pa(t)}) := M(q_{t|pa(t)})$$

   where $q_{t|pa(t)}$ is the structural morphism at t in $\mathcal{T}$.

(c) If v is the variable in $\mathcal{T}(\text{do}\, T)$ corresponding to $t \in T$ with no parents, then let

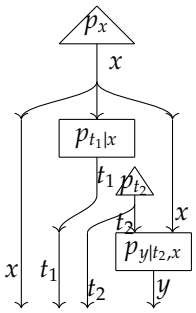
$$M(\text{do}\, T = \tau)(p_v) := \delta_{\tau_i}$$

where $\delta_{\tau_i} : I \to M(t)$ is the distribution that assigns probability mass 1 to the point $\tau_i \in M(t)$.

3. For the extra isomorphisms $v_1 \cong v_2$ in the theory $\mathcal{T}(\text{do}\, T)$, $M(\text{do}\, T = \tau)$ sends them to the identity maps $M(\text{do}\, T = \tau)(v_1) = M(\text{do}\, T = \tau)(v_2)$.

Example 6.5. (Accupill SWIFT) Recall from example 6.2 that we gave a causal model M to the accupill causal theory $\mathcal{T}$. I will not reiterate the content of that model here.

Now, suppose that an experimenter comes along, and assigns the treatment τ to the patient, with certainty. The patient no longer gets to choose what treatment they take. We know that the causal theory representing this situation is $\mathcal{T}(\text{do}\, t)$. That theory has sampling morphism



We now also know the causal model representing this situation. It is the functor

$$M(\text{do}\, t = \tau) : \mathcal{T}(\text{do}\, t) \to \mathit{Markov}$$

1. The functor $M(\text{do}\, t = \tau)$ acts on objects in $\mathcal{T}(\text{do}\, t)$ as follows:
   (a) $M(\text{do}\, t = \tau)(x)$ is the measurable space of all possible pre-treatment features of the patient.
   (b) $M(\text{do}\, t = \tau)(t_1) = M(\text{do}\, t = \tau)(t_2)$ are both the discrete space consisting of two points: a point representing "accupill received", and a point representing "accupill not received".
   (c) $M(\text{do}\, t = \tau)(y)$ is the measurable space of all possible outcomes (recovered, not recovered).
2. The functor $M(\text{do}\, t = \tau)$ acts on morphisms in $\mathcal{T}(\text{do}\, t)$ as follows:
   (a) $\mathbb{P}_x = M(\text{do}\, t = \tau)(p_x)$ is the population distribution of the pre-treatment features from which the patient is drawn.

(b) $\mathbb{P}_{t_1|x} = M(\mathrm{do}\, t = \tau)(p_{t_1|x})$ is the Markov kernel where, given a patient's pre-treatment feature, there is a probability that the patient *would wish* to receive accupill.

(c) $\mathbb{P}_{t_2} = M(\mathrm{do}\, t = \tau)(p_{t_2}) = \delta_\tau$ is the distribution assigning probability 1 to the treatment τ. This is, in other words, the distribution of the treatment received by the patient (possibly contrary to her wishes).

(d) $\mathbb{P}_{y|t_2,x} = M(\mathrm{do}\, t = \tau)(p_{y|t_2,x})$ is the Markov kernel where, given the treatment actually received, and given the patient's pre-treatment feature, there is a definite probability for every possible outcome.

(e) $M(\mathrm{do}\, t = \tau)$ sends the additional isomorphism $\alpha : t_1 \to t_2$ in $\mathcal{T}(\mathrm{do}\, t)$ to the identity map of the space $M(t_1) = M(t_2)$.

Theorem 6.6. *Let $M : \mathcal{T} \to$ Markov be a causal model generated by a DAG. Let $M(\mathrm{do}\, T = t)$ be the SWIFT of M with respect to the intervention $T = t$. Let $H : \mathcal{T} \to \mathcal{T}(\mathrm{do}\, T)$ be the causal theory morphism described in theorem 6.5. Then*

$$M = H^* M(\mathrm{do}\, T = t).$$

Proof. Since $\mathcal{T}$ is a causal theory generated by a DAG, so it suffices to show that M and $H^*M(\mathrm{do}\, T = t)$ agree on all structural morphisms.

First, suppose $v \notin T$ is a non-treated variable. Let $p_{v|pa(v)}$ be the structural morphism at v in $\mathcal{T}$. Then $H : \mathcal{T} \to \mathcal{T}(\mathrm{do}\, T)$ sends $p_{v|pa(v)}$ to the structural morphism $q_{v'|pa(v')}$ in $\mathcal{T}(\mathrm{do}\, T)$, where v' is the unique variable in $\mathcal{T}(\mathrm{do}\, T)$ corresponding to v. So

$$\begin{aligned} H^*M(\mathrm{do}\, T = t)(p_{v|pa(v)}) &= M(\mathrm{do}\, T = t)(H(p_{v|pa(v)})) \\ &= M(\mathrm{do}\, T = t)(q_{v'|pa(v')}) \\ &= M(p_{v|pa(v)}) \end{aligned}$$

as required.

Next, suppose $v \in T$ is a treated variable. Then H sends $p_{v|pa(v)}$ to the morphism $\alpha \circ q_{v_1|pa(v_1)}$, where v_1 is the variable corresponding to v that has no children, and $\alpha : v_1 \to v_2$ is the isomorphism. Then

$$\begin{aligned} H^*M(\mathrm{do}\, T = t)(p_{v|pa(v)}) &= M(\mathrm{do}\, T = t)(H(p_{v|pa(v)})) \\ &= M(\mathrm{do}\, T = t)(\alpha \circ q_{v_1|pa(v_1)}) \\ &= M(\mathrm{do}\, T = t)(\alpha) \circ M(\mathrm{do}\, T = t)(q_{v_1|pa(v_1)}) \\ &= id_{M(v)} \circ M(p_{v|pa(v)}) \\ &= M(p_{v|pa(v)}) \end{aligned}$$

as required.

Thus, $M = H^*M(\mathrm{do}\, T = t)$. □

Now that we have an account of the SWIFT operation of models of causal theories generated by DAGs, we know the interventional content of those

models. However, it is natural at this point to ask: can we generalize this account to all causal theories, including those not generated by DAGs?

To do that, we require the following conjecture.

Conjecture 6.1. *Let $\mathcal{T}$ be a causal theory, and G be the DAG associated with it. Let $\mathcal{T}_G$ be the causal theory generated by G. Let $\mathcal{T}_G(\mathrm{do}\, T)$ be the SWIT of $\mathcal{T}$ with respect to the set of variables T. Then there is a push-out in the category of causal theories*

$$\begin{array}{ccc} \mathcal{T}_G & \xrightarrow{\ i\ } & \mathcal{T} \\ {\scriptstyle H}\downarrow & & \vdots\downarrow \\ \mathcal{T}_G(\mathrm{do}\, T) & \dashrightarrow & \mathcal{T}' \end{array}$$

where i is the strict causal theory morphism defined in theorem 6.1, and H is the lax causal theory morphism defined in theorem 6.5.

We shall call this pushout $\mathcal{T}'$ the SWIT of $\mathcal{T}$ with respect to the set of variables T, and denote it $\mathcal{T}(\mathrm{do}\, T)$.

Proving this conjecture, and extending the SWIFT operation to general causal theories, will be beyond the scope of this thesis. We leave this to be addressed in a further work.

6.5 Concluding Remarks

This new formalism of FCMs retains the important properties of the SCM formalism: it articulates the generative dependencies between variables as Markov kernels, and it specifies the composite structure built out of these generative dependencies. However, it extends it. An FCM contains more data than an SCM does. Indeed, as we have seen, every SCM can be represented as an FCM, but the converse is not true.

This extra data allows us to do several things. First, it allows us to do away with the requirement that all Markov kernels admit conditional density functions. There are, as we've mentioned, important examples where Markov kernels fail to be absolutely continuous, because they are discrete over some parts of the domain, and continuous over other parts (think of example 3.6). The FCM formalism gives a fully rigorous language for dealing with the compositionality of such Markov kernels.

Second, it allows us to articulate relations between FCMs in a formal way. We developed two such relations: morphisms between causal theories, and morphisms between models of the same theory. We have also seen that this notion can be applied to make rigorous the relations between a model and its SWIFT.

However, because FCMs are more complex data structures than SCMs, they have become even more burdensome to work with. It is more difficult to visualize, interact with, or discover good FCMs. In addition to this, several further difficulties remain, including the fact that conjecture 6.1 has yet to be proven. These difficulties are discussed in detail in section 7.2.

7 Conclusion

7.1 Summary

We began with a philosophical survey of the concept of causation, which gave us three normative insights about mathematical causal models:

1. A good causal model should not only describe the world as it is observed, but it should also describe what *would* happen, if we were to *intervene* in this world as agents. In other words, it must encode the *generative dependencies* between variables, and not just the *correlational regularities*.

2. A good causal model should not be a black box, but should have *internal structure*. In other words, it must tell us how various generative dependencies come together to form a chain or network, and how this overall structure generates consequences that we can, in principle, observe.

3. Differing causal models should not be viewed as stand-alone ideas, but should bear *meaningful relations* to each other.

We then surveyed three different approaches to the mathematical modelling of causation:

1. Potential outcome models, or RCMs, are very powerful in clinical settings, in which exactly one cause-effect pair is being modelled. This framework articulated for us the notion of a Markov kernel, which allowed us to fulfill requirement number 1 from above. However, it does not allow for the composing of causal structure, nor does it formalize any notion of relations between causal models.

2. Structural causal models, or SCMs, extended the RCM framework by articulating the compositionality of Markov kernels, at least in the case where these kernels admit conditional density functions. They are powerful in situations where many variables come together to form a system of causal effects. However, it still does not formalize any meaningful relations between differing models.

3. Functorial causal models, or FCMs, extend the SCM framework. It reinterprets causal models as *models* (in the logical sense of *model*) of causal *theories*. By doing so, it endows both causal theories and causal models with categorical structures, and therefore naturally yields a formalism for the meaningful relations between models.

Thus, the three normative insights are gradually fulfilled by the three different formalisms presented in this thesis.

However, the price we paid is complexity. With each more powerful formalism came a more burdensome set of machinery. The more burdensome the machinery, the more difficult it becomes to visualize, interact with, and discover good causal models. Thus, a balancing act is required.

7.2 Limitations and Future Work

We will now discuss some directions for future work in regards to FCMs, starting from the most concrete, towards the more abstract directions.

First, as noted in section 6.4, the interventional content for those causal theories not generated by DAGs still remains to be developed. This work will have to start with a proof (or an adjustment) of the statement in conjecture 6.1.

Second, as we observed in section 6.1, the structure of a causal theory is quite a lot more cumbersome than that of a simple DAG. Of course, it is precisely this structure that allowed us to articulate the meaningful relations between causal theories. Nevertheless, it is necessary to develop the graphical calculus of causal models further, so that one can interact with them more easily. We believe that the language of the sampling morphism and the marginal sampling morphisms goes a long way to simplifying the understanding of causal theories, but more work must be done here.

Third, the notion of a lax causal theory morphism is, as defined in section 6.1, very permissive. It does not place any constraints on the transitioning morphisms f_v and g_v, as they figure in definition 6.6. It only requires that there exist such morphisms, and that such morphisms are included in the data of a lax causal theory morphism. Although strong and strict causal theory morphisms serve as less permissive notions, it would be helpful to develop some intermediate notion, less permissive than a lax causal theory morphism, and more permissive than a strong causal theory morphism, by imposing some relevant naturality condition on the transitioning morphisms. A correct articulation of these naturality conditions requires further inquiry.

Fourth, we must take a look at whether the problematic cases encountered in chapter 2 have really been addressed. I claim that they have not all been addressed satisfactorily. The problem of early pre-emption (see example 2.4) is successfully addressed by SCMs and FCMs alike, because both of these formalisms allow for causation to be transitive. However, this feature also makes these formalisms vulnerable to the problem of non-transitivity (see example 2.5). In both SCMs and FCMs, if a variable A is a parent of variable B, and B is a parent of C, then A is an ancestor of C, and therefore C must be generatively dependent on A. Does that mean that the boulder rolling towards the hiker is a cause for the hiker's ability to continue the hike? In a sense, SCMs and FCMs bite the bullet: yes, they say, because once we know of the possibility of a boulder rolling towards the hiker, we cannot compute the probability that the hiker can continue the hike, without first knowing whether the boulder does roll towards the hiker!

Both SCMs and FCMs provide a sort of dodgy answer to the problem of late pre-emption (see example 2.6). They say, in response to that problem, that both Sarah's and Taro's rock throwing are causes of the breaking of the window, even though Taro's rock did not even touch the window. This is because any successful causal model of the situation must encode the fact that, if we were to intervene in the situation, and stop Sarah from throwing the rock, then Taro's rock would have broken the window. Thus, any successful causal model

will necessarily make both Sarah's rock throwing, and Taro's rock throwing, parents of the variable encoding the breaking of the window. The breaking of the window, in other words, is generatively dependent on the *joint* behaviour of the two.

Notice that the failure to adequately answer the problem of non-transitivity and the problem of late pre-emption yield a common consequence: if a notion of causation can be modelled by SCMs and FCMs, then that notion of causation is inadequate for the assignment of blame and praise in a moral context. It is unreasonable to bite the bullet and say that the boulder is praiseworthy for causing the hiker to continue on her journey, and it is unreasonable to say that Sarah and Taro are equally blameworthy for the breaking of the window. Thus, much further inquiry in both philosophy and mathematics needs to be made before these problems can be fully addressed.

References

[1] D. Armstrong. *A Theory of Universals*. Cambridge University Press, 1978.

[2] John Carroll. The humean tradition. *The Philosophical Review*, 99(2):185–219, 1990.

[3] Brendan Fong. Causal theories: A categorical perspective on bayesian networks. *arXiv preprint arXiv:1301.6201*, 2013.

[4] Tobias Fritz. A synthetic approach to markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, 2020.

[5] Dan Geiger, Thomas Verma, and Judea Pearl. Identifying independence in bayesian networks. *Networks*, 20(5):507–534, 1990.

[6] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[7] David Hume. *A treatise of human nature*. Courier Corporation, 2003.

[8] Guido W Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000.

[9] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[10] Arthur B Kahn. Topological sorting of large networks. *Communications of the ACM*, 5(11):558–562, 1962.

[11] F William Lawvere. Functorial semantics of algebraic theories. *Proceedings of the National Academy of Sciences of the United States of America*, 50(5):869, 1963.

[12] David Lewis. *Counterfactuals*. John Wiley & Sons, 1973.

[13] David Lewis. Causation as influence. *The Journal of Philosophy*, 97(4):182–197, 2004.

[14] John Stuart Mill. *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation*. Longmans, Green, Reader, and Dyer, 1884.

[15] Evan Patterson. *The algebra and machine representation of statistical models*. Stanford University, 2020.

[16] Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

[17] Emily Riehl. *Category theory in context*. Courier Dover Publications, 2017.

[18] Sheldon M Ross. *Introduction to probability models*. Academic press, 2014.

[19] Gavin J Seal. Tensors, monads and actions. *arXiv preprint arXiv:1205.0101*, 2012.

[20] Bas C Van Fraassen. Armstrong on laws and probabilities. *Australasian Journal of Philosophy*, 65(3):243–260, 1987.