UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

Master's Thesis

Eigil Fjeldgren Rischel

The Category Theory of Causal Models

Advisor: Sebastian Weichwald

August 30, 2020

Abstract

Building on work of Rubenstein et al, we consider a notion of structure-preserving transformations between structural causal models. We describe a *category*, in the sense of category theory, of such models, and explore the properties of this category, FinMod. We further generalize this theory to encompass *approximate* transformations, which only preserve the causal structure up to a certain error, and a larger class of *profinite models*, which includes for example graphical models on infinite graphs. The language of category theory provides a natural setting for these generalizations, and allows for certain definitions, such as the error of a transformation between infinite models, to be made automatically.

Contents

1 Introduction

> By convention sweet is sweet, bitter is bitter, hot is hot, cold is cold, color is color; but in truth there are only atoms and the void.
>
> Democritus

It is a common aphorism in statistics that "all models are wrong" (Often attributed to George Box, see [23]). Any statistical model should be not regarded as a precise description of the universe, or even of the domain under consideration, but as a useful approximation. It is natural to ask, in this situation, which properties of a model makes it "useful". Of course, this is a question about the relationship of a model to "the universe", and since we do not have a mathematical description of the universe, it is not clear that we can expect this question to have a precise answer. Nevertheless, we can ask a related question: when is it permissible to replace one model with another, simplified one? Instead of asking whether a model is useful relative to the universe, we ask whether it's useful relative to some larger model, which stands in for the universe. In [21], Rubenstein et al consider a potential answer to this question, in the context of structural causal models: one model should be a (measurable) function of the other, in a way that preserves not only the observed probability, but also the interventional probabilities.

Inherent in this answer is the perspective that "utility" or "exactness" is not just a relation between two models, but rather a *property* of a *transformation* between two models. This leads to the idea that one could study causal models by studying the system of such transformations. This is the perspective of *category theory* — that to study mathematical objects, we should study the transformations between them. The goal of this thesis is to use this language to develop a theory of transformations between causal models.

Though the application of category theory to the domain of causal modeling is not exactly mainstream, there is an existing body of work in this domain. In [4], Fong develops the theory of directed acyclic graph (DAG) models using "syntactical categories" associated to a DAG. This is further developed by Jacobs–Kissinger–Zanasi in [7], where they show how to use *string diagram* manipulations to carry out causal inference. More broadly, Fritz has begun an ambitious program of developing probability theory in the language of categories, which is described in [5]. Working in this framework, Patterson ([17]) has developed a precise analogy between statistics and universal algebra, where a *statistical model* becomes a model of a *theory*, in the sense of logic. This analogy is also present in the work of Fong and Jacobs–Kissinger–Zanasi, although not discussed explicitly.

In this thesis, we take a different approach from these authors — where the transformations they are concerned with are generally Markov kernels of some sort, and a causal model is represented as a collection of such transformations, we are concerned with transformations *between* causal models. At first, we consider only transformations between finite models — i.e. models with a finite number of variables, each with a finite number of possible values. This simplifies the theory considerably. For instance, it allows us to use Jensen–Shannon distance (Definition 3.18) without getting into the weeds of differential entropy. We consider a general notion of transformation with no requirements on "consistency" between the two models — such a transformation has an associated *error*, which is defined as the prediction error when using the high-level model rather than the low-level model. The "exact" transformations of Rubenstein et al are then those transformations with error zero. This provides a partial answer to a question posed in [21], about how to define a useful notion of "approximate" transformation between causal models.

Finally, we look for a generalization of this theory to more general models than just finite ones. Rather than looking ad-hoc for some sufficiently constrained class of graphical models where the theory still makes sense, we use a piece of categorical technology called *(enriched) copresheaves* to construct a useful generalization in a single line. These models, which we call *profinite models*, are quite abstract, but still contain enough structure to reasonably be interpreted as a model.

1.1 Guide to reading

Because this thesis concerns two fields which do not have a lot of contact, I've put a lot of effort into structuring it in such a way that it will be useful and

readable for people with a wide variety of backgrounds. Those aspects of the theory of causal modeling that I touch here are elementary enough that they can probably be picked up along the way, even by people with no significant experience with the field[1]. However, category theory is notoriously abstract and hard to wrap your head around, so in section 2, I try to sketch the basics for the benefit of novices. People with some familiarity with category theory can probably skip most of this section at first and simply return when needed. The exceptions are Definitions 2.8, 2.24 and 2.43, concerning the category Err of "error spaces", which is probably unfamiliar. In subsection Section 2.5, I prove some necessary theorems about the enriched category theory of Err. I suggest that novices skip this section — the relevant consequences will be spelled out elsewhere in the text. Section 3 gives a brief discussion of certain aspects of probability theory, including how to treat them in a categorical language. The most important result here is Proposition 3.19, which we might summarise as saying that the Jensen-Shannon distance on probability distributions is *compositional*, in the sense that it is compatible with the composition of stochastic matrices. This is a key result in setting up a useful theory of approximate abstractions. This section should probably be read by everyone — the probability theory here is not that interesting to experts, but it will be helpful to familiarize yourself with the categorical approach that I will use in the rest of the thesis. In section 4, we are finally ready to approach the main topic of the thesis, by defining the Err-enriched category of finite causal models and abstractions. I also give several natural examples of abstraction. In section 5, I develop a broad generalization of the technology in section 4, namely the so-called profinite models. These contain, for example, graphical models with an infinite number of nodes (Example 5.9), discrete dynamical systems (Example 5.14), and certain continuous-time systems like Poisson processes (Example 5.13). The correct notion of abstraction between such models, and the correct measure of error, is derived naturally from categorical machinery. I also describe how to work concretely with this very abstract notion of model, in terms of "variables" and "probabilities". Sections 4 & 5 last two sections contain the main ideas of the thesis, and should be read by everyone.

When certain ideas are very abstract, I try to provide reasonable explanations. Hence, if some proposition seems impossibly technical, hopefully you can skip it and still get by in the rest of the thesis. I have marked certain bits with ⊙⊙ two concerned eyes, as here on the left. This means that the marked sections are extra-technical or abstract, and might be skipped by people who are not familiar with category theory.

1.2 Acknowledgements

First of all, I would like to thank my advisor, Sebastian Weichwald. Thank you for taking me up on the crazy idea that eventually became this thesis, and for consistently useful discussions about the causal side of things. This certainly

[1] After all, the thesis was written by such a person.

would not have existed without you. Second, I would like to thank Tobias Fritz for useful conversations during the development of this thesis. Not everything we discussed made it in, but the finished product still owes a lot to his input. Third, thanks to Fosco Loregian for pointing me to a crucial reference at exactly the right time. I've had useful mathematical conversations with a large number of people in the six months it took to write this thesis. If I've forgotten to mention you here, rest assured that your input was appreciated nonetheless. Last, but certainly not least, I would like to thank Julie for keeping me sane through the past six months. The combination of a global pandemic and a thesis deadline was not the best thing I've tried, but you managed to get me through it. For that, I'm thankful.

2 Category Theory

The goal of this section is to present the categorical technology that will be needed in the rest of the thesis. Since there are quite a few fairly technical ideas here, it is probably not possible for readers with no prior knowledge to assimilate them all in such a short amount of time. Therefore, this section will not attempt to teach you all the category theory that's used. The goal is instead to present a clear enough *sketch* of the topic that most of the rest of the thesis can be followed. As an instance of this philosophy, I won't describe *enriched categories* (Section 2.3) in full detail. Instead, I try to indicate the general idea, and spell out those definitions we need in the cases that we consider. The hope is that those without a background in category theory will be able to follow the material, while all the technical details are still available to those who do have this knowledge.

There are several excellent, readable introductions to category theory available — I will mention in particular Riehl's book [20], and Perrone's lecture notes [18]. Both are very approachable even to readers with no background in "abstract" mathematics.

The basic idea of category theory is

1. To study mathematical objects (like groups, vector spaces, topological spaces, etc) by studying the "structure-preserving maps" between them (group homomorphisms, linear maps, continuous maps).

2. To make this study systematic, by abstracting out the general features of a "system of objects and morphisms".

Such a system of objects and morphisms is called a *category*:

Definition 2.1. A category C consists of the following data:

1. A collection of *objects*, denoted $\operatorname{ob}\mathsf{C}$. We usually just write $X \in \mathsf{C}$ for $X \in \operatorname{ob}\mathsf{C}$

2. For each pair of objects $X, Y \in \mathsf{C}$, a set of *morphisms* $\mathsf{C}(X, Y)$. When $f \in \mathsf{C}(X, Y)$, we write $f : X \to Y$.

3. A *composition*, a function $\circ : \mathsf{C}(X,Y) \times \mathsf{C}(Y,Z) \to \mathsf{C}(X,Z)$, which maps $f : X \to Y, g : Y \to Z$ to $g \circ f : X \to Z$.

4. For each object X, an *identity* $1_X : X \to X$.

This data must satisfy the following properties:

1. $1_Y \circ f = f \circ 1_X = f$ for all morphisms $f : X \to Y$. (This is called *unitality*)

2. $f \circ (g \circ h) = (f \circ g) \circ h$ for all triples of morphisms where this composition is defined (i.e $h : X \to Y, g : Y \to Z, f : Z \to W$). (This is called *associativity*).

Remark 2.2.

1. We often just write fg for the composite $f \circ g$ in a category. The associativity of composition makes it safe to omit the parentheses when composing three or more maps, so we simply write fgh for the composite $f \circ (g \circ f)$ and so on.

2. We also call morphisms *maps* (as we did just above), or *arrows*. I have tried to reserve the word *function* for actual functions between sets in the usual sense.

3. We may occasionally write $\mathrm{Hom}(A, B)$ for the set of maps $A \to B$ if the category is understood — Hom is short for *homomorphism*, and this notation comes from algebra, although we use it even in cases where the morphisms of the category are not really "homomorphisms" in any sense.

4. When $f : A \to B$, we call A the *domain* and B the *codomain*. A bit of a subtlety is that we require that each map in a category has a specific domain and codomain — for example, we distinguish between the identity map $1_{\mathbb{Z}} : \mathbb{Z} \to \mathbb{Z}$ and the inclusion map $i : \mathbb{Z} \hookrightarrow \mathbb{R}$ (in the category of sets, see below). These are *different morphisms*, even though they have the same value at every point.

Example 2.3. There is a category Set, where the objects are sets, the morphisms $X \to Y$ are simply the functions with domain X and codomain Y, and the composition is just the ordinary composition of functions.

Remark 2.4. In many cases, the composition and identities are more or less obvious once the objects and morphisms have been defined. Thus, we might have described Set simply as "the category of sets and functions", or even "the category of sets", for brevity. Below, we will frequently use this convention.

Example 2.5.

1. There is a category Fin of *finite* sets and functions.

2. There is a category Grp of groups and group homomorphisms.

3. There is a category $\mathsf{Vect}_{\mathbb{R}}$ of real vector spaces and linear maps.

4. For any partially ordered set $(S, \leq)$, there is a category with objects given by points of S, and a unique morphism $s \to s'$ if $s \leq s'$ (and otherwise no morphisms).

5. For any group $G = (G, \cdot)$, there is a category BG with one object, $*$, and $BG(*, *) = G$, with composition given by the group law (i.e $f \circ g = f \cdot g$). The identity 1_* is simply the neutral element of the group.

6. For any group G, there is a category $G\mathsf{Set}$ of (left) G-sets, sets with a left action of G. The morphisms are equivariant functions.

7. There is a category SmMan of smooth manifolds and smooth maps.

8. There is a category Meas of measurable spaces and measurable functions.

9. There is a category Prob of probability spaces (i.e measurable spaces equipped with a probability measure) and measure-preserving measurable maps.

Remark 2.6. In many of the above cases, there is a technical issue due to the fact that the collection of objects does not form a set (for instance, there is no set of all sets). There are various ways of resolving these issues — for a brief discussion, see [20, p. 6]. The distinction between a category whose objects and morphisms fit inside a set, called a *small* category, and one without this property, a *large* category, is not entirely irrelevant, but it should probably be ignored while first getting a feel for the subject.

Category-theorists often use so-called *commutative diagrams* to reason about categories. Here is an example:

$$\begin{array}{ccc} \mathbb{R} & \xrightarrow{+2} & \mathbb{R} \\ \downarrow{\scriptstyle \cdot 2} & & \downarrow{\scriptstyle \cdot 2} \\ \mathbb{R} & \xrightarrow{+4} & \mathbb{R} \end{array} \tag{1}$$

Such a diagram depicts objects of a category and morphisms between them — in this case, the category is simply Set. This square is said to *commute* if both ways of going around result in the same composite morphism, as is the case in the diagram above. (Somewhat confusingly, the term "commutative" is used both to distinguish this type of diagrams from other types — e.g. string diagrams, Section 2.2 — as well as the specific property that certain compositions are equal, which may or may not hold in a given diagram.)

We now list a few categories that will play an important role later

Definition 2.7. The category Met of metric spaces is defined as follows:

- The objects are metric spaces.

- The maps $f : (X, d_X) \to (Y, d_Y)$ are *short* or 1-*Lipschitz* functions, i.e those satisfying $d_Y(f(x), f(x')) \leq d_X(x, x')$ for all $x, x' \in X$.

Definition 2.8. An *error space* is a set S equipped with a function $e_S S \to [0, \infty]$, called the error. A map of error spaces $(S, e_S) \to (S', e_{S'})$ is a map $f : S \to S'$ so that $e_{S'}(f(s)) \leq e_S(s)$ for all $s \in S$ — in other words, a map can *decrease* the error of a point, but not increase it. We usually omit the error function and simply write "S is an error space". We also write simply e for the error function if there is no chance of confusion. This defines the category Err of error spaces.

Definition 2.9. We let $*_e \in \mathsf{Err}$ denote the set $\{*\}$ with $e(*) = e$. We let $* = *_0$.

Note that there is a unique map $S \to *_e$ if and only if each $s \in S$ has $e(s) \geq e$, otherwise there are no such maps. In particular there are maps $*_e \to *_{e'}$ if $e > e'$.

Definition 2.10. A map $f : X \to Y$ is an *isomorphism* if there exists $g : Y \to X$ so that $fg = 1_Y, gf = 1_X$. If such a g exists, it is necessarily unique and we denote it f^{-1}. It is called an *inverse* of f. If there exists an isomorphism $X \to Y$, we write $X \cong Y$ and call them *isomorphic*.

Example 2.11.

1. An isomorphism of sets is a bijection.
2. In BG, all the maps are isomorphisms (with inverses given by inverses in the group).
3. An isomorphism of groups is precisely what's normally called a group isomorphism — a bijective group homomorphism (the inverse is automatically a homomorphism).
4. An isomorphism in Met is *not* the same thing as a bijective short map — instead, it is a bijective *isometry*, i.e a bijection where $d(f(x), f(y)) = d(x, y)$.
5. An isomorphism in Err is a bijective error-preserving map f, i.e one with $e(f(x)) = e(x)$ for all x.

One of the guiding principles in category theory is that objects which are isomorphic should be interchangable — a property or construction which does not respect isomorphisms is sometimes said to be *evil*.

Example 2.12. Let Prob^* be the category where objects are probability spaces, and maps are equivalence classes of measure-preserving maps under the relation of almost-certain equivalence[2]. Then a map f is an isomorphism if it has a

[2]For this to make sense, we actually have to verify that a.e. equality is preserved by composition. To see this, note that $\{x \mid fg(x) \neq f'g'(x)\} \subseteq \{x \mid g(x) \neq g'(x)\} \cup g^{-1}(\{y \mid f(y) \neq f'(y)\})$. If two compositions differ at x, either the first functions differ there, or the latter functions differ at $g(x)$. Then we just note that these two sets are null and g is by assumption measure-preserving.

measurable "almost certain inverse", i.e so that $ff^{-1}(y) = y$ for almost all y, and $f^{-1}f(x) = x$ for almost all x.

If $(\mathcal{X}, P)$ is a probability space, and $E \subseteq \mathcal{X}$ is measurable with probability 1, then the inclusion $(E, P|_E) \hookrightarrow (\mathcal{X}, P)$ is an isomorphism in this category — reflecting the intuition that sets of probability 0 "don't matter".

Example 2.13. Let Met be as above, and let Met^c denote the category of metric spaces and *continuous* maps. There is an obvious inclusion functor $\mathsf{Met} \to \mathsf{Met}^c$ (since any Lipschitz map is continuous). This of course implies that isometrically metric spaces are isomorphic in Met^c. But the converse is not true — metric spaces which are isomorphic in Met^c are merely homeomorphic. For example, the open interval $(0, 1)$ and the real line $\mathbb{R}$ (both in the usual metric) are known to be homeomorphic, by the map $x \mapsto \tan(x\pi - \pi/2)$, but they are obviously not isometric.

This example shows that the requirement that "isomorphic objects are interchangable" is largely a matter of perspective — it depends on what notion of morphism you're considering! This also motivates the collection of 1-Lipschitz functions as the "correct" notion of transformation between metric spaces — it ensures that isomorphic metric spaces really have "the same" metric.

(We could also have chosen, for example the collection of all Lipschitz morphisms. This would lead to metric spaces where isomorphic metric spaces have "equivalent" metrics, in the sense that

$$cd_X(x, x') \leq d_Y(f(x), f(x')) \leq Cd_X(x, x'),$$

for all $x, x' \in X$ and for some positive constants c, C. But this is not what we're interested in here.)

Example 2.14. Given a category with one object (call it $*$), and where each map is an isomorphism, the set $\mathrm{Hom}(*, *)$ acquires the structure of a group, with multiplication given by composition. This is an inverse to the construction BG discussed above.

Definition 2.15. Let C be a category. Then the *opposite* or *dual* category, denoted C^{op}, is defined by reversing the arrows of C. Formally, $\mathrm{ob}\,\mathsf{C}^{\mathrm{op}} = \mathrm{ob}\,\mathsf{C}$, but $\mathsf{C}^{\mathrm{op}}(A, B) = \mathsf{C}(B, A)$. Composition is the same as composition in C, but with the arguments reversed.

2.1 Functors

Since the whole idea of category theory is that often the transformations between objects are as important as the objects themselves, it's natural to apply this idea to categories. The structure-preserving transformation between categories are called *functors*.

Definition 2.16. Let C, D be categories. A *functor* $F : \mathsf{C} \to \mathsf{D}$ consists of

1. A function $F : \mathrm{ob}\,\mathsf{C} \to \mathrm{ob}\,\mathsf{D}$

2. For each pair of objects $X, Y \in \mathsf{C}$, a function $\mathsf{C}(X,Y) \to \mathsf{D}(FX, FY)$, which we also denote by F.

3. So that $F(1_X) = 1_{FX}$ and $F(fg) = F(f)F(g)$.

In a sense, a functor is a lot like a group homomorphism — it must preserve the composition law and the identities of the category. Another way of thinking about functors is that they are precisely those maps between categories which preseve commutative diagrams like Eq. (1)

Example 2.17. There is a functor $U : \mathsf{Vect}_{\mathbb{R}} \to \mathsf{Set}$ which assigns a vector space its underlying set, and a linear map its underlying function. Functors of this form, which "forget" the structure of some object, are often called forgetful functors. There are for example also forgetful functors $\mathsf{Met} \to \mathsf{Set}$, $\mathsf{Grp} \to \mathsf{Set}$.

Example 2.18. Let G be a group. A functor $A : BG \to Set$ consists of this data:

1. A set $A(*)$

2. For each group element g, a function $A(g) : A(*) \to A(*)$

3. Such that $A(1) = 1_{A(*)}$, and $A(gg') = A(g)A(g')$.

If you stare at this for a bit, it's clear that this is exactly the data of a set with a (left) action of G. Similarly, a functor $BG \to \mathsf{Vect}_{\mathbb{R}}$ is precisely the data of a vector space with a linear action of the group G (i.e a representation of G). In general, it makes sense to think of a functor $BG \to \mathsf{C}$ as an object of C equipped with an "action" of G.

Example 2.19. Given any object $X \in \mathsf{C}$, the construction $Y \mapsto \mathsf{C}(X,Y)$ defines a functor $\mathsf{C} \to \mathsf{Set}$. Functors of this form are called *representable*. We also denote this functor $\mathsf{C}(X, -)$

Remark 2.20. We will in general use the notation "$-$" for an "anonymous" variable. For example, we let $f(x,-)$ denote the function $y \mapsto f(x,y)$, and so on.

Example 2.21. Let $\Omega = (\Omega, P) \in \mathsf{Prob}$ be a background probability space.

1. $\mathsf{Prob}(\Omega, -)$ is a functor which sends a probability space to the set of random variables *with that distribution*. For instance, $\mathsf{Prob}(\Omega, (\mathbb{R}, \mathcal{N}(0,1)))$ is the set of standard Gaussian random variables. (If we use notation in a slightly unorthodox way and let $\mathcal{N}(0,1)$ denote the measure on $\mathbb{R}$ corresponding to the standard Gaussian).

2. There is also a functor $RV_\Omega : \mathsf{Meas} \to \mathsf{Set}$ which takes a measurable space to the set of random variables valued in that space, quotiented by the equivalence relation of P-almost certain equality. A measurable function $f : \mathcal{X} \to \mathcal{Y}$ is taken to the operation $X \mapsto f(X)$ on random variables.

Example 2.22. Let Prob^* again be the category of probability spaces and a.e. equivalence classes of measure-preserving maps. A functor $B\mathbb{Z} \to \mathsf{Prob}^*$ consists of

1. A probability space (Ω, P)
2. Equipped with measure-preserving operations $T^n : \Omega \to \Omega$ for each $n \in \mathbb{Z}$
3. So that $T^0 = 1_\Omega$ almost-surely and $T^n T^m = T^{n+m}$ almost surely.

It's clear that every T^n is determined uniquely (up to almost sure equivalence) by the choice of T^1, and also that T^1 can be any almost-surely invertible and measure-preserving transformation (the inverse is given by T^{-1}). Hence this is the same thing as a probability space equipped with an almost-surely invertible, measure-preserving transformation. This is essentially a measurable discrete-time dynamical system, where P is an equilibrium distribution.

Example 2.23. A functor $BG \to BH$, for two groups G, H, is the same thing as a group homomorphism $G \to H$ — there is only one possible function $\operatorname{ob} BG \to \operatorname{ob} BH$, and the condition on the map $BG(*,*) \to BH(*,*)$ is precisely that it defines a group homomorphism.

Definition 2.24. There are two important functors $\mathsf{Err} \to \mathsf{Set}$. The first, which we will denote ex, takes an error space S to the set of elements with error 0. In other words, $\operatorname{ex}(S) = \{s \in S \mid e(s) = 0\}$. Given a morphism $f : S \to S'$, $\operatorname{ex}(f)$ is simply the restriction of f to $\operatorname{ex}(S)$ (which, by assumption, has image contained in $\operatorname{ex}(S')$).

The second, which we will denote $|-|$, simply forgets the error, so that $|(S, e)| = S$ and $|f : S \to S'| = f$.

Remark 2.25. In fact, $\operatorname{ex} \cong \mathsf{Err}(*, -)$, and $|-| \cong \mathsf{Err}(*_\infty, -)$ (as in definition Definition 2.9)

Remark 2.26. Any functor preserves isomorphisms — specifically, $F(f^{-1})$ is an inverse of $F(f)$. In this sense a construction which is functorial is "not evil".

Definition 2.27. A functor is *full* if each map $\mathsf{C}(X, Y) \to \mathsf{D}(FX, FY)$ is surjective. It is *faithful* if each map $\mathsf{C}(X, Y) \to \mathsf{D}(FX, FY)$ is injective. A functor which is full and faithful is called *fully faithful*.

Example 2.28. The forgetful functor $\mathsf{Grp} \to \mathsf{Set}$ is obviously faithful, since group homomorphisms are equal if and only if they are equal as functions — a group homomorphism does not involve any extra data than a function. It is not full, since there are obviously functions between the underlying sets that are not group homomorphisms (outside of trivial cases like maps into $\{*\}$).

Example 2.29. The functor $|-| : \mathsf{Err} \to \mathsf{Set}$ is faithful, but not full. The functor $\operatorname{ex} : \mathsf{Err} \to \mathsf{Set}$ is not faithful or full. It is not faithful because two maps $f, g : S \to T$ may differ, but agree on those points with error zero. It is not full because, if $S = *_1, T = *_2$, there are no maps $S \to T$, but there is one map $\operatorname{ex}(S) = \emptyset \to \operatorname{ex}(T) = \emptyset$.

Remark 2.30. Given a fully faithful functor $F : \mathsf{C} \to \mathsf{D}$, we may treat C as a "subcategory" of D, consisting of a certain subset of the objects, but having the same morphisms and composition. Interestingly, we can do this even if F is not injective on objects, because if $F(X) = F(Y)$, then the preimage in $\mathsf{C}(X, Y)$ of $1_{F(X)} : F(X) \to F(Y)$ is an isomorphism, so $X \cong Y$. (In other words, a fully faithful functor is necessarily injective on *isomorphism classes* of objects.)

There is also a notion of transformation between functors:

Definition 2.31. Let $F, G : \mathsf{C} \to \mathsf{D}$ be functors. A *natural transformation* $\alpha : F \to G$ is a collection of maps $\alpha_X : FX \to GX$ for each $X \in \mathsf{C}$, such that for each arrow $f : X \to Y$ in C, the diagram

$$\begin{array}{ccc} FX & \xrightarrow{F(f)} & FY \\ \downarrow{\scriptstyle \alpha_X} & & \downarrow{\scriptstyle \alpha_Y} \\ GX & \xrightarrow{G(f)} & GX \end{array}$$

commutes.

The category of functors $\mathsf{C} \to \mathsf{D}$ and natural transformations is denoted $[\mathsf{C}, \mathsf{D}]$.

Example 2.32. Let $U : \mathsf{Vect}_{\mathbb{R}} \to \mathsf{Set}$ be as in Example 2.17. For any $\alpha \in \mathbb{R}$, the map $v \mapsto \alpha v$ is a natural transformation $U \to U$. To see this, note that by definition, this means that for any linear map $f : V \to W$, we have $\alpha f(v) = f(\alpha v)$, which is part of the definition of linearity.

Example 2.33. For any group G, the category $[BG, \mathsf{Set}]$ is isomorphic to the category $G\mathsf{Set}$ of G-sets, in the sense that there is a functor $[BG, \mathsf{Set}] \to G\mathsf{Set}$ which has an inverse. This functor is essentially given by the construction of example Example 2.18

Example 2.34. Fix a background probability space (Ω, P). Then there is a functor $RV^1 : \mathsf{Vect}_{\mathbb{R}} \to \mathsf{Set}$ which takes a real vector space V to the set of random variables with finite expectation valued in V (using the Borel σ-algebra), and takes a linear transformation f to the map $X \mapsto f(X)$. There is also a forgetful functor $U : \mathsf{Vect}_{\mathbb{R}} \to \mathsf{Set}$ from Example 2.17. Then the expectation $\mathbb{E} : RV^1(V) \to U(V)$ is a natural transformation. This just reflects the well-known fact that linear transformations preserve expectation — $\mathbb{E}[f(X)] = f(\mathbb{E}X)$ when f is linear.

We have the following very important result about representable functors:

Proposition 2.35. The assignment $X \to \mathsf{C}(-, X)$ defines a functor $y : \mathsf{C} \to [\mathsf{C}^{\mathrm{op}}, \mathsf{Set}]$, which is moreover fully faithful.

This is a consequence of the celebrated *Yoneda lemma*, and the functor y is called the *Yoneda embedding*. The utility of this result is that the category $[\mathsf{C}^{\mathrm{op}}, \mathsf{Set}]$ is usually much more well-behaved than the base category C, and the Yoneda embedding allows us to work in the larger category instead.

Lemma 2.36 (Yoneda)**.** Let C be a category, $F : \mathsf{C}^{\mathrm{op}} \to \mathsf{Set}$ be any functor, and let $X \in \mathsf{C}$ be an object. Then $F(X) \cong [\mathsf{C}^{\mathrm{op}}, \mathsf{Set}](\mathsf{C}(-, X), F)$.

To prove the proposition, simply apply the lemma in the case $F = \mathsf{C}(-, Y)$. See e.g. [20, Thm. 2.2.4] for a proof.

Functors $\mathsf{C}^{\mathrm{op}} \to \mathsf{Set}$ are called *presheaves on* C

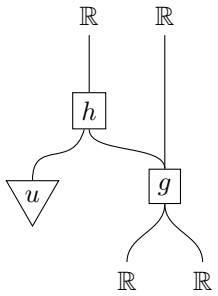
Remark 2.37. Essentially, we can see the Yoneda lemma as saying

1. A presheaf $F : \mathsf{C}^{\mathrm{op}} \to \mathsf{Set}$ is a sort of generalized object of C (in the sense that the presheaves contain the objects as a full subcategory).

2. To map from an object of C into a generalized object, you simply apply the functor that is the generalized object.

2.2 Monoidal categories and string diagrams

We will make use on several occasions of the graphical notation called *string diagrams*. Here is an example of a string diagram.



(2)

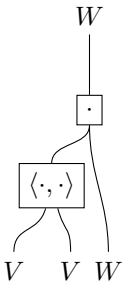
The meaning of this diagram is fairly clear — it shows how to build a composite function by "wiring together" functions. g is some function $\mathbb{R}^2 \to \mathbb{R}^2$, for example $g(x, y) = (x + y, (x - y)^2)$, and h is some function $\mathbb{R}^2 \to \mathbb{R}$, say $h(a, b) = a \cdot b$, while u is a function $\{*\} \to \mathbb{R}$ — which is the same thing as an element of $\mathbb{R}$, say 5. Note that this diagram is oriented "bottom to top" — the domain of a map is at the bottom, and the codomain at the top. We use this convention throughout.

It does not, in general, make sense to wire together maps in a general category like this. This is because a category does not come with a built-in notion of "multivariate morphism". The simplest way to add this structure to a category is C to add a "product", $\otimes : \mathsf{C} \times \mathsf{C} \to \mathsf{C}$. Then we can think of a "morphism in two variables" as a map $A \otimes B \to C$, and so on[3].

To intepret things like u, we must further have a notion of element, or zero-variable operation. This is provided by specifying a *unit object*, usually denoted $I \in \mathsf{C}$.

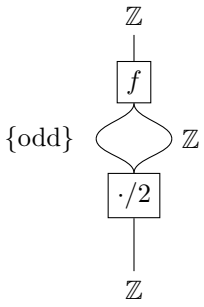
[3]There are other approaches to this, where the notion of morphism with multiple inputs is taken as primitive. These are usually called something like multi- or polycategories, see eg [11]

This data, $\otimes$ and I, must satisfy a number of conditions which ensure that the interpretation of diagrams like Equation (2) is well-defined — $\otimes$ must be a functor, be associative, I must be a unit up to natural isomorphism, and satisfy the so-called *coherence axioms*. See [15, Chap. 7] for a technical treatment. A category equipped with data like this is called a *symmetric monoidal category*.



(3)

This diagram depicts a map in vector spaces, which takes the inner product of two vectors and scales a third vector by it. This is a *multilinear* map, which means it's a map $V \otimes V \otimes W \to W$[4], and not a map $V \times V \times W \to W$. In other words, this is a diagram in the symmetric monoidal category $(\mathsf{Vect}_{\mathbb{R}}, \otimes, \mathbb{R})$.



(4)

This diagram depicts a function of sets, but in an unusual way. The function $\cdot/2$ halves an integer if it is even, but "returns an error" if it's odd. Hence it's a function $\mathbb{Z} \to \mathbb{Z} \sqcup \{\text{odd}\}$, where $\sqcup$ denotes the "disjoint union" of sets[5]. Then the function $f : \mathbb{Z} \sqcup \{\text{odd}\} \to \mathbb{Z}$ handles this error in some way (it doesn't really matter — let's say $f(\text{odd}) = 0$). This diagram can be interpreted in a symmetric monoidal category $(\mathsf{Set}, \sqcup, \emptyset)$. The two paths through the diagram show the two possible ways the function can be evaluated, depending on the input.

This example shows that we can often put multiple distinct monoidal structures on a category — this choice reflects a decision about what it should mean for a function to be "multivariate". In $(\mathsf{Set}, \times, \{*\})$, it means that it takes in a pair of values. In $(\mathsf{Set}, \sqcup, \emptyset)$, it means that it takes in a value from one of two possible sets. In both $(\mathsf{Vect}_{\mathbb{R}}, \otimes, \mathbb{R})$ and $(\mathsf{Vect}_{\mathbb{R}}, \times, 0)$, a multivariate morphism can be said to accept two variables — in the former, it must be *bilinear*, while in the latter, it must be simply linear.

[4]Here, the $\otimes$ symbol denotes the usual tensor product of vector spaces

[5]The normal union $\cup$ is an evil operation — $A \cong B$ does not imply $A \cup C \cong B \cup C$

In the presence of a symmetric monoidal structure, we can also make sense of "function spaces".

Definition 2.38. Let $(\mathsf{C}, \otimes, I)$ be a symmetric monoidal category. Let $X, Y \in \mathsf{C}$ be objects. We say that $[X, Y]$ is a *hom object*[6] for X and Y if there is a natural isomorphism

$$\mathsf{C}(- \otimes X, Y) \cong \mathsf{C}(-, [X, Y])$$

of functors $\mathsf{C}^{\mathrm{op}} \to \mathsf{Set}$.

It turns out that hom objects, if they exist, are unique up to isomorphism, so that we can reasonably speak of *the* hom object $[X, Y]$.

Definition 2.39. A symmetric monoidal category is *closed* if all hom objects exist.

Remark 2.40. The coincidence between our notation for hom objects and functor categories is no coincidence — there's a symmetric monoidal "category of categories", Cat, so that the functor category is the hom object.

Example 2.41. The category $(\mathsf{Set}, \times, *)$ is closed, and $[X, Y]$ is the set of functions $X \to Y$.

Definition 2.42. The tensor product $X \otimes Y$ of two metric spaces is their Cartesian product equipped with the ℓ^1-metric: $d((x, y), (x', y')) = d(x, x') + d(y, y')$. This equips the category of metric spaces with a symmetric monoidal structure, with unit the singleton metric space.

See [6] for a proof of this.

Definition 2.43. The *tensor product* of two error spaces is defined by $S \otimes S' = S \times S'$ and $d_{S \otimes S'}(s, s') = d_S(s) + d_{S'}(s')$. The *function space*, written $[S, S']$ is the set of all functions $f : S \to S'$, with $e(f) = \max(0, \sup_s e(f(s)) - e(s))$. This defines a closed monoidal category Err. The monoidal unit is $*_0$.

Proposition 2.44. The above definitions equip Err with the structure of a closed monoidal category.

⊙⊙ *Proof.* We first verify that $(\mathsf{Err}, \otimes, *_0)$ is a symmetric monoidal category. This consists of the following claims:

1. $\otimes$ defines a functor $\mathsf{Err} \times \mathsf{Err} \to \mathsf{Err}$
2. There exists natural isomorphisms $\sigma_{X,Y} : X \otimes Y \cong Y \otimes X$, $\eta_X : X \otimes \{*\} \cong X$, and $\alpha_{X,Y,Z} : X \otimes (Y \otimes Z) \cong (X \otimes Y) \otimes Z$
3. These natural isomorphisms satisfy the MacLane coherence axioms.

[6] for "homomorphism"

This part of the proof is entirely routine.

Given maps $f : X \to X'$ and $g : Y \to Y'$, the map $(f \otimes g) : X \otimes Y \to X' \otimes Y'$ is simply defined by $f \otimes g(x, y) = (f(x), g(y))$. In other words, on morphisms the functor $\otimes$ simply acts as the functor $\times : \mathsf{Set} \times \mathsf{Set} \to \mathsf{Set}$. The natural isomorphisms are similarly lifted from the symmetric monoidal structure $(\mathsf{Set}, \times, \{*\})$:

$$\eta_X(x, *) = x$$

$$\sigma_{X,Y}(x, y) = (y, x)$$

$$\alpha_{X,Y,Z}(x, (y, z)) = ((x, y), z)$$

A routine calculation suffices to verify that these formulae do in fact define natural isomorphisms of error spaces, and to verify the coherence axioms.

To prove that the function space as defined above defines a closed structure, it suffices to verify the adjunction — that is, to check that

$$\mathsf{Err}(X, [Y, Z]) \cong \mathsf{Err}(X \otimes Y, Z).$$

If we drop the requirement of being a map of error spaces on both sides, we clearly have a bijection between functions $X \to [Y, Z]$ and $X \otimes Y \to Z$, where the map $f(x, y)$ corresponds to the map $x \mapsto f(x, -)$. We must verify that this correspondence preserves the property of being error nonincreasing, in each direction. Suppose $f : X \otimes Y \to Z$ is error nonincreasing. This means that $e(f(x, y)) \leq e(x) + e(y)$. Now take the map $f(x, -) \in [Y, Z]$. By the inequality above, $e(f(x, y)) - e(y) \leq e(x)$ for all y, so that map has error at most $e(x)$ — meaning that the map $X \to [Y, Z]$ is error nonincreasing.

For the opposite direction, suppose $f : X \to [Y, Z]$ is error nonincreasing. Then $e(f(x)(y)) - e(y) \leq e(x)$ for all x, y, but that means $e(f(x)(y)) \leq e(y) + e(x)$, as desired. □

Remark 2.45. Observe that the error of a function $f \in [X, Y]$ can be infinite, even if all elements of both X and Y have finite error. This is one big technical advantage of allowing infinite errors — the category of error spaces would not be closed if we did not. We will also see later (Lemma 2.65) that this is necessary in order for Err to have all limits and colimits.

2.3 Enriched categories

An *enriched* category is like a category, but where the morphism sets $\mathsf{C}(X, Y)$ carry some extra structure. For instance, the set of linear maps between two vector spaces, $\mathsf{Vect}_{\mathbb{R}}(V, W)$, can itself be endowed with the structure of a (real) vector space (by pointwise operations). Then we say that $\mathsf{Vect}_{\mathbb{R}}$ is *enriched in* $\mathsf{Vect}_{\mathbb{R}}$ (it's quite common for categories to be enriched in themselves). In general, a category C is enriched in another category V if each morphism set $\mathsf{C}(X, Y)$ carries the structure of an object of V, and the composition respects this structure (for instance, composition of linear maps is bilinear). This definition is somewhat vague — a precise definition is a bit subtle, and requires a more

precise treatment of monoidal categories. We will spell out what an enriched category is in each of the cases that we need it. See [8] for a comprehensive treatment of enriched categories.

Example 2.46.

1. A category "enriched in Set" is just an ordinary category.
2. A category enriched in $(\mathsf{Vect}_{\mathbb{R}}, \otimes, \mathbb{R})$ has vector spaces of maps, with composition bilinear (the identity is a linear map $1_A : \mathbb{R} \to \mathsf{C}(A, A)$, which can be identified with the value $1_A(1)$)
3. A category enriched in Fin is a category where every morphism set is finite.
4. A category enriched in Meas is a category where every morphism set carries a σ-algebra, so that composition is a measurable operation.

Remark 2.47. A Met-category consists of the following data:

1. An ordinary category C,
2. with a metric on each mapping set $\mathsf{C}(X, Y)$,
3. such that the compositions $\mathsf{C}(X, Y) \times \mathsf{C}(Y, Z) \to \mathsf{C}(X, Z)$ are short in each variable, i.e $d(fg, fg') \leq d(g, g')$ and $d(fg, f'g) \leq d(f, f')$.

The idea here is that we can control the distance between composite morphisms in terms of the distances between their components.

Readers should feel free to regard this as the definition of a Met-enriched category.

The following lemma will be useful later — it tells us how to compose "almost commutative" squares.

Lemma 2.48. Let C be a Met-category and consider a diagram of this form (not necessarily commutative):

$$\begin{array}{ccc} A & \xrightarrow{a} & A' \\ \downarrow{\scriptstyle f} & & \downarrow{\scriptstyle f'} \\ B & \xrightarrow{b} & B' \\ \downarrow{\scriptstyle g} & & \downarrow{\scriptstyle g'} \\ C & \xrightarrow{c} & C \end{array}$$

Suppose $d(f'a, bf) = \epsilon$ and $d(g'b, cg) = \delta$. Then $d(g'f'a, cgf) \leq \epsilon + \delta$.

Proof. By Remark 2.47, we have

$$d(g'f'a, g'bf) \leq \epsilon$$

$$d(cgf, g'bf) \leq \delta$$

Now we apply symmetry and the triangle inequality for the desired conclusion. □

Remark 2.49. An Err-category consists of the following data.

1. An ordinary category,
2. with an error $e(f)$ for each morphism $f : X \to Y$,
3. such that $e(fg) \leq e(f) + e(g)$.

Similar to the case of Met-categories, the idea of an Err-category is that we can control the error of a composite morphism in terms of the component morphisms.

An Err-functor is a functor which does not increase the error of morphisms, i.e $e(F(f)) \leq e(f)$ for any morphism f.

Given two Err-categories C, D, there is an Err-category of Err-functors $[\mathsf{C}, \mathsf{D}]$. The objects are Err-functors $\mathsf{C} \to \mathsf{D}$. The morphisms are natural transformations. The error of a natural transformation α is $e(\alpha) = \sup_{X \in \mathsf{C}} e(\alpha_X)$.

Again, readers should feel free to regard the above as definitions on a first reading.

Definition 2.50. Given an Err-category C, we obtain a subcategory C_{ex} of *exact morphisms*. Its objects are the same as C, and its morphisms are those morphisms in C with error zero.

⊛⊛ **Remark 2.51.** C_{ex} is just the "underlying category" of C, in the sense of enriched category theory, usually denoted C_0 (see eg [8, section 1.3]). The construction $(-)_{\text{ex}}$ has a left adjoint, which equips an ordinary category with the Err-enrichment where $e(f) = 0$ for all f. This has a further left adjoint, which we might call the *inexact category*, which simply forgets the errors of all the morphisms — we could reasonably denote this functor $|(-)|$.

Definition 2.52. Define the Err-category $\underline{\mathsf{Err}}$ as follows:

1. The objects are the error spaces.
2. A map $S \to S'$ is any function $S \to S'$.
3. Composition is just ordinary function composition.
4. The error of a map $f : S \to S'$ is $\max(0, \sup_x e(f(x)) - e(x))$

In other words, $\underline{\mathsf{Err}}(X, Y) = [X, Y]$.

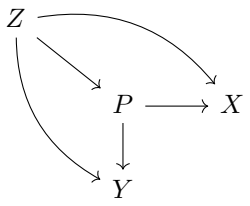
Observe that $\underline{\mathsf{Err}}_{\text{ex}} = \mathsf{Err}$. For this reason we think of $\underline{\mathsf{Err}}$ as the "canonical" enrichment of Err in itself. (It may perhaps seem a bit strange that the morphism sets of $\underline{\mathsf{Err}}$ contain a large number of functions that aren't in Err, but from the point of view of enriched categories, this is actually natural). Since there is usually no chance of harmful confusion, we will often abuse notation and use the symbol Err for the enriched category as well, omitting the underline.

Remark 2.53. All the technology we have developed so far also works for enriched categories. That is, there is a notion of enriched functor, which is the same as an ordinary functor except the maps $\mathsf{C}(X,Y) \to \mathsf{D}(FX,FY)$ are maps in the enriching category. There is an enriched functor category (which we simply denote $[\mathsf{C},\mathsf{D}]$), and an enriched Yoneda lemma.

2.4 Limits

We now discuss the construction known as a *limit*. A limit is a way of defining a new object in a category in terms of "smaller" objects. The simplest example is the so-called *product*

Definition 2.54. Let $X, Y \in \mathsf{C}$ be objects of a category. A *product* of X and Y consists of a third object P equipped with two maps $\pi_X : P \to X, \pi_Y : P \to Y$, with the following "universal" property: given a fourth object Z and maps $f : Z \to X, g : Z \to Y$, there is a unique map $(f,g) : Z \to P$ such that the following diagram commutes



Proposition 2.55. Given two products (P, π_X, π_Y) and (P', π'_X, π'_Y), there is a unique isomorphism $i : P \to P'$ so that $\pi_X = i\pi'_X, \pi_Y = i\pi'_Y$.

Proof. Apply the universal property for P' to the maps π_X, π_Y to construct i. Apply the universal property for P to the maps π'_X, π'_Y to construct $j : P' \to P$. The uniqueness in the statement implies that $ij = 1_{P'}$ and $ji = 1_P$. □

Since products are unique (up to isomorphism), we usually talk about "the" product of two objects, and denote it $X \times Y$.

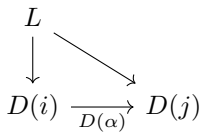
Example 2.56. In Set, the product of two sets X, Y always exists and equals the usual Cartesian product $X \times Y$, with $\pi_X(x,y) = x, \pi_Y(x,y) = y$ the ordinary projections. Given $f : Z \to X, g : Z \to Y$, the induced unique map $Z \to X \times Y$ is simply $z \mapsto (f(z), g(z))$.

The idea behind the definition of product is that it generalizes this concept to other categories. The two projections in the product pick out the two "coordinates", and the universal property ensures that the product is entirely given by the two coordinates, in some sense.

The idea of a "limit" is to generalize this to more complicated systems of objects than just a pair. In general, the input to a limit is a "diagram", which in fact just means a functor $D : \mathsf{I} \to \mathsf{C}$ from some other "index" category I. The limit of such a thing is the "universal" object with a map $L \to D(i)$ for each

$i \in \mathsf{I}$ which is compatible with all the maps $D(f) : D(i) \to D(j)$. Here is how to make this precise:

Definition 2.57. Given a functor $D : \mathsf{I} \to \mathsf{C}$, a *limit* consists of an object L equipped with maps $\pi_i : L \to D(i)$ for each i, so that all the triangles



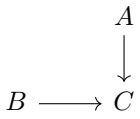
commute, for each $\alpha : i \to j$ in I. Furthermore L must be universal, in the sense that given another such compatible collection $(Z, f_i : Z \to D(i))$, there is a unique map $\hat{f} : Z \to L$ such that $f_i = \pi_i \hat{f}$.

One can prove, in a way completely analogous to the previous proof, that limits are unique up to unique isomorphism if they exist. For that reason we often speak of "the" limit, denoted $\lim_{i \in \mathsf{I}} D(i)$, or just $\lim D$. We recover the case of products by choosing for I the category with two objects and only identity morphisms.

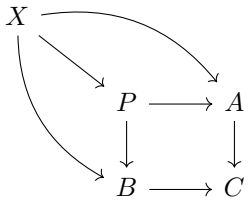
Example 2.58. If I is the *empty* category, with no objects, there is a unique functor $\mathsf{I} \to \mathsf{C}$ (for any category C). An object $T \in \mathsf{C}$ is a limit of this diagram if and only if, for each other object $X \in \mathsf{C}$, there is a unique map $X \to T$. Such an object T is called *terminal*.

For example, a terminal object of Set is precisely a singleton set.

Example 2.59. If I is the category with three objects, a, b, c, and two nonidentity morphisms, $a \to b$ and $c \to b$, a diagram $D : \mathsf{I} \to \mathsf{C}$ looks like this



(where $A = D(a)$ and so on). A limit of this diagram is an object P equipped with maps $P \to A, P \to B$, such that



1. The inner square above commutes
2. Given any X equipped with maps $X \to A, B$ so that the outer cell commutes, there exists a unique map $X \to P$ making the whole diagram commutative.

Such a limit P is called a *pullback*, and typically denoted $A \times_C B$

Example 2.60. If $\mathsf{I} = BG$ for some group G, recall that a functor $A : BG \to \mathsf{Set}$ is the same thing as a set $A(*)$ with an action of G. The limit $\lim A$ is the set of *fixpoints* of the action, i.e those $a \in A(*)$ so that $g.a = a$ for all $g \in G$.

In the enriched case, things get significantly more complicated. In general, one has to deal with something called "weighted limits". However, we can simplify things significantly in the case of Err-categories, which is the only case of interest.

Definition 2.61. Let C and I be Err-categories. Let $D : \mathsf{I} \to \mathsf{C}$ be a functor, and let $W : \mathsf{I} \to \mathsf{Err}$ be a functor with $W(i)$ a singleton for each i. The data of such a functor amounts to specifying the error $e_{W(i)}(*)$, which we denote e_i (this is subject to certain conditions, coming from the functoriality). Then a *W-weighted limit of D* is a universal object $L \in \mathsf{C}$, equipped with maps $\pi_i : L \to D(i)$, with $e(\pi_i) \leq e_i$. Here "universal" is taken to mean that, given another such object (Z, f_i), there is a unique *exact* morphism $Z \to L$ making the diagrams commute.

The concept of weighted limit makes sense for a general functor $W : \mathsf{I} \to \mathsf{Err}$, with a significantly more complicated definition. However we will show in the following section that this generality is redundant in our situation. The purpose of the "weights" $W(i)$, in this situation, is just to specify the error that should be allowed on the maps π_i — this is a necessary part of pinning down the universal object up to exact isomorphism. In concrete examples, we'll generally only be concerned with limits that are "exact", i.e those where $W(i) = *_0$ for all i and where all the morphisms of I are exact. In this case a weighted limit is the same thing as a limit in the category C_{ex}.

One class of limits is particularly important to us: the so-called *cofiltered* limits. Being "cofiltered" is actually a property of the index category:

Definition 2.62. A category I is *cofiltered* if

1. There is at least one object of I
2. Given any two objects $x, y \in \mathsf{I}$, there exists an object z and maps $z \to x, z \to y$.
3. Given any two parallel arrows $f, g : x \to y$, there exists an arrow $h : z \to x$ so that $fh = gh$.

A "cofiltered limit" is just a limit of a functor from a cofiltered category. Cofiltered limits have certain theoretical properties that make them convenient. A simple example of a cofiltered category is the category $(\mathbb{Z}, \leq)$. A functor from this category looks like a sequence of objects

$$\cdots \to X_{-1} \to X_0 \to X_1 \to \cdots$$

Taking the limit of this diagram is in some sense like taking the limit of X_n as $n \to -\infty$[7].

Example 2.63. Given a nonincreasing $\mathbb{Z}$-indexed sequence $x_n \in [0, \infty]$, we obtain a functor $*_{x_{(-)}} : (\mathbb{Z}, \leq) \to \mathsf{Err}$. The limit of this sequence is

$$*_{\lim_{n\to\infty} x_n} = *_{\sup_n x_n}.$$

In the Err-enriched case, we need an additional consideration of the weights. This turns out to be fairly simple: the weights $W(i)$ need to contain arbitrarily small (but not necessarily zero) errors. This is a slightly strange condition, which I don't know exactly how to think about. In the cases of interest, all the errors are zero, so this is not an issue. The technical term for this type of W is *flat* — we will later encounter this term in the definition of *profinite models*. There are deep reasons for this "coincidence", but they're probably beyond this level of explanation. We record the upshot of this discussion precisely for convenience:

Remark 2.64. A functor $W : \mathsf{I} \to \mathsf{Err}$ where $W(i)$ is a singleton for each i is flat if and only if I is cofiltered in the usual sense, and the infimum of errors $\inf_i e_{W(i)}(*) = 0$. A cofiltered limit in an Err-category is a limit weighted by a flat functor.

2.5 Limits in Err-categories

We now study several properties of weighted limits in Err-enriched categories. This section is written assuming significant familiarity with the theory of enriched categories, weighted limits and so forth. The standard reference for that topic is [8]. The goal is to show

1. That Err has sufficiently good properties for the theory of enriched flat functors, developed in [1], to apply. This is contained in Lemma 2.65 and the subsequent definitions.

2. That finite weighted limits can, as normal, be built out of pullbacks and a terminal object. This is Lemma 2.74.

3. That enriched left Kan extensions interact nicely with finite limits. This is Proposition 2.76.

These are the results necessary to understand the *profinite models* of Section 5.

Lemma 2.65. Err has all limits and colimits. Moreover, the underlying set functor $\mathsf{Err} \to \mathsf{Set}$ preserves both limits and colimits.

Proof. It suffices to prove this for products, coproducts, equalizers and coequalizers. Let two parallel arrows $A \rightrightarrows B$ in Err be given. Let $i : S \to A$ be the equalizer of this diagram in Set. Equip S with an error structure by setting

[7] This is where the terminology "limit" comes from

$e(s) = e(i(s))$. Then one easily checks that this has the universal property of an equalizer in Err.

On the other hand, let $p : B \to Q$ be the coequalizer in Set. Equip Q with an error structure by setting $e(q) = \inf_{p(b)=q} e(b)$. Then this is again easily seen to have the universal property of a coequalizer in Err.

Let $\{A_i\}_{i \in I}$ be a family of objects of Err. Equip the product $\prod_i A_i$ with the error structure $e((a_i)) = \sup_i e_{A_i}(a_i)$. Then one easily checks that this satisfies the universal property of a product. Equip the disjoint union $\coprod_i A_i$ with the error structure $e(a) = e_{A_i}(a)$ when $a \in A_i$. Then one can, again, check that this has the right universal property.

It is clear from this construction that the underlying set functor preserves all limits and colimits. □

Definition 2.66. An Err-category D is *finite* if the underlying category $|\mathsf{D}|$ is finite, i.e has finitely many objects and morphisms. A *finite* limit or colimit in an Err-category is a (co)limit weighted by a functor $W : \mathsf{D} \to \mathsf{Err}$, where D is finite and each weighting set $W(d) \in \mathsf{Err}$ is a finite set.

Remark 2.67. Observe that infinite colimits of spaces with finite error may have points of infinite error. For example, the colimit of the digram $* \hookrightarrow *_1 \hookrightarrow *_2 \hookrightarrow \dots$ is $*_\infty$.

We now turn to weighted limits and colimits in Err-categories. Our first result shows that we may replace any such (co)limit with one where each weight is a singleton. Thus, the additional information of a weight reduces to associating an error with each object of the diagram.

Proposition 2.68. Given any Err-category I with a weight $W : \mathsf{I} \to \mathsf{Err}$, there is a replacement $\mathsf{I}', W' : \mathsf{I}' \to \mathsf{Err}$, equipped with a functor $a : \mathsf{I}' \to \mathsf{I}$ and a natural transformation $\alpha : W' \to aW$, such that for any Err-category C and any functor $D : \mathsf{I} \to \mathsf{C}$,

$$\operatorname{colim}^{W}_{i \in \mathsf{I}} D(i) \cong \operatorname{colim}^{W'}_{i \in \mathsf{I}'} D(a(i)),$$

in the sense that if either of these colimits exists, the other one does, and the canonical map is an exact isomorphism.

Proof. We simply obtain I' as the category of elements of W — an object in I' is a pair $(i \in I, x \in W(i))$, and a map $(i, x) \to (i', x')$ is a map $f : i \to i'$ so that $W(f)(x) = x'$ — the error of such a map is just the error of f in I. The functor $a : \mathsf{I}' \to \mathsf{I}$ is simply the obvious forgetful functor. The weighting $W' : \mathsf{I} \to \mathsf{Err}$ takes (i, x) to $*_{e(x)}$, where e is the error on $W(i)$. The natural transformation $\alpha : W' \to aW$ takes $* \in W'(i, x)$ to $x \in W(a(i, x)) = W(i)$. It is essentially obvious that the universal property of the two colimits are the same, giving the desired result. □

(Of course, by duality, the analogous result holds for limits)

Now we describe weighted limits and colimits in Err itself.

Proposition 2.69. Let $W : \mathsf{I} \to \mathsf{Err}$ be a weighting functor, and let $D : \mathsf{I} \to \mathsf{Err}$ be any (enriched) diagram. Suppose $|W(i)|$ is a singleton for each $i \in \mathsf{I}$. Then the weighted limit $\lim^W D$ is given by the following

1. As underlying set, $\left|\lim^W D\right| = \lim_i |D(i)|$, the limit computed in sets.
2. The error function is given by

$$e((x_i \in |D(i)|)_{i \in \mathsf{I}}) = \max(0, \sup_i e_{D(i)}(x_i) - e_{W(i)}(*))$$

If $W : \mathsf{I}^{\mathrm{op}} \to \mathsf{Err}, D : \mathsf{I} \to \mathsf{Err}$ are as before, then the weighted colimit $\mathrm{colim}^W D$ is given by

1. As underlying set, $\left|\mathrm{colim}^W\right| = \mathrm{colim}|D(i)|$, the colimit computed in sets.
2. The error function given by

$$e([x \in D(i)]) = \inf_{j, x' \in D(j), x' \sim x} e_{D(j)}(x') + e_{W(j)}(*)$$

   In other words, the error of an equivalence class $[x]$ in the colimit is the infimum of the value $e_{D(j)}(x') + e_{W(j)}(*)$ over all representatives $x' \in D(j)$ of the equivalence class.

Proof. Both statements follow immediately from writing out the universal property. Let's see this in a bit more detail.

The universal property of the described limit says that a map $f : X \to \lim^W D$ is equivalent to a family of maps $f_i : X \to D(i)$, compatible with the maps $D(i) \to D(j)$. Clearly the described set has this property. The error on the morphism space $\mathsf{Err}(X, \lim^W D)$ described by the universal property says that, under the above correspondence,

$$e(f) = \max(0, \sup_i (e(f_i) - e_{W(i)}(*))).$$

Again, it's clear that this is the same error function we get if we put the error described in the proposition on $\lim^W D$.

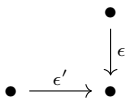
The result about colimits follows in the same way, by considering the universal property coming from the weighted colimit. □

Definition 2.70. An object X of an Err-category C is *terminal* if, for each other object Y, there is a unique morphism $Y \to X$, and that morphism is exact.

Definition 2.71. Let C be an Err-category and consider this diagram

$$\begin{array}{ccc} & & A \\ & & \downarrow f \\ B & \xrightarrow{g} & C \end{array}$$

Suppose $e(f) = \epsilon, e(g) = \epsilon'$ This diagram depicts a functor from the category $P_{\epsilon,\epsilon'}$ to C, where $P_{\epsilon,\epsilon'}$ is the Err-category of this shape:



(Where the edge labels denote the errors). A *pullback* of the diagram above is a limit of this diagram weighted by a functor $W : P_{\epsilon,\epsilon'} \to \mathsf{Err}$, where each $W(-)$ is a singleton. A pullback is *exact* if each weight $W(-)$ is the terminal object $*$.

The point of this definition is that a pullback may be a non-conical limit — i.e, we may specify some specific nonzero error tolerance on the universal maps into the pullback.

Definition 2.72. Let C be an Err-category. A product is a limit weighted by $W : \mathsf{D} \to \mathsf{Err}$, where D has no morphisms and each $W(-)$ is a singleton.

Lemma 2.73. If an Err-category has pullbacks and a terminal object, it also has finite products and equalizers.

Proof. Finite products may be constructed in the usual way from terminal objects and binary products. Moreover, the pullback $A \times_* B$ is easily verified to have the same universal property as the binary product $A \times B$, regardless of the weights.

An equalizer of $f, g : A \rightrightarrows B$ is the same thing as a pullback $A \times_{A \times B} A$, with the two maps $A \to A \times B$ being given respectively by $(1_A, f)$ and $(1_A, g)$ (with the weights being all zero in this case). □

Lemma 2.74. An Err-category has finite limits if and only if it has pullbacks and a terminal object. A functor preserves these limits if and only if it preserves the terminal object and pullbacks.

Proof. The forward direction of both statements is obvious, since pullbacks and terminal objects are special cases of finite limits.

Let $W : \mathsf{I} \to \mathsf{Err}$ be a functor from a finite Err-category, with $W(i)$ a finite set for all $i \in \mathsf{I}$. Let $D : \mathsf{I} \to \mathsf{C}$ be another Err-functor. Unwinding the definition, a W-weighted limit of D consists of an object $L \in \mathsf{C}$, equipped with maps $W(i) \to \mathsf{C}(L, D(i))$ for each i, so that the natural diagrams commute, and so that given another such object L', there is a unique map $L' \to L$ with error 0 making all the diagrams commute.

This is the same universal property as an equalizer of the two parallel morphisms

$$\prod_{i \in \mathsf{I}, x \in W(i)} D(i) \overset{f}{\underset{g}{\rightrightarrows}} \prod_{\varphi : i \to j \in \mathsf{I}, x \in W(i)} D(j)$$

where

$$f((a_{ix})_{i \in \mathsf{I}, x \in W(i)}) = (D(\varphi)(a_{ix}))_{\varphi : i \to j, x \in W(i)}$$

$$g((a_{ix})_{i\in \mathsf{I}, x\in W(i)}) = (a_{jW(\varphi)(x)})_{\varphi:i\to j, x\in W(i)}$$

so it suffices that the category has equalizers and finite products. This is a direct consequence of the assumptions and Lemma 2.73.

An analogous argument shows the claim about functors: a cone is a limit cone if and only if the induced map into the expression of the limit as an equalizer of products is an isomorphism. Hence it suffices to show that the functor preserves equalizers and finite products. The expression of these in terms of pullbacks and the terminal object, and the fact that it preserves these, establishes the claim. □

Lemma 2.75. In the category $[\mathsf{C}^{\mathrm{op}}, \mathsf{Err}]$, pullbacks preserve colimits, in the sense that $A \times_B \operatorname{colim}_i C_i \cong \operatorname{colim}_i A \times_B C_i$.

Note that in the notation above, we have suppressed the weights.

Proof. This is true for Err by applying Proposition 2.69 and writing out the resulting error sets. It follows for functor categories since (weighted) limits and colimits in such are computed pointwise. □

Proposition 2.76. Let $f : \mathsf{C} \to \mathsf{Err}$ be an Err-functor, and let $F : [\mathsf{C}^{\mathrm{op}}, \mathsf{Err}] \to \mathsf{Err}$ be the left Kan extension, i.e the unique colimit-preserving Err-functor with $F(yX) = f(X)$. Then F preserves finite limits if and only if it preserves pullbacks of representables and the terminal object.

The following proof is essentially a rewriting of [13, prop. 6.1.5.2] to our case. It is quite likely that a form of this result already exists in the literature for enriched categories, but we were unable to find it.

Proof. For convenience, we introduce the notation $P := [\mathsf{C}^{\mathrm{op}}, \mathsf{Err}]$. Note that a pullback of representables $yA \times_{yC} yB$ is not necessarily itself representable — this is the case if and only if C has pullbacks itself. It suffices by Lemma 2.74 to show that F preserves all pullbacks and the terminal object — the latter it does by assumption, so it suffices to verify that it preserves pullbacks. Let a map $X \to Y$ in $[\mathsf{C}^{\mathrm{op}}, \mathsf{Err}]$ be "good" if F preserves pullbacks of the form $X \times_Y Z$. (Note that we are not restricted to exact maps).

First, we claim that all maps between representables $A \to B$ are good. Note that pullbacks preserve colimits by Lemma 2.75, and every object of P is a colimit of representables. Since F also preserves colimits, and pullbacks of representables, it preserves every pullback of the form $A \times_B X$. So $A \to B$ is good.

Now let's say an object $B \in P$ is good if every morphism $A \to B$ is good. By repeating the above argument, we see that an object is good as long as every morphism $yA \to B$ from a representable is good. It follows that every representable is good, since we already saw that every map $yA \to yB$ was good.

Now we wish to show that every object is good. It suffices to show that the class of good objects is stable under coequalizers and coproducts.

First, consider the case of coproducts. Assume $Z = \coprod_i Z_i$, where each Z_i is good, and suppose we're given a pullback diagram

$$\begin{array}{ccc} S & \longrightarrow & yA \\ \downarrow & & \downarrow \\ yB & \longrightarrow & Z \end{array}$$

By the Yoneda lemma, mapping out of yA is the same as evaluating at A, which preserves colimits. Hence $\mathrm{Hom}(yA, Z) \cong \coprod_i \mathrm{Hom}(yA, Z_i)$. This implies $yA \to Z$ factors as $yA \to Z_i \to Z$. By assumption, $yA \to Z_i$ is good. Therefore it's enought to show that the map $Z_i \to Z$ is good. A similar argument applied to the other leg reduces us to showing that pullbacks of the form

$$\begin{array}{ccc} S & \longrightarrow & Z_i \\ \downarrow & & \downarrow \\ Z_j & \longrightarrow & Z \end{array}$$

are preserved by F. If $i = j$, one can show that the maps $S \to Z_i, Z_j$ are equivalences. It's now easy to see that this is also carried to a pullback. If $i \neq j$, the object S is the initial object. Since F preserves colimits, this is carried to the initial object of Err. Moreover, the coproduct $Z = \coprod Z_i$ is also carried to a coproduct in Err, where it also holds that the pullback of two distinct components of the coproduct is the initial object. Hence this pullback is preserved in every case. □

We now consider filtered colimits in Err-categories. (We will actually be interested mainly in cofiltered limits, but we stick with the conventional direction to be consistent with [1]). By definition a colimit is "filtered" if it is a colimit weighted by a flat presheaf — but in the case of Err-categories, the situation is much simpler. In fact any pair of index category and flat presheaf can be replaced (in the sense of Proposition 2.68) with one in which each weight $W(i) \in \mathsf{Err}$ is a singleton, and in this situation the weighting is flat if and only if the indexing category is filtered in the ordinary sense.

Proposition 2.77. Let $\mathsf{I}, W : \mathsf{I} \to \mathsf{Err}$ be an Err-category and a weighting, such that $W(i)$ is always a singleton — write $W(i) = *_{e_i}$ Then W is flat if and only if I is filtered in the usual sense, and $\inf_i e_i = 0$. (These morphisms are not supposed to be exact)

Proof. Flatness means that the left Kan extension $\mathrm{Lan}_y W : [\mathsf{I}^{\mathrm{op}}, \mathsf{Err}] \to \mathsf{Err}$ preserves finite limits. By standard results (see e.g. [12, Prop. 2.3.5]) the formula for that Kan extension is $\mathrm{Lan}_y W(F) = \int^i W(i) \otimes F(i)$. The terminal functor is $F(-) = *$. This is taken to $\int^i W(i) = \mathrm{colim}_i W(i)$. By Lemma 2.65, the underlying set of this colimit is a singleton. Hence it has the form $*_e$ for some e. There are inclusions $*_{e_i} \to *_e$, so $e \leq e_i$ for all i, implying $e = 0$, as desired.

Now consider a pullback $F \times_G H$. We consider

$$\int^i W(i) \otimes F(i) \times_{G(i)} H(i).$$

Using Lemma 2.65 again, since coends can be expressed as colimits, we can use the same coend to compute the undelying set. It's $\int^i F(i) \times_{G(i)} H(i) = \mathrm{colim}_i F(i) \times_{G(i)} H(i)$. Since filtered colimits commute with finite limits, we can rewrite this as

$$\int^i F(i) \times_{\int^i G(i)} \int^i H(i)$$

This is also the underlying set of

$$\int^i W(i) \otimes F(i) \times_{\int^i W(i) \otimes G(i)} \int^i W(i) \otimes H(i)$$

Writing out the definitions shows that these sets also have the same error, as desired. □

3 Probability theory

We will now develop certain preliminaries from probability theory which we will need in the rest of the thesis. There is nothing seriously deep here, but we go through the material for ease of reading. The main points of interest are

1. How to do probability theory with diagrams in the symmetric monoidal category FinStoch.
2. Technical properties of the Jensen-Shannon divergence.

We also note here, for convenience, some notation for a few distributions we'll use later:

Definition 3.1.

1. We denote by $\mathrm{Bern}(p)$ the Bernoulli distribution with parameter p — it is 1 with probability p and 0 with probability $1 - p$.
2. We denote by $\mathrm{Pois}(\lambda)$ the Poisson distribution with rate λ. It takes the value n with probability
$$\frac{\lambda^n e^\lambda}{n!},$$
for each $n \in \mathbb{N}$ (our natural numbers include 0)
3. For some $N \in \mathbb{N}$, we denote by $\mathrm{Pois}_{\leq N}(\lambda)$ a distribution which we call a *truncated Poisson distribution*. It takes the value n with probability
$$\frac{\lambda^n e^\lambda}{n!},$$
for each $n = 0, \ldots, N-1$, and takes the value N with the residual probability
$$1 - \sum_{n=0}^{N-1} \frac{\lambda^n e^\lambda}{n!}$$

3.1 FinStoch

Definition 3.2. The monoidal category FinStoch has

1. Objects the finite sets.
2. Maps $\mathcal{X} \to \mathcal{Y}$ given by $\mathcal{X} \times \mathcal{Y}$ stochastic matrices.
3. Monoidal product given by Cartesian product of sets.

By convention, there is a unique stochastic matrix $\emptyset \to \mathcal{X}$ for all $\mathcal{X}$, and no stochastic matrix $\mathcal{X} \to \emptyset$ for nonempty $\mathcal{X}$ (because an empty column can't sum to 1).

Definition 3.3. Given a function $f : \mathcal{X} \to \mathcal{Y}$ between two finite sets, there is an obvious associated stochastic matrix $\mathcal{X} \to \mathcal{Y}$ given by $f_{xy} = \delta_{f(x),y}$. We will generally abuse notation and simply denote this stochastic matrix by f as well.

This assignment defines a functor $\mathsf{Fin} \to \mathsf{FinStoch}$.

In addition, we have the following important operations in FinStoch:

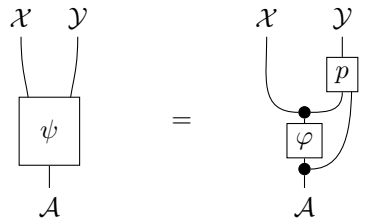
Definition 3.4. Let $\mathcal{X} \in \mathsf{FinStoch}$ be a finite set.

1. The unique stochastic matrix $\mathcal{X} \to *$ is denoted $\mathrm{del}_{\mathcal{X}}$
2. The stochastic matrix associated to the diagonal $\mathcal{X} \to \mathcal{X} \times \mathcal{X}$ (i.e the map $x \mapsto (x, x)$) is denoted $\mathrm{copy}_{\mathcal{X}}$.

Remark 3.5. A stochastic matrix $f : \mathcal{X} \to \mathcal{Y}$ is equivalent to a Markov kernel $\mathcal{X} \to \mathcal{Y}$, where both sets carry the powerset σ-algebra (see e.g. [10] for a discussion of Markov kernels). We can think of such a matrix as a "stochastic function", which receives an input $x \in \mathcal{X}$ and produces an output in $\mathcal{Y}$, but does so nondeterministically, according to the distribution $f(x)$. The composition rule for stochastic matrices is then "independent composition" — composing the two functions under the assumption that they use independent sources of randomness. We will sometimes call stochastic matrices "kernels" or "maps", when we want to emphasize their role as "functions", rather than matrices.

Note that a kernel $* \to \mathcal{X}$ is the same thing as a probability distribution on $\mathcal{X}$.

Definition 3.6. Let $\psi : \mathcal{A} \to \mathcal{X} \times \mathcal{Y}$ be a Markov kernel. We say that a kernel $p : \mathcal{A} \times \mathcal{X} \to \mathcal{Y}$ is a *conditional distribution of* $Y \in \mathcal{Y}$ *given* $A \in \mathcal{A}$ *and* $X \in \mathcal{X}$ if there exists $\varphi : \mathcal{A} \to \mathcal{X}$ so that we have the following identity:



(5)

In this diagram, the black circles denote either del or copy, as applicable.

Remark 3.7. Definition 3.6 is a definition of conditional distributions suitable for parameterized joint distributions. Dealing with such distributions is necessary if we want to combine conditional and interventional distributions. (We don't delve into that in this thesis). In the case $\mathcal{A} = \{*\}$, we recover the usual situation of a joint distribution on a product set.

To spell out the connection with the normal definition of conditional distribution: a map $p : \mathcal{A} \times \mathcal{X} \to \mathcal{Y}$ is a conditional distribution for $\psi : \mathcal{A} \to \mathcal{X} \times \mathcal{Y}$ if and only if, for all $a \in \mathcal{A}$, and for all $x \in \mathcal{X}$ with nonzero probability given a, the distribution $p(a, x)$ is the conditional distribution of Y given $X = x$ and $(X, Y) \sim \psi(a)$. This is also the reason we say *a* conditional distribution and not *the* conditional distribution.

For a more thorough discussion of this point (from a categorical point of view), see e.g [5, section 11] (in particular definition 11.5 and remark 11.6).

Note that the operation $(\mathcal{X}, \mathcal{Y}) \mapsto \mathcal{X} \times \mathcal{Y}$ is *not* a categorical product, in the sense of Definition 2.54. This is for the simple reason that a joint distribution is not uniquely determined by the marginals! Given distributions on $\mathcal{X}$ and $\mathcal{Y}$ — that is, maps $* \to \mathcal{X}, \mathcal{Y}$ — there are in general many possible pairings $* \to \mathcal{X} \times \mathcal{Y}$.

(On the other hand, $\times$ does give a categorical product in Fin, the category of finite sets and ordinary functions).

3.2 Jensen-Shannon divergence

We will later want to compare distinct probability distributions, and we would like a quantitative measure of "how close" they are to one another. We defer a discussion of the pros and cons of various such measures to Remark 4.20, after we have seen the context in which we'll be using this measure. For now, we just note that it it extremely natural from the categorical point of view to ask that this distance measure extends in some way to an enrichment of $\mathsf{FinStoch}$ in Met. The object of this subsection is to show that the square root of the *Jensen-Shannon divergence* JSD does this. We refer to [3] or [24] for information on JSD, but we recount the basic facts here.

We will make use of some basic notions of information theory — for an introduction to these terms, see [14]. We recall the necessary definitions here:

Definition 3.8. Let p be a probability distribution on a finite set $\mathcal{X}$. Then the *entropy* of p is

$$H(p) := \sum_{x \in \mathcal{X}} -p(x) \lg p(x)$$

If X is a random variable, $H(X)$ denotes the entropy of the distribution of X, which we also call simply the entropy of X.

Remark 3.9. Here lg denotes the base two logarithm. There is some disagreement about whether to use base 2 or base e for information theory. The natural logarithm, base e, has the obvious advantage of being more canonical. Base 2 has the advantage that the unit of information becomes the familiar bit. Thus

for example $H(\text{Bern}(0.5)) = 1$. Of course, the only difference is a constant factor, so this is not a particularly important choice. We use base 2.

Remark 3.10. The idea of entropy is that $H(X)$ is the amount of information, in expectation, received by learning the value of X. Thus $H(\text{Bern}(0))$ is zero, since the value is already known, while, as remarked above, $H(\text{Bern}(0.5))$ is one — learning the outcome of a random coin flip, we learn exactly one bit of information.

Definition 3.11. Given two random variables X, Y, with joint distribution $p(x, y)$, the *conditional entropy* is defined as

$$H(X|Y) = \sum_{x,y} -p(x, y) \ln \left(\frac{p(x, y)}{p(y)} \right)$$

Their *mutual information* is defined as

$$I(X; Y) = H(X) + H(Y) - H(X, Y) = H(X) - H(X|Y)$$

(The claimed identity is easily verified)

The idea here is that relative entropy measures the expected information contained in X, given that we've already learned the value of Y. For example, if $X = Y$, relative entropy is zero. Then the mutual information is the amount of information that's "shared" between the variables — how much less information is in X if Y is already known?

Definition 3.12. For p, q two distributions on a finite set $\mathcal{X}$, $\text{JSD}(p, q)$ denotes the Jensen-Shannon divergence between them, which is defined as the entropy of an equal mixture of p and q, minus the average entropy of p and q:

$$\text{JSD}(p, q) := H \left(\frac{1}{2}p + \frac{1}{2}q \right) - \frac{1}{2} \left(H(p) + H(q) \right)$$

Note that there are many equivalent definitions of this quantity. We will be particularly interested in the following:

Lemma 3.13. Let p, q be distributions on $\mathcal{X}$. Let B be an unbiased random coin, i.e B is a random variable with $B \sim \text{Bern}(0.5)$ Let X be a random variable which distributed according to p if $B = 0$ and according to q if $B = 1$. Then $\text{JSD}(p, q)$ is the mutual information between X and B

We refer to [24] for a proof — it is essentially a straightforward calculation from the definition. Since $I(X; B) = H(B) - H(B|X) \leq H(B)$, we have the following property of JSD:

Corollary 3.14. $\text{JSD}(p, q) \leq 1$

Lemma 3.15. Let p, q be two distributions on $\mathcal{X}$. Let $f : \mathcal{X} \to \mathcal{Y}$ be a kernel. Then $\text{JSD}(p, q) \geq \text{JSD}(fp, fq)$

Proof. Let X, B be as in Lemma 3.13. Observe that $f(X)$ is a random variable distributed according to fp if $Z = 0$ and fq if $Z = 1$. Hence $\mathrm{JSD}(fp, fq) = I(f(X), Z)$. But clearly this is less that $I(X, Z)$ — postprocessing X with an independent kernel can only remove some of the mutual information, not add it. □

Lemma 3.16. Let $f_0, f_1 : \mathcal{X} \to \mathcal{Y}$ be kernels. Let p be a distribution on $\mathcal{X}$. Then $\mathrm{JSD}(f_0 p, f_1 p) \leq \sup_x \mathrm{JSD}(f_0(x), f_1(x))$.

Proof. Let $X \in \mathcal{X}$ be distributed according to p. Let B be an unbiased random coin, and let $Y \in \mathcal{Y}$ be distributed according to $f_B(X)$. Then

$$\mathrm{JSD}(f_0 p, f_1 p) = I(Y; B),$$

$$\mathrm{JSD}(f_0(x), f_1(x)) = I(Y; B|X = x).$$

Hence our claim is that there exists x such that $I(Y; B) \leq I(Y; B|X = x)$. We insert the entropy formula for mutual information:

$$H(B) - H(B|Y) \leq H(B|X = x) - H(Z|B, X = x)$$

Since B and X are independent, we can cancel the first term on each side, giving the inequality

$$H(B|Y) \geq H(B|Y, X = x)$$

Observe that the expected value of $\mathbb{E}_x H(B|Y, X = x)$ is precisely $H(B|Y, X) \leq H(B|Y)$ (since conditioning always reduces entropy). Hence there is at least one x so that this inequality holds. □

Lemma 3.17. Let p_1, p_2 be distributions on $\mathcal{X}$, and q be a distribution on $\mathcal{Y}$. Then $\mathrm{JSD}(p_1 \otimes q, p_2 \otimes q) = \mathrm{JSD}(p_1, p_2)$.

Proof. This is an immediate consequence of Lemma 3.13 We learn the value of a random variable valued in $\mathcal{X} \times \mathcal{Y}$, distributed according to $p_Z \otimes q$. Since the second coordinate is independent of the first, it confers no information about Z (but of course, it also doesn't confer negative information). □

Definition 3.18. We define the distance $d_{\mathrm{JSD}}(f, g) = \sup_x \left(\sqrt{\mathrm{JSD}(f(x), g(x))}\right)$ for $f, g \in \mathsf{FinStoch}(\mathcal{X}, \mathcal{Y})$

Proposition 3.19. The distance d_{JSD} defines an enrichment of $\mathsf{FinStoch}$ in Met.

Proof. It is a standard result that the square root of the Jensen-Shannon entropy defines a metric (see eg [3]). It's a straightforward consequence of this that taking the supremum over the codomain defines a metric as well. The last condition to define an enriched category is that the composition map

$$\mathsf{FinStoch}(\mathcal{X}, \mathcal{Y}) \times \mathsf{FinStoch}(\mathcal{Y}, \mathcal{Z}) \to \mathsf{FinStoch}(\mathcal{X}, \mathcal{Z})$$

is short in each variable — this follows from Lemmas 3.15 and 3.16. □

⊙⊙ **Proposition 3.20.** The monoidal structure of $\mathsf{FinStoch}$ is compatible with the enrichment Proposition 3.19, in the sense that it defines a monoidal Err-category, see e.g. [16].

Proof. The content of this proposition is that the tensor product map

$$\mathsf{FinStoch}(\mathcal{X}, \mathcal{Y}) \times \mathsf{FinStoch}(\mathcal{X}', \mathcal{Y}') \to \mathsf{FinStoch}(\mathcal{X} \times \mathcal{X}', \mathcal{Y} \times \mathcal{Y}')$$

is short in each variable. This is a direct consequence of Lemma 3.17 □

4 Finite Graphical models

We're now ready to start delving into the main topic of the thesis: transformations between graphical models. As noted in the introduction, the starting point of our theory is *finite DAG models*. We have chosen this class of models mainly to avoid technical issues involving convergence, almost-everywhere equality, and so on. Our grahical models are equivalent to structural causal models (SCMs, see e.g. [19, Def. 6.2]) with finite, acyclic underlying graph, and where each variable takes values in a finite set. Note that we will generally not concern ourselves with the *counterfactual* ([19, Sec. 6.4]) behavior of our models, but only the interventional distributions. This decision means our transformations correspond to Rubenstein et al's exact transformations — but see Remark 4.17.

Definition 4.1. A *graphical model* M consists of the following data:

1. A directed acyclic graph $G = G(M)$.
2. For each vertex $X \in G$, a finite set $M[X]$, and a stochastic matrix
$$M[\varphi_X] : \prod_{Y \in \mathrm{pa}_G(X)} M[Y] \to M[X]$$

Given a subset $A \subseteq G$ of the vertices, we write $M[A] := \prod_{X \in A} M[X]$.

Remark 4.2. In our models, the vertices are generally the most important part. Therefore, we write things like $X \in G$ to mean "X is a vertex of G", and denote by $f : G \to G'$ a function from the vertices of G to the vertices of G'.

Definition 4.3.

1. Given a model M, a subset $X \subseteq G$, and a variable y such that all its direct parents are in X, we abuse notation and denote by $M[\varphi_y]$ the map $M[X] \to M[y]$ given by applying $M[\varphi_y]$ to the relevant coordinates of $M[X]$, and throwing away the rest.
2. In the situation above, there is an obvious map
$$(1_{M[X]}, M[\varphi_y]) : M[X] \to M[X \cup \{y\}] = M[X] \times M[y]$$
given by retaining the values from $M[X]$ and generating $y \in M[y]$ by $M[\varphi_y]$.

3. Given a set $X \subseteq G$, we can select an ordering of $G \setminus X$, $y_1, \dots y_n$ such that all of y_i's parents are contained in $X \cup \{y_1, \dots y_{i-1}\}$, for $i = 1 \dots n$. (It is clear that such a set exists, since the graph is finite and acyclic, although it is equally clear that the ordering is not necessarily unique). Then we can construct the composed map

$$M[X] \to M[X \cup \{y_1\}] \to M[X \cup \{y_1, y_2\}] \to \cdots \to M[G]$$

We call this map the *interventional distribution given* X and denote it by $P(-|\operatorname{do}(X = \cdot))$ or $P_{\operatorname{do}(X=\cdot)}$.

4. For the case $X = \emptyset$, we find a map $* \to M[G]$, which is called the *observational distribution.*

5. Given two subsets $X, Y \subseteq G$, we can define the interventional marginal distribution $M[X] \to M[Y]$ as the composition $M[X] \to M[G] \to M[Y]$ of the interventional distribution with the map that simply discards the values not in Y.

Remark 4.4. The notation for these kernels can get a little clunky — in practice, once you know the domain and codomain, there is usually no chance of confusion.

Of course, it is not clear that the interventional distributions are well-defined. The intuition is clear enough: the interventional distributions describe the distributions that obtain if we fix the value of the variables in X and then "run" the model to produce the remaining values. The fact that disparate parts of the model "run independently" implies that it doesn't matter which order one does this in. The following proposition shows that it is, in fact, well-defined.

Proposition 4.5. The interventional distribution is well-defined.

For this proposition, we will need the following lemma.

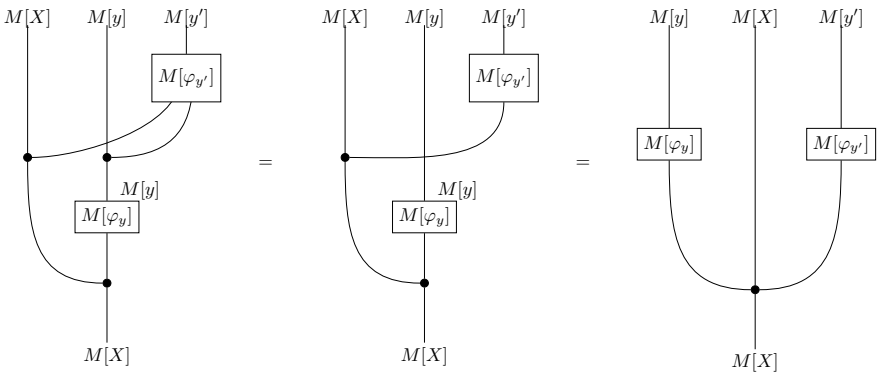
Lemma 4.6. Let S be a finite partially ordered set. Let $A : s_1 \dots s_n$ and $B : s'_1 \dots s'_n$ be two totalizations of the ordering on S — in other words, two ways of arranging the elements of S in a nondecreasing sequence. Then one can turn A into B by a finite sequence of transpositions, where each transposition exchanges two adjacent, incomparable elements.

Proof. Let's show that any nondecreasing sequence can be turned into B by such a sequence of transpositions — this is really the content of the lemma. Define the error of a sequence $s_1 \dots s_n$ as the total number of pairs i, j so that s_i and s_j are not in the same order as in B. Clearly if the error is zero, we must already be in sequence B. Suppose the error is greater than zero. Then there must be a pair of consecutive elements, s_i, s_{i+1}, that are in the wrong order compared to the ordering B. They must also be incomparable — we can't have $s_{i+1} \leq s_i$, since it's a nondecreasing sequence, and we can't have $s_i \leq s_{i+1}$: since B is nondecreasing, if this was true, they would be in the same order as

in B. Hence we can swap s_i and s_{i+1} — this decreases the error by 1. After a finite number of steps the error must be zero, and we have obtained B.

□

Proof of Proposition 4.5. By applying Lemma 4.6 to the vertices of G, partially ordered by causal dependence, we see that we can move between any two constructions of the interventional distribution by swapping two consecutive variables at a time. Hence it suffices to show that we may swap the order of two consecutive y_is, neither dependent on the other, without changing the final distribution. Consider the following diagram manipulation:



In the first step we use the fact that y' does not depend on y, so we may delete the $M[y]$ input to $M[\varphi_{y'}]$. Then we just rearrange the wires.

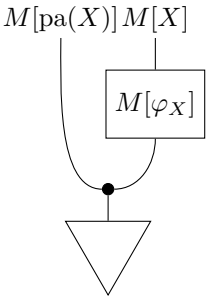
This shows that the composition $M[X] \to M[X \cup \{y\}] \to M[X \cup \{y, y'\}]$ is equal to another map $M[X] \to M[X \cup \{y, y'\}]$. Clearly a similar argument will show that the composition $M[X] \to M[X \cup \{y'\}] \to M[X \cup \{y, y'\}]$ is equal to the same map. This concludes our proof. □

Proposition 4.7. Each mechanism $M[\varphi_X] : M[\mathrm{pa}_G(X)] \to M[X]$ is a conditional distribution for the observational distribution $* \to M[\mathrm{pa}_G(X) \cup \{X\}]$.

Proof. Recall that given a distribution $* \to \mathcal{X} \otimes \mathcal{Y}$, a kernel $\mathcal{X} \to \mathcal{Y}$ is a conditional distribution if and only if we have the identity Eq. (5). After marginalizing out the other variables, the observational distribution on $M[\mathrm{pa}_G(X) \cup \{X\}]$ factors as

$$* \to M\mathrm{pa}_G(X) \xrightarrow{(1_{M[\mathrm{pa}_G(X)]}, M[\varphi_X])} M[\mathrm{pa}_G(X) \cup \{X\}].$$

Diagramatically, this looks like



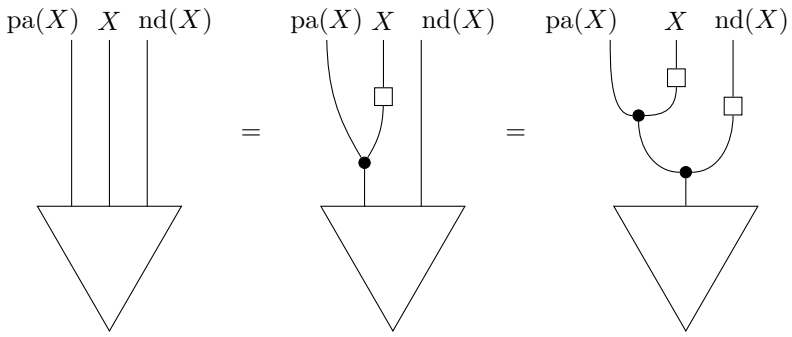
Which is exactly the statement we wanted. □

Of course, there is nothing surprising about this proposition — just as in the classical theory of graphical models, it holds more or less by construction. In classical treatments, this is usually not even considered interesting enough to make into a theorem, although it's implicit in most treatments of the Markov property for structure causal models, eg [19, prop. 6.31].

Proposition 4.8. The observational distribution satisfies the directed Markov property with respect to the DAG G.

Proof. We must prove that any variable X is independent of its nondescendants given its parents. Let's introduce the somewhat awkward notation $\mathrm{nd}(X)$ for the nondescendants of X, minus the parents. Then we are trying to show $\mathrm{nd}(X) \perp X \mid \mathrm{pa}(X)$.

Observe this diagram manipulation:



In the first step, we are factoring the observational distribution on $X \cup \mathrm{pa}(X) \cup \mathrm{nd}(X)$ as "sample the parents and nondescendants of X, then sample X conditional on the parents" — according to the definition, this is a possible choice for how to construct the observational distribution.

In the second step, we are factoring the distribution on $\mathrm{pa}(X) \cup \mathrm{nd}(X)$ as "sample the parents of X, then sample the nondescendants of X *according to the conditional distribution*". This is always possible, and gives us the diagram on the right.

By [5, Remark 12.2], this implies the conditional independence we wanted. □

Remark 4.9. In fact, Propositions 4.7 and 4.8 characterize the observational distribution uniquely. This can be proven diagramatically by using a diagramattic formulation of Proposition 4.8 to show that the observational distribution factorizes as a certain diagram, and then using Proposition 4.7 to show that the morphisms in this diagram may be replaced by the mechanisms.

In classical terms, this argument corresponds to arguing that the probability factorizes according to the graph, and that the factors must be precisely the mechanisms.

Definition 4.10. An *abstraction of models* $M \to M'$ consists of the following data:

1. A subset $R \subseteq G(M)$ of *relevant* variables.
2. A surjective map $f : R \to G(M')$.
3. For each $Y \in G(M')$, a surjective function $M[f^{-1}(Y)] \to M[Y]$

Definition 4.11. We let $* \in \mathsf{FinMod}$ denote the unique model with an *empty* underlying graph — in other words, $*$ has no variables.

It is clear that, given any other model M, there is a unique abstraction $M \to *$, given by $R = \emptyset$, the map $\emptyset \to \emptyset$, and so on. (In other words, $*$ is terminal in FinMod)

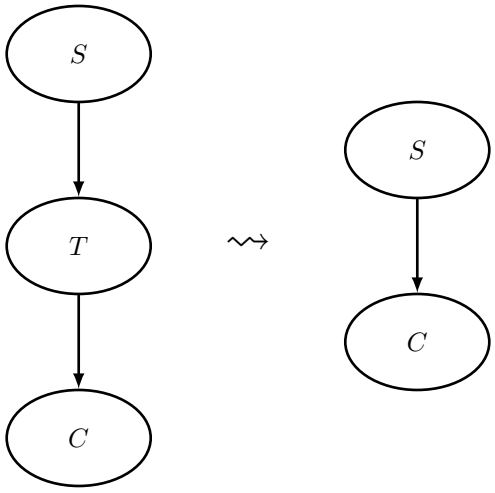


Figure 1: Example 4.12

Example 4.12. Consider a very simple model containing three binary variables. S measures whether or not a given person smokes. T measures the presence of

tar in that given person's lungs. C measures whether or not that person ends up with lung cancer. The causal structure of this model is as in Fig. 1. Suppose that $S \sim \text{Bern}(0.2)$, $T \sim \text{Bern}(0.8S)$, and $C \sim \text{Bern}(0.3T + 0.1)$[8]. We can consider a simplified model M' containing only S and C, with $C \sim \text{Bern}(0.24S + 0.1)$. There is an obvious abstraction $M \to M'$, with $R_f = \{S, C\}$, $f : \{S, C\} \to \{S, C\}$ the identity, and the maps $M[S] \to M'[S], M'[C] \to M'[C]$ identities.

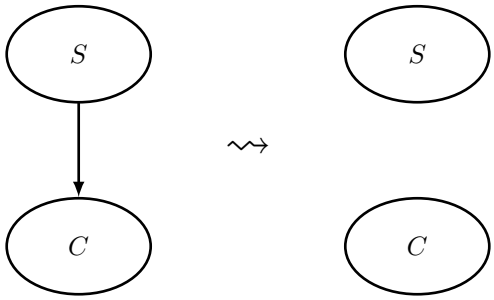


Figure 2: Example 4.13

Example 4.13. Consider the model M' from above — we have two variables, smoking and cancer, in a simple causal relationship. We may abstract this to a different model with the same variables and the same marginal distributions, but with $S \perp C$. Let's call this model $M_\perp$. In other words, $M_\perp$ has two variables S, C, a discrete causal graph, $S \sim \text{Bern}(0.2)$ and $C \sim \text{Bern}(0.148)$. See Fig. 2. There is again an obvious map $M \to M_\perp$ which is just the identity everywhere. This is a valid map of models — but of course, it distorts the distributions involved quite severely. To quantify the error of the map $M \to M_\perp$, we can calculate the Jensen-Shannon divergence between the two distributions on $\{0, 1\}^2$.

We will elide this annoying calculation, and simply note that the actual d_{JSD} (Definition 3.18) is approximately 0.0850.

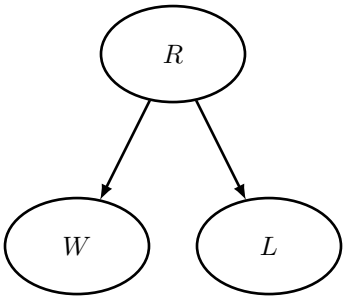


Figure 3: Example 4.14

[8]These numbers are made up, and probably *wildly* different from the actual statistics

Example 4.14. We can consider a model with three binary variables. W, which measures whether a given person drinks wine regularly. L, which meansures whether or not a person lived for a long time, and R, which measures whether or not a person is rich. We assume that the causal structure looks like Fig. 3. $R \sim \text{Bern}(0.1)$, $W \sim \text{Bern}(0.1 + 0.1R)$, $L \sim \text{Bern}(0.1 + 0.5R)$.

We can abstract this to a model containing only $W \to L$ — but now there is a non-obvious decision to make about the distribution. One natural choice is to maintain the observational distribution. This means that the causal distribution $P(L = 1|\operatorname{do}(W = w))$ is chosen to be the same as the conditional distribution $P(L = 1|W = w)$ in the original model. This maintains the observational distribution, but it introduces error in the interventional distribution, since the true interventional distribution $P(L = 1|\operatorname{do}(W = w))$ is *not* the same — the dependency between L and W is being confounded by R, which we have forgotten. Another choice would be to retain the causal relationship — since there is no causal relationship between L and W, this would mean setting them to be independent, with $L \sim \text{Bern}(0.15), W \sim \text{Bern}(0.11)$. This would destroy the statistical relationship between them, but provide the correct predictions about interventions. (Of course, one could also interpolate between these two models).

As exemplified above, our notion of abstraction places no requirements of *consistency* between the two models. The two models may lead to completely different predictions. The following *error* measure captures this difference:

Definition 4.15. Let $\alpha : M \to M'$ be an abstraction. Let X, Y be two sets of variables in $V(M')$. Then α induces a diagram

$$\begin{array}{ccc} M[\alpha^{-1}(X)] & \longrightarrow & M[\alpha^{-1}(Y)] \\ \downarrow{\scriptstyle \alpha} & & \downarrow{\scriptstyle \alpha} \\ M'[X] & \longrightarrow & M'[Y] \end{array}$$

which does *not* necessarily commute. We can compute the distance, in the sense of Definition 3.18, between the two paths around the square. Call this quantity $E_\alpha(X, Y)$. Now define the *error* of α to be

$$e(\alpha) := \sup_{X,Y \subseteq V(M'), \text{disjoint}} E_\alpha(X, Y).$$

Example 4.16. The calculation in Example 4.13 shows that the error of the map $M \to M_\perp$ is at least ≈ 0.085. The arrow $M \to M'$ in Example 4.12 har error zero — as long as we're only interested in the causal relationship between smoking and cancer, we may ignore the presence of tar.

Remark 4.17. Note that our error measure only accounts for differences between the *interventional* distributions — not between conditional distributions. The simplest example of this is to consider a model M with two variables X, Y, where X causes Y. Suppose simply that X is binomial distributed with $p = 0.5$,

and $Y := X$. We can transform this to a model M', also with two variables, X', Y', where $X' = X$ and $Y' = *$ (no matter what Y is). This abstraction is exact (it has error zero), because for any intervention on X, the passage to X' loses no information, while an intervention on Y has no effect on the transformed distribution. On the other hand, if we want to predict the *conditional* distribution $P(X = x|Y = y)$, it is clear that the passage to the high-level model loses relevant information — it is just not *causal* information.

An error measure which accounted for both conditional and causal information would correspond to the requirement that our transformations preserve not just interventional distributions, but also counterfactuals as well.

In [2], Chalupka et al consider a similar situation involving transformations between models — their case is slightly different, since it involves marginalizing out a causal node, but retaining information about its confounding effects, something which is not captured in our framework. They prove a theorem, which they call the *causal coarsening theorem*, saying that in their situation, if two values of the same variable lead to the same interventional distribution, they also lead to the same conditional distribution[9]

This suggests that there may be a strong connection between preserving internventional distributions and counterfactuals, which would certainly be worth investigating further.

Remark 4.18. We've seen in Examples 4.13 and 4.14 that certain abstrations are prevented from being exact just by the nature of the map $G \supseteq R \to G'$. This happens when the two causal structures are incompatible in some way. It is difficult to write down an exact list of criteria for when this happens, but we can list some essential cases:

1. When $R = V$, the structures "are compatible" as long as, for each edge $x \to y \in G$, there exists an edge $f(x) \to f(y) \in G'$, *or* $f(x) = f(y)$.

2. When G' is the subgraph obtained by deleting a vertex x from G, and $R = V(G')$, then the two structures are compatible if and only if x has at most one child.

Here "compatible" just means that the graphs do not present an obstruction to the existence of an exact transformation.

Lemma 4.19. Error satisfies $e(\alpha\beta) \leq e(\alpha) + e(\beta)$.

Proof. Let $\alpha : M \to M', \beta : M' \to M''$ be two abstractions. Let $X, Y \subseteq V(M'')$ be disjoint sets of variables.

Consider this diagram

[9] This is not entirely accurate — in fact, this result only holds for almost all (in the sense of Lebesgue measure) possible choices of conditional distributions.

$$\begin{array}{ccc} M[\alpha^{-1}(\beta^{-1}(X))] & \longrightarrow & M[\alpha^{-1}(\beta^{-1}(Y))] \\ \downarrow{\scriptstyle \alpha} & & \downarrow{\scriptstyle \alpha} \\ M'[\beta^{-1}(X)] & \longrightarrow & M'[\beta^{-1}(Y)] \\ \downarrow{\scriptstyle \beta} & & \downarrow{\scriptstyle \beta} \\ M''[X] & \longrightarrow & M''[Y] \end{array}$$

By assumption, the error of the top square is at most $e(\alpha)$, and of the bottom square is at most $e(\beta)$. It now follows from Lemma 2.48 that the error of the outer square is at most $e(\alpha) + e(\beta)$ as desired. □

Remark 4.20. At this point, let us discuss the choice of $\sqrt{\text{JSD}}$ as the distance function on probability distributions. There are several good candidates in the literature for such a distance function. Perhaps the most natural is the *Kullback-Leibler divergence* ([9]), which measures the information inefficiency from assuming that your data is distributed according to p, when it's really distributed according to q. The main issue with the Kullback-Leibler divergence is that it is asymmetric and does not satisfy a triangle inequality (not even when raised to some power — in the way that $\text{JSD}^{1/2}$ satisfies the triangle inequality). The asymmetry, while it certainly makes things more complicated, is not really a serious defect from our point of view — in our application, there is actually a "ground truth" and an "imperfect prediction" that we are comparing, so the asymmetry is not so unnatural. The main problem is the lack of a triangle inequality, which means Lemma 2.48 can't work. This means that we don't know how to prove Lemma 4.19 for KL-divergence. Without some version of this lemma, the whole notion of a compositional approach to abstraction is unworkable.

Another imporant class of distance functions are the *Wasserstein metrics* (See eg [22, Chapter 6]). As the name implies, these are a class of metrics on the space of (sufficiently "nice") probability distributions on a metric space X. These metrics do have the theoretical properties we require, but they require the choice of a metric on the sets in question. There is no technical reason we could not have developed our theory for finite metric spaces — the distance $d(x, y)$ measuring the cost of predicting x when the true value was y. However, we have decided not to add this layer of complication, finding the information-theoretic Jensen-Shannon distance sufficient for the present note.

Lemma 4.19 lets us make the following definition:

Definition 4.21. We define the category FinMod of finite models and abstractions to be the Err-enriched category where

- Objects are finite graphical models.
- Morphisms are abstractions
- The error of a morphism is its error as in Definition 4.15.

Example 4.22. Applying Definition 2.50, we obtain a (unenriched) category $\mathsf{FinMod}_{\mathrm{ex}}$ of *exact abstractions*. Its morphisms are those abstractions with zero error. This means that, as long as we are only concerned with the value of high-level variables, there is no predictive error in using the high-level model instead of the low-level one. This is essentially the notion of *exact transformation* as developed by Rubenstein et al in [21]. We note that our version enforces a strict correspondence between high-level and low-level variables which is not implied by their framework (and of course, we are considering a much less rich notion of causal model).

Remark 4.23. We emphasize that the error of an abstraction is not exactly a measure of "information loss" — the existence of an exact transformation $\alpha : M \to M'$ does not imply that M' contains "as much" information as M. It only implies that the value $\alpha(X) \in M'[x]$ contains as much causal information *about other variables of the form* $\alpha(Y)$ as does the untransformed variable X. As a degenerate example, if the model M' contains only one variable, any measure-preserving function is an exact transformation, because, trivially, $\alpha(X)$ contains all relevant information about itself.

Proposition 4.24. Let (M, G) and (M', G') be models, and let $f : M \to M'$ be an abstraction. Then f is an isomorphism in $\mathsf{FinMod}_{\mathrm{ex}}$ if and only if the following all hold:

1. $R_f = G$ — in other words, f defines a map $G \to G'$.
2. The map $G \to G'$ is a bijection.
3. The function $f_X : M[X] \to M[f(X)]$ is a bijection for each $X \in G$.
4. For each possible intervention $\mathrm{do}(X_1 = x_1, X_2 = x_2, \dots)$, when $M[G]$ is equipped with the inteventional distribution

$$P(- \mid \mathrm{do}(X_1 = x_1, X_2 = x_2, \dots)),$$

and $M'[G']$ with the distribution

$$P(- \mid \mathrm{do}(f(X_1) = f_{X_1}(x_1), \dots))$$

the map $f : M[G] \to M'[G']$ is measure-preserving.

Proof. First, suppose f is an exact isomorphism. Clearly, if f is only defined on a proper subset of G, so is gf for any g, so gf cannot equal the identity. So 1. holds. Similarly obvious, given an inverse f^{-1}, we have $f^{-1}f$ and ff^{-1} both equal to the identity. In particular their underlying map on graphs are equal to the identity. So $f : G \to G'$ has an inverse, and is therefore a bijection. The same argument shows that the maps on sets must be bijections. Finally, the last part is just the definition of exactness in this situation.

For the other direction, suppose a map satisfies 1-4. Then we can construct an inverse simply by putting $f^{-1} : G' \to G$ the inverse of f (existing because it's a bijection), $f^{-1}_{f(X)} : M[f(X)] \to M[X]$ an inverse of f_X, and so on. Condition 4. implies that this transformation will again be exact. □

The content of this proposition is that isomorphic finite graphical models are exactly those which have the same variables (up to renaming), the same outcome spaces (up to bijection), and the same probabilistic behavior under any intervention. This means for instance that isomorphic models can have distinct graphs — edges that play no causal role can be deleted (or added) freely.

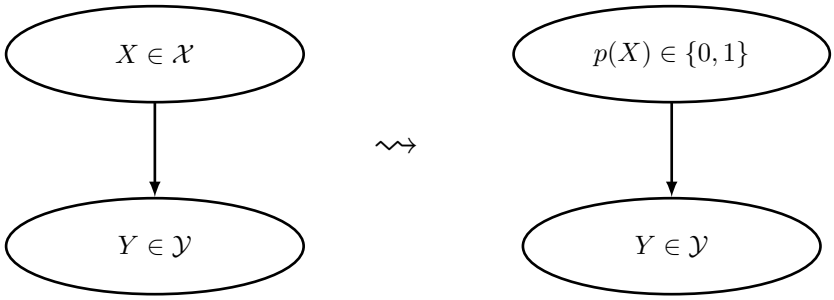


Figure 4: Example 4.25

Example 4.25. Let $\mathcal{X}, \mathcal{Y}$ be two finite sets, and let $f : \mathcal{X} \to \mathcal{Y}$ be a (deterministic!) function, and let ψ be a distribution on $\mathcal{X}$. This data described a model of two variables, $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ as depicted on the left of Fig. 4. Let furthermore $p : \mathcal{X} \to \{0,1\}$ be a partition of $\mathcal{X}$ into two classes. Then we can cook up a new model of two variables $p(X), Y$, where the distribution on X is given by $p(\psi)$, and the kernel $\{0,1\} \to \mathcal{Y}$ is given by $\bar{f} : \{0,1\} \to \mathcal{Y}$, sending $i \in \{0,1\}$ to the conditional distribution $P(f(x) = y | p(x) = i)$. This is the model on the right.

This example shows that we can "abstract a deterministic mapping into a stochastic one" — in this situation, the randomess in $\bar{f}$ reflects our uncertainty about which element of $\mathcal{X}$ we are actually holding.

The error of this abstraction is $\sup_{x \in \mathcal{X}} d(f(x), \bar{f}(p(x)))$ - the magnitude of the largest mistake we can make by inferring from the abstracted model instead of the unabstracted model.

We note that models with exactly one variable are essentially the same thing as finite probability spaces, in the following sense:

Definition 4.26. Let $\mathsf{FinProb} \subseteq \mathsf{Prob}$ be the category of finite probability spaces and measure-preserving maps.

Proposition 4.27. There is a functor $F : \mathsf{FinProb} \to \mathsf{FinMod}_{\mathrm{ex}}$ which takes a finite probability space $(\mathcal{X}, P)$ to the model $F(\mathcal{X})$ with $G(F(\mathcal{X})) = *$ (a DAG with exactly one vertex, $*$), $F(\mathcal{X})[*] = \mathcal{X}$, and the unique mechanism $F(\mathcal{X})[\varphi_*] : * \to \mathcal{X}$ given by P.

A measure-preserving map $f : \mathcal{X} \to \mathcal{Y}$ is taken to the map $F(f) : F(\mathcal{X}) \to F(\mathcal{Y})$ given by $R_{F(f)} = \{*\}$, the mapping on vertices being the only possible thing, and the map $\mathcal{X} \to \mathcal{Y}$ being f.

Moreover, every finite model with exactly one variable is in the image of this functor (up to isomorphism), and a map $f : M \to M'$ between such models is in the image of this functor if and only if $R_f = G(M)$.

Proof. Verifying functoriality of the described construction is more or less trivial. The fact that it respects identities and composition is a direct consequence of the definitions. It is also clear that the described map is, in fact, an exact transformation between models — since the interventional distributions are all degenerate, there is not much to check.

The claim that every finite model with one variable is isomorphic to one of this form is also easy to verify — simply take $\mathcal{X}$ to be $M[*]$ (where $*$ is the unique variable), and P to be the observational probability. The last thing to verify is that all maps with $R_f = G(M)$ between such models have this form (the converse is clear). Again, given such a map $f : F(\mathcal{X}) \to F(\mathcal{Y})$, it's easily verified that $f = F(f_*)$, with $f_* : F(\mathcal{X})[*] = \mathcal{X} \to \mathcal{Y} = F(\mathcal{Y})[*]$. □

This proposition means it makes some sense to treat a finite probability space as if it were a one-variable model. Hence we may pretend that $\mathsf{FinProb} \subseteq \mathsf{FinMod}$ — although we have to remember that there may be extra maps with $R = \emptyset$ between them, when considered as models.

We could have considered an Err-enriched category of finite probability spaces, with the error of $f : (\mathcal{X}, P) \to (\mathcal{Y}, P')$ given by $d_{\mathrm{JSD}}(f(P), P')$. The above construction extends to this larger category. However, we won't be thinking about this extra generality.

Remark 4.28. Rather than taking the supremum over the domain when defining our error measure, we could also have taken the expected value (or perhaps the square root of the expected Jensen-Shannon divergence, or something to that effect). Of course, if we had done this, we couldn't have used the slick definition in terms of the Met-enrichment of $\mathsf{FinStoch}$, but we could certainly still have made sense of a definition of this type.

The reason to consider worst-case, rather than expected error, is that we are trying to measure the precision of our predictions *under interventions*. We don't have access to a probability distribution on interventions — and moreover, since in general the character of the predictions may affect the interventions happening, it's not even clear that such a thing would be meaningful. To elaborate on this point, think of the example of predicting the effects of smoking on cancer. The prediction that smoking does or does not cause cancer may affect the prevalence of interventions do(Smoking = Yes). This is why the most sensible way of aggregating errors is to take the worst case.

5 Profinite models

So far, we have considered transformations between models only in the simplest possible case, namely the case of finite models. The obvious next step is to generalize this to cases involving infinite models, which are after all fairly common.

It turns out that the mere existence of a category of finite models gives us a natural way to define a category of potentially infinite models.

The basic idea is this: given an infinite model (supposing we had figured out some good definition of these), M, and a finite model M', we can consider the error space of transformations $M' \to M$, which we could denote $\mathsf{Mod}(M, M')$. Given a transformation $M' \to M''$, we obtain a map $\mathsf{Mod}(M, M') \to \mathsf{Mod}(M, M'')$. In other words, $\mathsf{Mod}(M, -)$ forms a functor $\mathsf{FinMod} \to \mathsf{Err}$. Now the claim is that this functor contains all the relevant information about M. This might be slightly controversial, and one can probably come up with reasonable definitions of "infinite model" for which it fails. The philosophical justification for such a claim is that the only properties of a model which have any practical statistical meaning are those that can be reflected in finite terms: we are always doing finitely many measurements (even if it's a huge number), and our measurement devices always have finitely many values (even if we measure with a huge number of decimals, it's a finite number!). Armed with this idea, we can *define* infinite models as certain functors $\mathsf{FinMod} \to \mathsf{Err}$. (We use the term "profinite" for these objects by analogy with objects such as profinite groups, which result from applying a similar construction to the category of finite groups)

Definition 5.1. A *profinite model* is a flat Err-functor $\mathsf{FinMod} \to \mathsf{Err}$. The Err-category of profinite models is $\mathsf{ProFinMod} := [\mathsf{FinMod}, \mathsf{Err}]^{\mathrm{op}}_{\mathrm{flat}} \subseteq [\mathsf{FinMod}, \mathsf{Err}]^{\mathrm{op}}$, the full subcategory of flat functors.

A profinite model simply *is* "something that can be approximated by finite models." We are applying the "Yoneda philosophy" of Lemma 2.36 backwards here — considering the functor $\mathsf{C} \to [\mathsf{C}, \mathsf{Err}]^{\mathrm{op}}$ given by $X \mapsto \mathsf{C}(X, -)$, instead of the functor $\mathsf{C} \to [\mathsf{C}^{\mathrm{op}}, \mathsf{Err}]$. This obviously means everything is the other way round, but it makes no technical difference for us.

Remark 5.2. "Flat" is a technical term, which we previously encountered in our discussion of limits in Err-categories. See [1] for a general treatment of flat functors in enriched categories (note that our Err satisfies their conditions on $\mathcal{V}$). We will soon see what it means in concrete terms for a functor $\mathsf{FinMod} \to \mathsf{Err}$ to be flat.

Remark 5.3. Given two profinite models X, Y, a transformation $\alpha : X \to Y$ between them consists of a natural family of functions $\alpha_M : Y(M) \to X(M)$ for each $M \in \mathsf{FinMod}$. The error of such a transformation is

$$e(\alpha) = \max(\sup_{M,m} e(\alpha_M(m)) - e(m), 0),$$

where M runs over all possible finite models, and m runs through $Y(M)$. In other words, for each finite model, given an abstraction of Y into M, we must specify an abstraction of X into M. And the error of this assignment is the supremal increase in error.

Remark 5.4. Each finite model M defines a profinite model $\mathsf{FinMod}(M, -)$. This assembles into an Err-functor $\mathsf{FinMod} \to \mathsf{ProFinMod}$, which we denote y.

This functor has the property that $\mathsf{FinMod}(M, M') \to \mathsf{ProFinMod}(yM, yM')$ is an isomorphism for all M, M' — in other words, it's fully faithful. Moreover, $\mathsf{ProFinMod}(X, yM) \cong X(M)$ for all X, M, by the (enriched) Yoneda lemma, Remark 2.53. (See also Remark 2.37). Note that we are working in the dual situation, considering an embedding $\mathsf{C}^{\mathrm{op}} \to [\mathsf{C}, \mathsf{Err}]$ instead.

This means we can essentially treat the profinite models as an extension of the collection of finite models — we can treat a finite model as a profinite model using the functor y, and this does not alter the relationship between finite models. Moreover, it makes the notion that $X(M)$ for $X \in \mathsf{ProFinMod}$ and $M \in \mathsf{FinMod}$ represents the set of abstractions $X \to M$ into a literal truth.

Remark 5.5. When considering maps $M \to M'$ from an infinite model to a finite model, we will sometimes take a slightly different perspective. Rather than viewing the finite model as an "approximation" or "abstraction" of the infinite model, we can instead think of it as a "view into M" — a way of focusing on a small piece of the big model. (Of course, we could also use this perspective for a finite model M). We may occasionally describe a map $M \to M'$ as a "view into M" when we wish to emphasize this point of view.

Proposition 5.6. Let $X : \mathsf{FinMod} \to \mathsf{Err}$ be a functor For each diagram in FinMod like this:

$$\begin{array}{ccc} & & A \\ & & \downarrow{\scriptstyle g} \\ B & \xrightarrow{f} & C, \end{array}$$

we can form the pullback $X(A) \times_{X(C)} X(B)$ in Err. We can also consider the set

$$S = \{(M, a : M \to A, b : M \to B, h \in X(M)) \mid ga = fb\}/\sim$$

where $\sim$ is the equivalence relation induced by identifying, for each quadruple

$$h \in X(M), p : M \to M', a : M' \to A, b : M' \to B$$

the two tuples

$$(M, ap, bp, h), (M', a, b, X(p)(h))$$

We can make this an error set by setting

$$e(x) = \inf_{(M', a', b', h') \sim x} \max(e(a), e(b)) + e(h).$$

In other words, the error of an equivalence class x is infimum of $\max(e(a), e(b)) + e(h)$ taken over all the representatives of the equivalence class. There is always a map $S \to X(A) \times_{X(C)} X(B)$ given by

$$(M, a, b, h) \mapsto (X(a)(h), X(b)(h)).$$

The functor X is flat if and only if this map is always an isomorphism of error sets, and $X(*) = *_0$.

⊙⊙ *Proof.* By definition X is flat if and only if the left Kan extension $L_y X : [\mathsf{FinMod}^{op}, \mathsf{Err}] \to \mathsf{Err}$ preserves finite limits. It preserves finite limits if and only if it preserves finite limits of representables, if and only if it preserves pullbacks of representables and the terminal object, by Proposition 2.76. Recall that the formula for the left Kan extension is

$$L_y X(F) = \int^M F(M) \otimes X(M)$$

If $F = yA$ is a representable, this is rewritten as

$$\int^M \mathsf{FinMod}(M, A) \otimes X(M),$$

which is equivalent to $X(A)$. If $F = yA \times_{yC} yB$ is a pullback of representables, we instead get

$$\int^M \mathsf{FinMod}(M, A) \times_{\mathsf{FinMod}(M,C)} \mathsf{FinMod}(M, B) \otimes X(M),$$

which is precisely the error set denoted S above. The statement that X preserves pullbacks of representables is then precisely the statement that the map from this set to $X(A) \times_{X(C)} X(B)$ is an isomorphism.

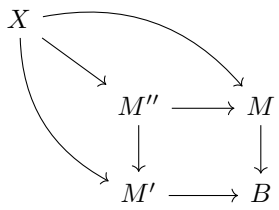
□

Remark 5.7. The intuition behind Proposition 5.6, which can be regarded as an alternative definitition of "profinite model", can be phrased as follows:

- Any finite family of finite abstractions of a profinite model admits a "common refinement", a single finite abstraction which contains all the information from those finite models.
- Any equations between the abstractions is "witnessed" by a common refinement.

To explain the second point, imagine that a profinite model X admits two views $\alpha \in X(M)$ and $\beta \in X(M')$, where M has causal graph $A \to B$ and M' has causal graph $B' \to C$ — in other words, we have given two ways of extracting two variables with a one-way causal relationship from X. Suppose further we have the equation $B = B'$, meaning those two random variables always agree (or agree with probability one, maybe). Then the second statement says we can find a common refinement $\gamma \in X(M'')$, where M'' has causal graph $A \to B \to C$, and α is the restriction to the first two variables, and β is the restriction to the second two. This refinement "witnesses" $B = B'$ in the sense that, given such a refinement, it obviously follows that $B = B'$ (because they are just two copies of the same variable).

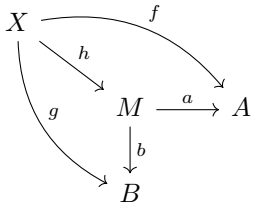
To phrase this diagramatically, consider the following diagram



If we know that the inner square commutes, and the maps $X \to M, M'$ factor over M'' as depicted, it obviously follows that the outer cell commutes. The flatness condition means that (up to some arbitarily small error), the converse holds — given that the outer square commutes, we can "fill out the middle" with a finite model M''.

We now give some corollaries of this somewhat abstract description.

Corollary 5.8. Let X be a profinite model. Given views $f, g : X \to A, B$ with error both less than or equal to ϵ. Let $\delta > 0$ be any real number. there exists a factorization diagram like so

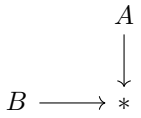


With M finite, and so that $\max(e(a), e(b)) + e(h) < \epsilon + \delta$.

In other words, we can form "common refinements" of models without increasing the error — the error gets "spread out" across the two steps of abstraction, but the total error does not increase.

Let us give this proof in some detail.

Proof. We consider the diagram in FinMod given by



(recall that $*$ is the terminal model, with zero variables)

Clearly the pullback $X(A) \times_{X(*)} X(B)$ in Err is just $X(A) \times X(B)$ The two given views identify a point (f, g) in this set with error less than or equal to ϵ.

On the other hand the set S from Proposition 5.6 can we rewritten as

$$\{M, a : M \to A, b : M \to B, h : X \to M\}/\sim$$

since the condition is trivial. The error on this set is

$$e(p) = \inf_{(M,a,b,h)\sim p} \max(e(a), e(b)) + e(h).$$

Hence the (f, g) corresponds to a point in this set with error less than ϵ — meaning we can find some representative (M, a, b, h) with $\max(e(a), e(b)) + e(h) < \epsilon + \delta$. This is exactly what we wanted. □

Example 5.9. Let G be a DAG which is not necessarily finite, but where the ancestral set[10] of each vertex is finite. Let for each variable $V \in G$ a finite set $M[V]$ be given, and a kernel $M[\mathrm{pa}(V)] \to M[V]$ in $\mathsf{FinStoch}$ be given. (This is simply the obvious meaning of a "model of G").

Then given a finite model M', we can define an abstraction $\alpha : M \to M'$ to be

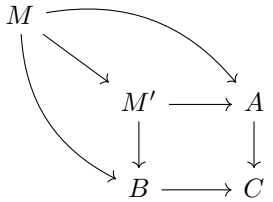
1. A finite, upwards-closed[11], full subgraph F_α of the vertices in G.
2. And an a an abstraction $M|_{F_\alpha} \to M'$, where $M|_{F_\alpha}$ is the obvious finite model obtained by restricting M to F_α.

We can define the error of such an abstraction to simply be the error of the "underlying" abstraction of finite models. There is an obvious way of composing such an abstraction with a further abstraction $M' \to M''$. This data gives a functor $\mathsf{FinMod} \to \mathsf{Err}$, which we can denote M — it sends the model M' to the set of abstractions $M \to M'$. (We are neglecting to verify that this satisfies the conditions of a functor).

This functor is in fact a profinite model. To see this, consider two abstractions $(F_\alpha, \alpha : M|_{F_\alpha} \to A)$ and $(F_\beta, \beta : M|_{F_\beta} \to B)$ Suppose that this fits into a diagram

$$\begin{array}{ccc} M & \longrightarrow & A \\ \downarrow & & \downarrow \\ B & \longrightarrow & C \end{array}$$

Then we can build a common factorization $\gamma : M \to M'$ by putting $F_\gamma = F_\alpha \cup F_\beta$, and $M' = M|_{F_\gamma}$. By construction, α and β factor over this, so that we have a diagram



[10]The ancestral set of y is the set of all vertices x so that there exists a directed path from x to y

[11]In the sense that, if $X \to Y$ is an edge and $Y \in F_\alpha$, then $X \in F_\alpha$

This diagram clearly commutes by construction, and it's also clear that the error of γ is zero, while the erros of the maps $M' \to A, B$ are exactly the errors of the maps $M \to A, B$. Hence the functor M satisfies the condition of Proposition 5.6.

In other words, we can define a view into M as a view into some finite subgraph of G.

5.1 Profinite models as limits

Profinite models provide a reasonable abstract theory of "infinitary" causal models. One big issue is that specifying the data of $X(M)$ for every possible graphical model is not always practical. We will now see a different way of constructing profinite models, namely as limits.

We have the following general result:

Proposition 5.10. The category $\mathsf{ProFinMod}$ has all cofiltered limits. Moreover, every profinite model is a cofiltered limit of finite models.

This is a case of an even more general result, [1, Cor. 2.2].

Remark 5.11. Unpacking the definitions, the previous proposition means that, for any profinite model M, we can find a system of finite models M_i indexed by a small (Err-)category I, such that there is a canonical abstraction $M \to M_i$ for each i, and so that an abstraction $M \to N$ for a general finite model N factors as $M \to M_i \to N$ — this choice of i being non-canonical.

Example 5.12. Let M be a model of an infinite graph as in Example 5.9. For each finite full upwards-closed subgraph $F \subseteq M$, we can consider the finite model $M|_F$. If $F \subseteq F'$, there is a natural marginalization $M|_{F'} \to M|_F$.

We can consider the Err-category $P(M)$ with objects the finite subsets F as above, a single exact morphism $F \to F'$ if $F \supseteq F'$, and no morphisms otherwise. Then the construction $F \to M|_F$ is a functor $P(M) \to \mathsf{FinMod}$. The limit of this functor in $\mathsf{ProFinMod}$ is M.

Example 5.13. Consider a Poisson process X with rate λ. This is a stochastic process taking place in continuous time. It takes values in $\mathbb{N}$. Given $X(t_0) = n$, the probability that $X(t_1) = m$ for $t_1 > t_0$

$$\frac{\lambda^{m-n}(t_1 - t_0)^{m-n} e^{\lambda(t_1 - t_0)}}{(m-n)!} \tag{6}$$

for $m \geq n$ and 0 for $m < n$. In other words, the value $X(t_1) - X(t_0)$ has distribution $\mathrm{Pois}(\lambda(t_1 - t_0))$.

We can make this into a causal model by allowing interventions at any time which set the current value X, and which don't affect the above conditional distributions.

Formally, given a finite set of times $0 = t_1 \leq t_1 \cdots \leq t_n$, and a natural number N, we define the finite model $X_{\{t_i\},N}$ to have n nodes labeled

$X(t_1), \dots X(t_n)$, an edge $X(t_i) \to X(t_{i+1})$ for each $i < n$, and no other edges. The outcome space of the variable $X(t_i)$ is in each case $\{0, 1, \dots, N\}$. The distribution of $X(t_0)$ is 0 with probability 1, and the kernels $X(t_i) \to X(t_{i+1})$ are given by the formula Eq. (6), except that every outcome above N is replaced with N — in other words, $X(t_{i+1}) - X(t_i)$ is distributed according to $\mathrm{Pois}_{\leq N - X(t_i)}(\lambda(t_{i+1} - t_i))$.

If $\{t_i\}_{i \in I} \subseteq \{s_j\}_{j \in J}$, and $N \leq N'$ there is an obvious exact abstraction $X_{s_j, N'} \to X_{t_i, N}$. The cofiltered limit $\lim_{\{t_i\} \subseteq \mathbb{R}_{\geq 0}, N \in \mathbb{N}} X_{t_i, N}$ "is" the (causal) Poisson process X.

Example 5.14. Consider a discrete dynamical system with variables $\{X_i\}$, taking values in the finite sets $\{\mathcal{X}_i\}$. This just means we have given a "timestep" function $T : \prod_i \mathcal{X}_i \to \prod_i \mathcal{X}_i$. Now we can form a graph G with the X_i as the vertices, and with an edge $X_{j'} \to X_j$ whenever $T((x_i))_j$ depends on the value of $x_{j'}$. Note that there may be self-edges $X_i \to X_i$ in this graph. Suppose we also have given an initial value $(x_i^0) \in \prod_i \mathcal{X}_i$

Given a natural number N, we let M_N be a finite model with variables $X_{i,n}$ for each $n = 1, \dots N$ and each i. We let the edges of the graph be given by $X_{i,n} \to X_{j,n+1}$ if there is an edge $X_i \to X_j$ in G (and no other edges). The kernel $\prod_{j' \to j} \mathcal{X}_{j'} \to \mathcal{X}_j$ is given by the deterministic function $T(-)_j$ — this makes sense because, by assumption, that function only depends on the given values. The distribution on $X_{i,0}$ is given by x_i with probability 1. There is an obvious (exact) abstraction $M_N \to M_{N-1}$, which simply discards the last value. The cofiltered limit $\lim_N M_N$ is a profinite model which represents the discrete dynamical system described by this data.

5.2 Variables in a profinite model

Since a profinite model only contains very abstract information about how it transforms into finite models, it would seem very difficult to work with it in a useful way. In this section, we'll see how the flatness property allows us to treat a profinite model very much like a finite model.

Definition 5.15. A *generalized variable* of a profinite model M is an exact element $X \in M(\mathcal{X})$ for some probability space $\mathcal{X} \in \mathsf{FinProb} \subseteq \mathsf{FinMod}_{\mathrm{ex}}$. The distribution of the variable is the underlying probability space $\mathcal{X}$.

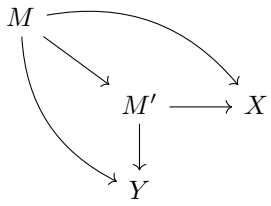
Note that all generalized variables in this sense have finite outcome space. It might be possible to define variables with values in *profinite sets* (cofiltered limits of finite sets), but we have not developed this idea further.

Example 5.16. For a finite model M, a generalized variable consists of a function $M[G] \to \mathcal{X}$ for some finite set $\mathcal{X}$. This is a random variable in the usual sense — one that only depends on the values of the model.

Since profinite models are flat, given a set of generalized variables, say X_i, we can find some finite view $M \to M'$ so that all the variables are variables in M', up to arbitrarily small error. In fact, we can do even better than this:

Lemma 5.17. Let X, Y be generalized variables of a profinite model M, valued in the sets $\mathcal{X}, \mathcal{Y}$. Let $\epsilon > 0$. Then at least one of the following is true:

1. There exists a model M' containing three variables, X, Y, C, with causal graph $C \to X \to Y, C \to X$, and a diagram

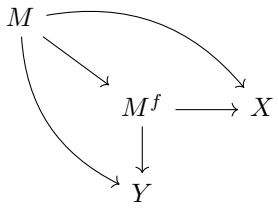


(7)

   with each map having error less than ϵ.

2. There exists a model as in case 1., but with the causal direction being $Y \to X$ instead.

Proof. First observe that there is some factorization



with M^f a finite graphical model, and all arrows having error less than ϵ. Given this, we can simply further transform M^f by aggregating all those variables that are causally upstream of X or Y into one, C, and forgetting all the others. This leads to the desired model M'. □

Note that both 1. and 2. may be true, if there is no causal relationship between the two variables. Also note that the confounding variable C does not have to actually have any effect on X or Y. The point of this proposition is that we can talk about the causal relationship between two variables in a profinite model, by placing them in an (approximate) finitary model. We can imagine C to simply contain all the information about X and Y that comes from variables upstream of them. The small ϵ of error may be necessary because there may be an infinite number of relevant confounding variables — we can approximate this with finitely many points, but not necessarily with complete precision.

In fact, we can strengthen this statement. The above statement essentially says that, given any positive bound on the error, we can fit a causal structure to X and Y within that error bound. But more is true: there exists some causal structure which can be made to fit within any error bound, and this causal structure is, in the following weak sense, unique:

Proposition 5.18. Let $X, Y, M, \mathcal{X}, \mathcal{Y}$ be as before. For any $\epsilon > 0$, we can consider the following statements

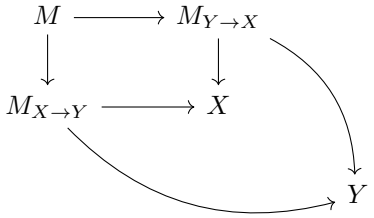
1. Statement 1. from Lemma 5.17.
2. Statement 2. from Lemma 5.17.
3. There exists a model M' with three variables X, Y, C, with causal graph $X \leftarrow C \rightarrow Y$, and a diagram as Eq. (7), with each map having error $< \epsilon$.
4. There exists a model M' with two variables X, Y, with causal graph $X \rightarrow Y$, and a diagram as Eq. (7), with each map having error $< \epsilon$.
5. As 4., but with M' having causal graph $Y \rightarrow X$.
6. As 4. and 5., but with M' having a discrete causal graph.

These statements are not mutually exclusive: For any fixed ϵ, 3 implies both 1 and 2, 4 implies 1, 5 implies 2, and 6 implies all the rest. For any model, there exists a uniuqe one of these statements, independent of ϵ, which holds for all $\epsilon > 0$, and so that every other statement which holds for all ϵ is a "formal consequence" of the first, in the sense that it is among the implications listed above.

Proof. First, let's prove that we can always find such a statement. Let $S(\epsilon) \subseteq \{1, 2, 3, 4, 5, 6, 7\}$ be the subset of numbers so that the statement holds for ϵ. Then $S(\epsilon) \subseteq S(\epsilon')$ if $\epsilon < \epsilon'$. Since each $S(\epsilon)$ is nonempty (by Lemma 5.17), this implies that $\cap_{\epsilon > 0} S(\epsilon)$ is also nonempty, which is precisely the claim. (This of course depends on the fact that there is a *finite* number of statements).

Now let's prove the "uniqueness" claim. We are claiming that if n and m both hold for all ϵ, then there is some statement which holds for all ϵ and which "formally" implies both n and m. For example, supose both 4 and 5 hold for all ϵ. Then we must show that 6 also holds for all ϵ.

By assumption, we have models, let's call them $M_{X \rightarrow Y}$ and $M_{Y \rightarrow X}$, with those causal graphs, and a commutative diagrams



where each arrow has error at most ϵ. Using the flatness property of M, we can replace M with a finite model M'. This may increase the error of the arrows, but only by an arbitrarily small amount, so that we can still have as small an error as we want.

Both X and Y are functions of some set of variables within M'. The upper path around the diagram means that intervening on a variable that Y depends on, can perturb the distribution of X by at most ϵ. Similarly, the lower path menas that intervening on X can perturb the distribution of Y by at most ϵ. This means we can construct an abstraction $M' \to M_{X \perp Y}$ which simply deletes the causal relationship between these two variables, and its error will be bounded by 2ϵ. Since ϵ was arbitrarily small, this proves the desired result.

A similar argument proves this result for the other cases. □

These seven statements are an enumeration of the possible causal structures on two variables (including a possible third confounding variable). The content of the proposition is that the causal structure is well-defined, in a suitable sense. Of course, if the two variables are causally independent, they can also be fit to the structure $X \to Y$ (with the arrow simply not being causally relevant). What the propositon says is that there is a well-defined "minimal" causal structure.

It's worth noting that results like this only work for *finite* numbers of variables. In general, asking "countably infinite questions" in a profinite model is not necessarily well-behaved — this is inherent in the definition of flat functors. In technical terms, we could ask instead that the left Kan extension $[\mathsf{FinMod}^{\mathrm{op}}, \mathsf{Err}] \to \mathsf{Err}$ preserves *countable* limits, instead of merely finite limits. This would correspond to changing the definition of "cofiltered" to require that, given a countable collection of objects x_i, there is an object y with maps $y \to x_i$ for all i.

References

[1] Francis Borceux and Carmen Quinteriro. "Enriched accessible categories". In: *Bulletin of the Australian Mathematical Society* 54.3 (1996), pp. 489–501. DOI: 10.1017/S0004972700021900.

[2] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. "Visual Causal Feature Learning". In: *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI 2015, July 12-16, 2015, Amsterdam, The Netherlands*. 2015, pp. 181–190. URL: http://auai.org/uai2015/proceedings/papers/109.pdf.

[3] D. M. Endres and J. E. Schindelin. "A new metric for probability distributions". In: *IEEE Transactions on Information Theory* 49.7 (2003), pp. 1858–1860. DOI: 10.1109/TIT.2003.813506.

[4] Brendan Fong. *Causal Theories: A Categorical Perspective on Bayesian Networks*. 2013. arXiv: 1301.6201 [math.PR].

[5] Tobias Fritz. *A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics*. 2019. arXiv: 1908.07021 [math.ST].

[6] Tobias Fritz and Paolo Perrone. *A Probability Monad as the Colimit of Spaces of Finite Samples*. 2017. arXiv: 1712.05363 [math.PR].

[7] Bart Jacobs, Aleks Kissinger, and Fabio Zanasi. "Causal Inference by String Diagram Surgery". In: *Foundations of Software Science and Computation Structures* (2019), pp. 313–329. ISSN: 1611-3349. DOI: 10.1007/978-3-030-17127-8_18.

[8] G. M. Kelly. *Basic concepts of enriched category theory*. 10. 2005.

[9] S. Kullback and R. A. Leibler. "On Information and Sufficiency". In: *Ann. Math. Statist.* 22.1 (Mar. 1951), pp. 79–86. DOI: 10.1214/aoms/1177729694.

[10] Steffen Lauritzen. *Lectures on Graphical Models*. 2019.

[11] Tom Leinster. "Higher Operads, Higher Categories". In: (2004). DOI: 10.1017/cbo9780511525896.

[12] Fosco Loregian. *Coend calculus*. 2015. arXiv: 1501.02503 [math.CT].

[13] Jacob Lurie. *Higher Topos Theory (AM-170)*. Princeton University Press, 2009. ISBN: 9780691140490. URL: https://www.math.ias.edu/~lurie/papers/HTT.pdf.

[14] David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Oct. 2003. ISBN: 0521642981. URL: http://www.inference.org.uk/mackay/itprnn/book.html.

[15] Saunders MacLane. *Categories for the Working Mathematician*. Graduate Texts in Mathematics, Vol. 5. New York: Springer-Verlag, 1971, pp. ix+262.

[16] Scott Morrison and David Penneys. "Monoidal Categories Enriched in Braided Monoidal Categories". In: *International Mathematics Research Notices* 2019.11 (Oct. 2017), pp. 3527–3579. ISSN: 1687-0247. DOI: 10.1093/imrn/rnx217.

[17] Evan Patterson. "The algebra and machine representation of statistical models". PhD thesis. Stanford University, Department of Statistics, 2020. arXiv: 2006.08945 [math.ST].

[18] Paolo Perrone. *Notes on Category Theory with examples from basic mathematics*. 2019. arXiv: 1912.10642 [math.CT].

[19] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA, USA: MIT Press, 2017.

[20] E. Riehl. *Category Theory in Context*. Aurora: Dover Modern Math Originals. Dover Publications, 2017. ISBN: 9780486820804. URL: http://www.math.jhu.edu/~eriehl/context.pdf.

[21] Paul K. Rubenstein et al. *Causal Consistency of Structural Equation Models*. 2017. arXiv: 1707.00819 [stat.ML].

[22] C Villani. "Optimal transport – Old and new". In: vol. 338. Jan. 2008. DOI: 10.1007/978-3-540-71050-9.

[23] Wikipedia contributors. *All models are wrong — Wikipedia, The Free Encyclopedia*. [Online; accessed 29-July-2020]. 2020. URL: https://en.wikipedia.org/w/index.php?title=All_models_are_wrong&oldid=966528057.

[24] Wikipedia contributors. *Jensen–Shannon divergence — Wikipedia, The Free Encyclopedia*. [Online; accessed 21-June-2020]. 2020. URL: https://en.wikipedia.org/w/index.php?title=Jensen%E2%80%93Shannon_divergence&oldid=960630930.