

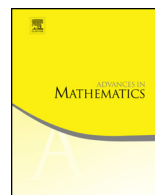


ELSEVIER

Contents lists available at ScienceDirect

Advances in Mathematics

www.elsevier.com/locate/aim



A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics

Tobias Fritz

Perimeter Institute for Theoretical Physics, Waterloo, Ontario, Canada

ARTICLE INFO

Article history:

Received 15 October 2019

Received in revised form 28 April 2020

Accepted 14 May 2020

Available online 28 May 2020

Communicated by Ross Street

MSC:

primary 60A05, 62A01

secondary 62B05, 18D10, 68Q55

Keywords:

Foundations of probability

Foundations of statistics

Monoidal categories with structure

Conditional independence

Almost surely

Sufficient statistic

ABSTRACT

We develop *Markov categories* as a framework for synthetic probability and statistics, following work of Golubtsov as well as Cho and Jacobs. This means that we treat the following concepts in purely abstract categorical terms: conditioning and disintegration; various versions of conditional independence and its standard properties; conditional products; almost surely; sufficient statistics; versions of theorems on sufficient statistics due to Fisher–Neyman, Basu, and Bahadur.

Besides the conceptual clarity offered by our categorical setup, its main advantage is that it provides a uniform treatment of various types of probability theory, including discrete probability theory, measure-theoretic probability with general measurable spaces, Gaussian probability, stochastic processes of either of these kinds, and many others.

© 2020 Elsevier Inc. All rights reserved.

Contents

1. Introduction	2
2. Markov categories	10
3. Example: Kleisli categories of monoidal monads	18
4. Example: measurable spaces and measurable Markov kernels	19
5. Example: compact Hausdorff spaces and continuous Markov kernels	25

E-mail address: tfritz@pitp.ca.<https://doi.org/10.1016/j.aim.2020.107239>

0001-8708/© 2020 Elsevier Inc. All rights reserved.

6.	Example: Gaussian probability theory	27
7.	Example: diagram categories and stochastic processes	30
8.	Example: hypergraph categories	32
9.	Example: categories of commutative comonoids	33
10.	Deterministic morphisms and a strictification theorem	33
11.	Further candidate axioms for Markov categories	43
12.	Conditional independence and the semigraphoid properties	62
13.	Almost surely	80
14.	Sufficient statistics and the Fisher–Neyman factorization theorem	90
15.	Complete morphisms, ancillary statistics, and Basu’s theorem	94
16.	Minimal sufficient statistics and Bahadur’s theorem	99
	References	101

1. Introduction

Probability theory and statistics are traditionally built on measure theory as a foundation. This has facilitated the development of a rich and diverse network of theoretical results and practical methods, comprising areas as distinct as stochastic PDEs, large deviation theory, and machine learning. In some of these areas, it is quite common to work with measure theory directly by applying its basic axioms such as σ -additivity. The position that we take in this paper is that this methodology is similar to programming a computer directly in terms of machine code. Following seminal ideas of Lawvere [79], our goal is to demonstrate that the appeal to such low-level machinery can frequently be avoided: there are simple and purely algebraic axioms for “systems of Markov kernels” which allow us to give high-level definitions of some of the basic concepts of probability and statistics, such as probability spaces, random variables, independence and conditional independence, almost surely, statistics and sufficient statistics, up to and including completeness and minimality of statistics. The definitions of these concepts also allow us to derive a surprising number of theorems on these notions purely abstractly: among other things, we prove abstract versions of the semigraphoid properties of conditional independence as well as of the Fisher–Neyman factorization theorem, Basu’s theorem and Bahadur’s theorem on sufficient statistics.

The advantages of this type of approach are various, and we now mention four points in favour of it. First, perhaps the least interesting advantage is that we achieve a novel *understanding* of the above mentioned concepts. Second, a more important point is the greater *generality* achieved through the development of such a framework: there are many different systems (categories¹) of Markov kernels which satisfy our axioms, ranging from probability theory on finite sets via Gaussian probability theory to full-fledged completely general Markov kernels between measurable spaces and even to stochastic processes, where probability theory takes place in time. Therefore our results immediately

¹ We use the words “category” and “categorical” exclusively in the sense of category theory, bearing no relation to categorical data in the sense of statistics.

specialize to different theorems in each one of these cases. There are other interesting categories which satisfy the same axioms but have little to do with probability theory at all, and again our results apply to them; a good example of such a Markov category is the category of sets and multivalued functions, where the morphisms are not Markov kernels, but display largely analogous behaviour.

Third, a related advantage of an abstract framework such as ours is that the axioms are more *high-level* than the standard measure-theoretic ones. If using the standard ones is analogous to programming a computer in machine code, then using the high-level ones is analogous to programming a computer in a language which provides higher abstraction. This enables the programmer to write larger and more complex programs whose functioning is easier to grasp for human minds. As we will see, something similar happens in our context: our proofs of theorems on sufficient statistics such as Basu's are very simple and visual. This should not be surprising, since a certain amount of proof difficulty already enters in the proofs that a given system, such as measure-theoretic probability, satisfies our axioms. Going beyond existing theorems, we expect that our abstract framework will also facilitate the development of new and more complex results in theoretical statistics. Fourth, there is a *modularity* advantage: most of our results only require a fragment of the axioms to be satisfied, so that these results can be instantiated also in Markov categories where merely this fragment holds.

The word “synthetic” in our title refers exactly to this idea of higher abstraction, suggesting that we are dealing with an instance of the *synthetic vs. analytic* dichotomy. Roughly speaking, a mathematical formalism or theory is *analytic* if it focuses on concrete and specific implementation details, often so specific that they allow the user to answer any particular question in a definite way (at least in principle). The paradigmatic example of an analytic theory is Descartes' *analytic geometry* based on coordinate computations. On the other hand, a formalism or theory is *synthetic* if it merely provides higher-level axioms or even just inference rules for reasoning about the structures of interest. Euclid's axioms for planar geometry is the analogous paradigmatic example here. And indeed, doing geometry in terms of Euclid's axioms has advantages over analytic geometry which are parallel to the ones outlined above.² In this case, the utility of these advantages has played out in a very interesting way historically, leading to the development of other analytic models of a fragment of the axioms, in the form of hyperbolic geometry and more generally Riemannian geometry.

Let us briefly discuss one more well-established analogue for the relation between our approach and the traditional measure-theoretic one. The theory of *abelian categories* is an abstraction of the theory of modules over a ring bearing some conceptual similarity to how our approach abstracts from the conventional foundations of probability and statistics. Again the theory of abelian categories has similar advantages (and disadvantages) over module theory. To see how these manifest themselves, consider the sentiment that

² Which is not to deny that analytic geometry, or analytic theories in general, have distinct advantages over synthetic ones as well.

abelian categories are in some sense the “right” level of abstraction for the development of homological algebra: the irrelevant implementation details of what the objects of an abelian category “really are” have been abstracted away, while the focus is precisely on those structures that are relevant to the development of homological algebra [58]. Moreover, the greater generality is useful in that not all applications of homological algebra are to categories of modules; this happens e.g. in algebraic geometry, where many categories of sheaves are abelian categories which are not categories of modules (not even up to equivalence) [98].

However, we certainly cannot claim that our formalism in its present form is as solid or as definite as the theory of abelian categories. Indeed, we do not expect it to have reached its final form, as some of the technical details in our definitions will almost surely be subject to revision. Nevertheless, we believe that the basic spirit will remain unchanged, and has the potential to become a new paradigm for the mathematical foundations of probability and statistics.

Summary. We now summarize the main definitions and results of the paper, one section at a time.

- ▷ Section 2 introduces Markov categories as our main protagonists (Definition 2.1), following work of Golubtsov as well as Cho and Jacobs. These are categories which axiomatize systems of Markov kernels. The axiomatized structures are sequential composition of Markov kernels, their tensor product, as well as distinguished Markov kernels which implement *copying* and *discarding* of values of random variables. These pieces of structure are subject to certain natural compatibility conditions. We discuss the significance of the axioms and introduce **FinStoch**, the category of Markov kernels (or stochastic matrices) between finite sets (Example 2.5) as our basic running example, and **FinSetMulti**, the category of finite sets and multivalued functions (Example 2.6).
- ▷ The next few sections are dedicated to examples of Markov categories. This starts with a general construction in Section 3, where we prove that the Kleisli category of a monoidal affine monad on a Markov category is again a Markov category (Corollary 3.2).
- ▷ Section 4 introduces our main examples besides **FinStoch**, namely the category of measurable spaces and Markov kernels **Stoch** and its full subcategory of standard Borel spaces **BorelStoch**. We recall how **Stoch** and **BorelStoch** arise via Corollary 3.2 as the Kleisli category of the Giry monad.
- ▷ Section 5 discusses the Kleisli category of the Radon monad, giving a Markov category again per Corollary 3.2. It contains continuous Markov kernels between compact Hausdorff spaces.
- ▷ Section 6 shows how Gaussian probability theory is also described by a Markov category denoted **Gauss**. Here, the morphisms are given by those Markov kernels

$\mathbb{R}^n \rightarrow \mathbb{R}^m$ which apply a linear transformation and add independent Gaussian noise with specified expectation and variance.

After these concrete examples, Sections 7 to 9 provide general categorical recipes for constructing Markov categories. A reader whose primary interest is in theoretical statistics may skip these without significant loss, and perhaps return to Section 7 on a second reading in order to see how our results automatically generalize to stochastic processes.

- ▷ Section 7 explains how we can construct new Markov categories by considering diagrams of a given shape in a given Markov category. This implies in particular that all of our definitions and results can be instantiated straightforwardly on stochastic processes, since these are diagrams given by infinite sequences of deterministic morphisms.
- ▷ Section 8 briefly describes the construction of Markov categories from hypergraph categories, recovering `FinStoch` and `FinSetMulti` as particular examples.
- ▷ Section 9 explains how commutative comonoids in *any* symmetric monoidal category form a Markov category with respect to the counit-preserving morphisms.

Having then treated examples of Markov categories in detail, at this point we turn to the development of probability and statistics in the abstract framework of Markov categories.

- ▷ Section 10 introduces deterministic morphisms in a Markov category as those morphisms which preserve the comultiplication (Definition 10.1). We show that in `FinStoch`, `Stoch` and `Gauss`, the deterministic morphisms are indeed exactly those which do not involve any randomness (Examples 10.3 to 10.5 and 10.7). The second half of the section is devoted to certain categorical technicalities for Markov categories: we prove that the structure morphisms are necessarily deterministic (Lemma 10.12), and show that the subcategory of deterministic morphisms is a *cartesian* monoidal subcategory (Remark 10.13). We then define the 2-category of Markov categories (Definition 10.14) and prove a strictification theorem (Theorem 10.17) which states that every Markov category is comonoid equivalent to another one in which the monoidal structure is strict.
- ▷ Section 11 introduces four possible additional axioms which a given Markov category may or may not satisfy, and which we believe to have some relevance for probability theory and statistics. The first one is the existence of *conditionals* (Definition 11.5), which we later relate to general disintegrations (Proposition 11.17). The second one is *randomness pushback*, formalizing the intuitive idea that any stochastic operation can be implemented by first generating randomness and then executing a deterministic function which uses the previously generated randomness as an additional input. Third, we discuss another property which we call *positivity* (Definition 11.22), closely related to the nonnegativity of probability, which will turn out to be useful in later proofs. Fourth, we consider a condition which we call *causality* (Definition 11.31),

expressing the idea that if a choice between two processes does not affect how their outcome is correlated with earlier outcomes, then there should not be any correlation with even earlier outcomes either.

For each of these axioms, we also discuss which ones of our particular Markov categories satisfy them. The existence of conditionals fails for *Stoch* (Examples 11.3 and 11.18) but holds in *FinStoch*, *BorelStoch* and *Gauss* (Examples 11.6 and 11.8). Randomness pushback likewise holds in *FinStoch* and *Gauss* (Examples 11.20 and 11.21), while we do not know whether it holds in *Stoch*. All of our main examples satisfy the positivity axiom (Lemma 11.24 and Example 11.25) as well as the causality axiom (Proposition 11.34 and Example 11.35).

- ▷ Section 12 treats conditional independence in our setting. We give various definitions of conditional independence for different kinds of morphisms (Definitions 12.1, 12.12, 12.16 and 12.19), including the difference on whether one wants to condition on an input or on an output (or both). We prove corresponding versions of the semigraphoid properties (Lemmas 12.5 and 12.13 and Propositions 12.17 and 12.20), and various results on how conditional independences propagate across composition of morphisms (Proposition 12.18). Our main innovation on conditional independence over previous work of Cho and Jacobs is that we do not assume the existence of conditionals.

Furthermore, in the presence of conditionals, we also develop the theory of conditional products of joint distributions, showing that these can be defined and satisfy axioms originally proposed by Dawid and Studený in any Markov category which has conditionals (Definition 12.8 and after).

- ▷ Section 13 treats almost sure equality (Definition 13.1), again following an idea of Cho and Jacobs. We prove a number of lemmas on the compositional structure of almost sure equality. We also construct the category of probability spaces and Markov kernels modulo almost sure equality in any Markov category satisfying causality (Proposition 13.9). This implies that Bayesian inversion is a symmetric monoidal dagger functor, solving an open problem posed by Clerc, Danos, Dahlqvist and Garnier (Remark 13.10). We then introduce the concept of almost sure determinism (Definition 13.11) as another example of how our setup allows for a systematic relativization of probabilistic concepts with respect to almost surely. Finally, we introduce a candidate definition of *support* of a morphism (Definition 13.20), conjecturally recovering the usual notion of support of a probability measure (Remark 13.21).
- ▷ The final three sections are devoted to the development of some of the basic notions and theorems of statistics within our formalism. This theory frequently makes use of aspects of conditional independence as studied in Section 12, in line with Dawid's arguments on the central role of conditional independence in statistics.

Section 14 starts by introducing statistical models (Definition 14.1) and statistics (Definition 14.2), leading up to the definition of sufficient statistic (Definition 14.3). This is followed by Theorem 14.5, which seems to be an abstract close relative of the Fisher–Neyman factorization theorem. This interpretation is corroborated by

showing that our result indeed induces the Fisher–Neyman factorization in the case of FinStoch (Example 14.6).

- ▷ Section 15 builds on the previous section by treating the completeness of statistics. Observing that the notion of completeness is secretly a property of Markov kernels in general (Remark 15.2), we find a simple definition of completeness within our abstract formalism (Definition 15.1). We show that it specializes to bounded completeness in the case of Stoch (Example 15.4). After defining ancillary statistics (Definition 15.7), we then prove an abstract version of Basu’s theorem (Theorem 15.8).

Theorem (Basu). *Any complete sufficient statistic for a given statistical model is independent of any ancillary statistic.*

- ▷ Section 16 goes further by introducing minimal sufficient statistics within our abstract setup (Definition 16.2), based on a preordering on the set of statistics for a given statistical model which compares statistics by their informativeness (Definition 16.1). We then state and prove an abstract version of Bahadur’s theorem.

Theorem (Bahadur). *If a minimal sufficient statistic exists for a given statistical model, then a complete sufficient statistic is minimal sufficient.*

This concludes our summary. Let us reemphasize that all of our definitions should be regarded as subject to further refinement. Of course, our results based on these preliminary definitions are nevertheless definite.

Discussion of prior work. The most important precursors to this paper are works by Golubtsov [53–56]³ as well as a recent paper by Cho and Jacobs [15]. The present paper can be regarded as a sort of follow-up to these, and we will make frequent reference to them. In particular, these prior authors already demonstrated that Markov categories⁴ can serve as a canvas for developing aspects of probability theory. Markov categories had also been used somewhat implicitly in Fong’s study of the categorical semantics of Bayesian networks [33].

However, synthetic approaches to probability and statistics in general have a long history going back further than these works, and have been pursued mostly independently in several scientific communities. We here provide a brief overview of the main literature that we are currently aware of.

- ▷ In category theory, categorical probability has been initiated by the work of Lawvere [79] and Giry [48]. It has traditionally focused on the study of *probability*

³ Due to the substantial overlap between those papers, we generally only refer to [54] from now on, which is easily digitally accessible. The references in those papers also point to earlier work of Golubtsov on particular categories of interest, where some of the relevant concepts have been studied concretely.

⁴ Or *affine CD-categories* in the terminology of Cho and Jacobs. The *categories of information transformers* or *IT-categories* in the sense of Golubtsov are technically slightly different.

monads [97, Chapter 1], providing formal descriptions of the measure-theoretic and functional-analytic aspects of probability [2,85]. This is closely related to the present paper insofar as Kleisli categories of probability monads are categories of Markov kernels, and should therefore generally be expected to be Markov categories in the sense of Definition 2.1 (see also Corollary 3.2).

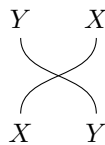
- ▷ Computer scientists have studied structural aspects of probability theory extensively, often with applications to the semantics of probabilistic programs in mind. The seminal paper of Cho and Jacobs [15] belongs to this area, as well as other work⁵ by Jacobs [16], Panangaden [3], Simpson [106], Staton [62], and others. Due to the affinity of theoretical computer scientists with category theory, there is a strong overlap here with work done by category theorists. This subject seems to be currently maturing, as witnessed by the first book-length treatment on structured probabilistic reasoning becoming available [66].
- ▷ In mathematical statistics, there is a large number of papers on its foundations, some of which have investigated approaches which are synthetic in flavour, and some of which use categorical concepts either explicitly or implicitly. For example, this includes works by Dawid on statistical operations [25], conditional independence [24, 26] and conditional products [27], by Lauritzen on combining statistical models [89] and composing Gaussian conditionals [78], as well as by McCullagh on the concept of statistic [90].
- ▷ In noncommutative probability theory, synthetic notions of independence which generalize various concrete notions of independence have been studied using categorical methods, such as free independence in free probability, for example in the works of Franz [37] as well as Gerhold, Lachs and Schürmann [47].
- ▷ In the foundations of quantum mechanics community, researchers have attempted to construct quantum analogues of Bayesian inference and Bayesian networks. This endeavour is crucially informed by having a synthetic approach to the classical case to begin with. For example, the work of Coecke and Spekkens [21] belongs here.
- ▷ In information geometry, Čencov had investigated the category *Stoch* as the “category of statistical decisions” independently of Lawvere and Giry [11,13], where it was used subsequently as a canvas for studying entropic distances on probability measures [12,91].

We have tried to integrate some discussion of what has been done in each of these areas, to the best of our knowledge, into the main text. Nevertheless, it is almost inevitable that some readers, in particular those coming from one of these particular fields, will feel that their own area remains underrepresented in our presentation. For these readers, we hope that the above list can provide some convenient entry points into the existing literature of other communities which they may not be aware of.

⁵ The references listed are but a sample of these authors’ extensive works on the subject.

Conventions and notation. Throughout, we use the term “Markov kernel” in the standard sense, referring to a map between measurable spaces which takes every element of the first space to a *probability measure* on the second space, such that a suitable measurability requirement holds; see e.g. [75, Definition 8.25] or [93, Définition III-2.1]. In other words, “Markov kernel” is more or less synonymous with other terms like “transition kernel”, “stochastic relation”, “stochastic map”, “probabilistic map”, “conditional distribution”, “channel”, “statistical operation”, “information transformer” and others (neither one of which we will use). We present a detailed technical treatment of Markov kernels in Section 4. However, precisely because our categorical formalism is more general and more abstract than the measure-theoretic one, most of our uses of the term “Markov kernel” will not require the reader to know the technical details. It is perfectly enough to associate the term “Markov kernel” merely with “noisy map”, i.e. a function whose output may be inherently random. The morphisms in those Markov categories which are of interest in probability and statistics are always some form of Markov kernels or noisy maps, even if the technical details may differ. We also think of a morphism in a general Markov category as an abstract version of a Markov kernel.

Our categorical setup is as follows. Throughout, $\mathcal{C}$ is a symmetric monoidal category with monoidal unit $I \in \mathcal{C}$ and symmetry morphisms $\text{swap}_{X,Y} : X \otimes Y \rightarrow Y \otimes X$. We freely use *string diagrams* as the usual graphical calculus for symmetric monoidal categories [104], as these are more intuitive and easier to parse for a human reader than long equations. Our convention is that these string diagrams are to be read from bottom to top: the domain of a composite morphism defined by a string diagram is at the bottom while the codomain is on top. Hence our string diagrams can be interpreted as processes with time flowing upwards. For example, (2.2) should be interpreted as stating that creating three copies of an object can be achieved either by first copying and then copying the second copy again, or first copying and then copying the first copy again. The symmetry morphisms $\text{swap}_{X,Y}$ are depicted graphically simply as crossings,



While we usually leave out the object labels on the wires of our string diagrams, we occasionally decorate the overall input and output wires by the corresponding object labels for clarity and emphasis.

While the possibility of using string diagrams is very convenient, we do not regard it as a central selling point of our framework. Rather, our main innovation lies in the greater generality and conceptual clarity facilitated by a categorical treatment, independently of which notation one uses for categorical reasoning.

Acknowledgments. We thank Jürgen Jost and Hông Vân Lê for the discussions which spurred this paper; Manuel Bärenz, Kenta Cho, Tomáš Gonda, Bart Jacobs, Dominique

Pastor, Paolo Perrone, Eigil Fjeldgren Rischel, Alex Simpson and David Spivak for useful discussions and detailed feedback on a draft; Brendan Fong, TC Fraser, Richard Garner, Paul Marriott, Arthur Parzygnat, Rob Spekkens, Sam Staton and Dario Stein for further helpful discussions; Sam Tenka for pointing out Basu’s theorem and Bahadur’s theorem to us; as well as Bob Coecke, Philip Dawid, Robert Furber, Peter Golubtsov, Alexander Kurz, Steffen Lauritzen, Nina Otter and MathOverflow user Steve for additional valuable pointers to the literature. Last but not least, we thank the creators of and contributors to the [nLab](#) and [TikZiT](#) for having made research and dissemination in category theory so much easier.

Part of this work was conducted while the author was with the Max Planck Institute for Mathematics in the Sciences.

2. Markov categories

Here, we introduce our main objects of study: *Markov categories*, which are symmetric monoidal categories with extra structure which makes them behave (in many ways) like categories of Markov kernels. As far as we know, these kinds of categories have first appeared in slightly different form⁶ in work of Golubtsov [54, Sections 2 and 3]. In the present form they first played a major role in Fong’s work on the categorical structure of Bayesian networks [33], even if somewhat implicitly. They crop up similarly implicitly in later work by Jacobs and Zanasi [68], and have subsequently been axiomatized explicitly by Cho and Jacobs in their categorical approach to conditional independence and Bayesian inference [15]. Cho and Jacobs call them *affine CD-categories*, where “CD” stands for “Copy/Discard”, describing the interpretation of the structure morphisms (2.1). Due to the central role that these categories seem to play in probability and statistics, we introduce a catchier term which hints at the idea that the morphisms in the categories under consideration behave like Markov kernels.

2.1. Definition. A *Markov category* $\mathbf{C}$ is a symmetric monoidal category in which every object $X \in \mathbf{C}$ is equipped with a commutative comonoid structure given by a comultiplication $\text{copy}_X : X \rightarrow X \otimes X$ and a counit $\text{del}_X : X \rightarrow I$, depicted in string diagrams as

$$\text{copy}_X = \begin{array}{c} \text{---} \text{---} \text{---} \\ \backslash \quad / \\ \bullet \\ / \quad \backslash \\ \text{---} \text{---} \end{array} \qquad \text{del}_X = \begin{array}{c} \bullet \\ | \\ \text{---} \end{array} \qquad (2.1)$$

and satisfying the commutative comonoid equations,

⁶ See Remark 10.19 for a discussion of the difference.

$$\begin{array}{c} \text{U-shape with dot on top} \\ \text{U-shape with dot on bottom} \end{array} = \begin{array}{c} \text{U-shape with dot on top} \\ \text{U-shape with dot on bottom} \end{array} \quad (2.2)$$

$$\begin{array}{c} \text{U-shape with dot on top} \\ \text{U-shape with dot on bottom} \end{array} = \begin{array}{c} | \\ | \\ | \end{array} = \begin{array}{c} \text{U-shape with dot on top} \\ \text{U-shape with dot on bottom} \end{array} \quad \begin{array}{c} \text{X-shape} \\ \text{U-shape with dot on bottom} \end{array} = \begin{array}{c} \text{U-shape with dot on bottom} \end{array} \quad (2.3)$$

as well as compatibility with the monoidal structure,

$$\begin{array}{c} \text{dot} \\ | \\ X \otimes Y \end{array} = \begin{array}{c} \text{dot} \\ | \\ X \end{array} \quad \begin{array}{c} \text{dot} \\ | \\ Y \end{array} \quad \begin{array}{c} X \otimes Y \quad X \otimes Y \\ | \quad | \\ \text{U-shape with dot on bottom} \\ | \\ X \otimes Y \end{array} = \begin{array}{c} X \quad Y \quad X \quad Y \\ | \quad | \quad | \quad | \\ \text{U-shape with dot on bottom} \quad \text{X-shape} \quad \text{U-shape with dot on bottom} \\ | \quad \quad \quad | \\ X \quad \quad \quad Y \end{array} \quad (2.4)$$

and naturality of del , which means that

$$\begin{array}{c} \text{dot} \\ | \\ \boxed{f} \end{array} = \begin{array}{c} \text{dot} \\ | \end{array} \quad (2.5)$$

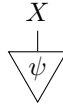
for every morphism f .

We usually think of a morphism in $\mathbf{C}$ as some form of Markov kernel, meaning that it assigns to every input value a probability distribution over output values. The comultiplication copy_X then represents the map which takes an input value, copies it, and outputs the two copies without introducing any randomness; while del_X is the map which simply discards its input and produces no output. The above equations then all acquire an intuitive meaning, such as that copying and discarding one copy is equal to simply outputting the input, a property sometimes known as *broadcasting* [4].

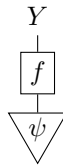
The Markov kernel interpretation, including the role of copy and del , will be made precise in Example 2.5 and Sections 4 and 5 by constructing actual categories of Markov kernels which are Markov categories. The main purpose of our paper is to demonstrate that these pieces of structure are enough to develop a substantial amount of probability theory and statistics.

The first concept from probability theory which acquires meaning in any Markov category is that of a probability measure, or *distribution*; we prefer the latter term due

to its more open-ended connotation. A distribution is a Markov kernel which takes no input and produces an output, interpreted as a “random element” of the codomain. This motivates the definition that a *distribution* in any Markov category $\mathbf{C}$ is a morphism $\psi : I \rightarrow X$ from the unit object I to some other object X , depicted in string diagram notation as



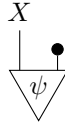
In the Markov category **Stoch**, our example from Section 4, these are precisely the probability measures on measurable spaces X . We can now say that the pair (X, ψ) is a *probability space* in $\mathbf{C}$; in **Stoch**, this indeed specializes precisely to the traditional notion of probability space. A *random variable* on (X, ψ) taking values in some other object Y can then be defined to be a morphism $f : X \rightarrow Y$, where one may want to impose the additional requirement that f must be deterministic (Definition 10.1), and perhaps identify two such morphisms as representing the same random variable as soon as they are ψ -almost surely equal (Definition 13.1).⁷ The distribution of the random variable f is then given by the composition $f\psi : I \rightarrow Y$,



so that the pair $(Y, f\psi)$ is also a probability space in $\mathbf{C}$. This string diagram can be interpreted conveniently in terms of processes happening in time: first, ψ generates a random element of X ; second, this element is fed into the function f , which subsequently outputs an element of Y . The overall effect is that we have produced a random element of Y , corresponding to the distribution of our random variable. One takeaway of this is that the formalism of Markov categories unifies the concepts of probability distributions and random variables by considering both of them as instances of the more general concept of Markov kernel.

For objects $X, Y \in \mathbf{C}$, morphisms $\psi : I \rightarrow X \otimes Y$ correspond to *joint distributions*. In the case of **Stoch**, these are exactly the probability measures on the product σ -algebra on the product of measurable spaces X and Y . In general, such a joint distribution can be *marginalized* over Y by composing with $\text{id}_X \otimes \text{del}_Y$. In terms of string diagrams, the resulting composite morphism $(\text{id}_X \otimes \text{del}_Y) \circ \psi$ can be conveniently depicted as

⁷ Since we will have no need for the term “random variable” in this paper, we do not commit ourselves to a particular definition of this concept yet, and mention it merely to illustrate how to work with Markov categories. But see Remark 13.14 for further treatment of probability spaces in Markov categories.



and similarly for marginalization over Y . Analogous statements apply to joint distributions on more than two factors. Of course, these interpretations are also central to the work of Cho and Jacobs [15], and have also been used prior to that e.g. in the work of Coecke and Spekkens [21].

2.2. Remark. Written out in a way which explicitly keeps track of the structure isomorphisms in $\mathbb{C}$, the first condition in (2.4) amounts to the commutativity of the diagram

$$\begin{array}{ccc} X \otimes Y & \xrightarrow{\text{del}_{X \otimes Y}} & I \\ \parallel & & \downarrow \cong \\ X \otimes Y & \xrightarrow{\text{del}_X \otimes \text{id}} I \otimes Y \xrightarrow{\text{id} \otimes \text{del}_Y} & I \otimes I \end{array} \quad (2.6)$$

and similarly the second one is

$$\begin{array}{ccc} X \otimes Y & \xrightarrow{\text{copy}_{X \otimes Y}} & (X \otimes Y) \otimes (X \otimes Y) \\ \parallel & & \downarrow \cong \\ & & (X \otimes (Y \otimes X)) \otimes Y \\ & & \downarrow (\text{id} \otimes \text{swap}) \otimes \text{id} \\ & & (X \otimes (X \otimes Y)) \otimes Y \\ & & \downarrow \cong \\ X \otimes Y & \xrightarrow{\text{copy}_X \otimes \text{id}} (X \otimes X) \otimes Y \xrightarrow{\text{id} \otimes \text{copy}_Y} & (X \otimes X) \otimes (Y \otimes Y) \end{array} \quad (2.7)$$

where the unnamed isomorphisms are the monoidal unitors and composite associators.

2.3. Remark. One may also think that we should impose $\text{del}_I = \text{copy}_I = \text{id}_I$, where I is the monoidal unit. These constraints are automatic if $\mathbb{C}$ is strict, because then we have

$$\text{del}_I \otimes \text{del}_I = \text{del}_I, \quad \text{copy}_I \otimes \text{copy}_I = \text{copy}_I,$$

by assumption, as well as

$$\text{del}_I \otimes \text{copy}_I = \text{del}_I \circ \text{copy}_I = \text{id}_I$$

in the commutative endomorphism monoid $\mathbb{C}(I, I)$. These properties imply

$$\mathrm{del}_I = \mathrm{del}_I \otimes \mathrm{del}_I \otimes \mathrm{copy}_I = \mathrm{del}_I \otimes \mathrm{copy}_I = \mathrm{id}_I,$$

and hence also $\mathrm{copy}_I = \mathrm{id}_I$.

If $\mathbf{C}$ is not strict, then the equation $\mathrm{copy}_I = \mathrm{id}_I$ is not actually well-typed. Rather, the correct statement is that $\mathrm{del}_I = \mathrm{id}_I$ and that $\mathrm{copy}_I : I \rightarrow I \otimes I$ is equal to the unique coherence isomorphism. The proof is similar.

By $\mathrm{del}_I = \mathrm{id}_I$ and (2.5), we can conclude that I is a terminal object. Put differently, our postulated properties of del can equivalently be phrased as saying that $\mathbf{C}$ should be a *semicartesian* symmetric monoidal category, for which there are several equivalent characterizations (see [82] and [47, Theorem 3.5]), with the simplest one being that I should be terminal⁸ [65, Definition 2.1(ii)], another one being the notion of *monoidal category with projections* as studied by Franz with a minor variation [37, Definition 3.3].⁹ In particular, the first equation in (2.4) can also be seen as a consequence of the terminality of I .

2.4. Remark. It is well-known that if $\mathbf{C}$ is cartesian monoidal, then there is a unique comonoid structure on every object. This makes $\mathbf{C}$ into a Markov category. However, the interesting Markov categories are those not of this form.

While we will consider a substantial range of particular Markov categories in Sections 3 to 9, we now introduce two basic examples, which we start with due to the technical simplicity of working with finite sets. We will use the first one throughout as our primary running example.

2.5. Example. $\mathbf{C}$ may be $\mathbf{FinStoch}$, the category of finite sets with Markov kernels. This means that a morphism $f : X \rightarrow Y$ between finite sets¹⁰ X and Y is a stochastic matrix

⁸ This condition has also been considered in [15,18,73]. However, their claimed equivalence with (2.5) is missing the assumption $\mathrm{del}_I = \mathrm{id}_I$.

⁹ The minor variation is that in order for Franz's definition to be equivalent to semicartesianness, one needs to amend it by the extra condition that at least one of the two projections $\pi_1, \pi_2 : I \otimes I \rightarrow I$ must coincide with the structure isomorphism $I \otimes I \xrightarrow{\cong} I$, as per [83] and [47, Theorem 3.5]. If π_1 does, then any $f : X \rightarrow I$ is equal to the composite $X \xrightarrow{\cong} I \otimes X \xrightarrow{\pi_1} I$ due to the diagram

$$\begin{array}{ccccc} X & \xrightarrow{\cong} & I \otimes X & \xrightarrow{\pi_1} & I \\ \downarrow f & & \downarrow \mathrm{id} \otimes f & & \parallel \\ I & \xrightarrow{\cong} & I \otimes I & \xrightarrow{\cong} & I \end{array}$$

where the right-hand square is a naturality square of π_1 . In other words, I is terminal. Conversely, it is easy to see that if I is terminal, then $\mathbf{C}$ is a monoidal category with projections satisfying the extra condition.

¹⁰ The restriction to finite sets is not essential and can easily be lifted. The most general Markov category relevant for discrete probability would be the one where the objects are sets (of arbitrary cardinality) and a morphism $f : X \rightarrow Y$ is a matrix $(f_{xy})_{x \in X, y \in Y}$ with nonnegative real entries such that $\sum_{y \in Y} f_{xy} = 1$ for every $x \in X$. In particular, for every x there are at most countably many y with $f_{xy} > 0$. We prefer to use $\mathbf{FinStoch}$ as our version of discrete probability because of its simplicity.

$(f_{xy})_{x \in X, y \in Y}$,¹¹ where the entry f_{xy} is interpreted as the probability that the Markov kernel f returns output y on input x . Composition of morphisms is given by matrix multiplication. In the following, we also use the suggestive notation $f(y|x)$ for such a morphism $f : X \rightarrow Y$, so that composition of f with some $g : Y \rightarrow Z$ turns into the Chapman–Kolmogorov equation,

$$(gf)(z|x) = \sum_y g(z|y)f(y|x). \quad (2.8)$$

There is a simple inclusion functor $\mathbf{FinSet} \rightarrow \mathbf{FinStoch}$, which takes an honest function between finite sets $f : X \rightarrow Y$ and turns it into a stochastic matrix: writing f for both by abuse of notation, we have

$$f(y|x) := \begin{cases} 1 & \text{if } y = f(x), \\ 0 & \text{otherwise.} \end{cases}$$

The symmetric monoidal structure on $\mathbf{FinStoch}$ is given by cartesian product at the level of objects, and the tensor product of stochastic matrices at the level of morphisms: for $f : A \rightarrow X$ and $g : B \rightarrow Y$,

$$(f \otimes g)(xy|ab) := f(x|a)g(y|b), \quad (2.9)$$

with the obvious structure morphisms, inherited from the symmetric monoidal inclusion functor $\mathbf{FinSet} \rightarrow \mathbf{FinStoch}$ based on the cartesian monoidal structure on $\mathbf{FinSet}$. The bifactoriality of this tensor product is easy to verify.

Finally, the comonoid operations are also inherited from $\mathbf{FinSet}$, which is a Markov category by virtue of being cartesian monoidal. More concretely, $\mathbf{del}_X : X \rightarrow I$ is the all-ones row vector indexed by the elements of X , which we formally write in the above probability notation as

$$\mathbf{del}_X(|x) = 1,$$

where the first argument is empty since $\mathbf{del}_X$ is a morphism with trivial output. $\mathbf{copy}_X$ is the stochastic matrix with entries

$$\mathbf{copy}_X(x_1, x_2|x_0) := \begin{cases} 1 & \text{if } x_1 = x_2 = x_0, \\ 0 & \text{otherwise.} \end{cases}$$

While it is routine to verify the relevant equations from Definition 2.1 directly, a more elegant and abstract argument is to use the fact that these equations are automatic in

¹¹ Recall that this means that $f_{xy} \geq 0$ and $\sum_y f_{xy} = 1$. While the primary case of interest is in matrices with nonnegative real entries, one can in principle work with nonnegative entries in any ordered field, and all results proven in this paper still hold just the same with the same proofs.

$\mathbf{FinSet}$, where the monoidal structure is cartesian, and then transfer to $\mathbf{FinStoch}$ by the functoriality and symmetric monoidality of the inclusion $\mathbf{FinSet} \rightarrow \mathbf{FinStoch}$. This works for all conditions except (2.5), which is obvious.

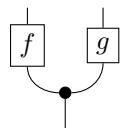
It may be instructive to see explicitly why the monoidal structure is not cartesian. A morphism $\psi : I \rightarrow X \otimes Y$ is a joint distribution, whose composition with the two projections $X \otimes Y \rightarrow X$ and $X \otimes Y \rightarrow Y$ gives the two marginal distributions $I \rightarrow X$ and $I \rightarrow Y$. Now if $\mathbf{FinStoch}$ was cartesian, the joints $I \rightarrow X \otimes Y$ would have to be in bijection with the pairs of marginals $I \rightarrow X$ and $I \rightarrow Y$. Thus the fact that a joint distribution is usually not uniquely specified by its marginals implies that $\mathbf{FinStoch}$ is not cartesian. More generally, this is the reason for why cartesian monoidal categories are not interesting as Markov categories.

2.6. Example (See also [52] and [54, Section 8.4]). $\mathbf{C}$ may be $\mathbf{FinSetMulti}$, the category of finite sets with multivalued functions, defined in more detail as follows. A morphism $f : X \rightarrow Y$ between sets X and Y is a function from X to the set of nonempty subsets of Y , or equivalently a matrix $(f(y|x))_{x \in X, y \in Y}$ with entries in $\{0, 1\}$ such that for every $x \in X$ there is $y \in Y$ with $f(y|x) = 1$, where again we use notation suggestive of conditional probabilities as in Example 2.5. We think of the statement $f(y|x) = 1$ as saying that the output y is *possible* given the input x , without specifying a further quantification of this possibility, while $f(y|x) = 0$ states that the output y is *impossible* on input x . The extra condition is then that for every input, at least one output is possible. It is in this sense that we are dealing with multivalued functions.

The definition of composition is then just the usual composition of relations: for $f : X \rightarrow Y$ and $g : Y \rightarrow Z$, we put $(gf)(z|x) := 1$ if and only if there is $y \in Y$ such that $g(z|y) = 1$ and $f(y|x) = 1$, meaning that an output z is possible for gf if and only if an output y is possible for f such that the output z is possible for g on input y . Using the convention $1 + 1 := 1$, this composition operation again takes the form of the Chapman–Kolmogorov equation (2.8). The tensor product of morphisms is also as in (2.9), making an output xy possible on input ab if x is possible for f on a and y is possible for g on b . The comonoid morphisms $\mathbf{del}$ and $\mathbf{copy}$ take the exact same form as in $\mathbf{FinStoch}$.

Although $\mathbf{FinSetMulti}$ looks like an interesting example for the concepts and results developed in the rest of this paper, we will leave its more detailed investigation in our context to future work.

2.7. Remark. Given morphisms $f : X \rightarrow Y$ and $g : X \rightarrow Z$ in a Markov category, we can form a morphism $X \rightarrow Y \otimes Z$ given by


(2.10)

It is not hard to see that this equips the coslice category $X/\mathbf{C}$ with a monoidal structure, where the monoidal unit is given by del_X . This hints at a possible relation between Markov categories and monoidal fibrations [105] by virtue of equipping the domain fibration $\mathrm{dom} : \mathbf{C}^\rightarrow \rightarrow \mathbf{C}$ with a monoidal structure, and possibly hints at an equivalent but more abstract definition of Markov categories. For a more “pedestrian” definition of a categorical structure very similar to Markov categories based on this operation, see [54, Definition 1].

In categories of Markov kernels like $\mathbf{FinStoch}$, the monoidal structure on the coslices $X/\mathbf{C}$ is closely related to *conditional products* of probability distributions, which we will investigate in Definition 12.8.

2.8. Notation. Although we will do this only with explicit mention, it can sometimes be convenient to use notation like “ $f(x, y|a, b)$ ” to indicate codomain and codomain of a morphism $f : A \otimes B \rightarrow X \otimes Y$ in a *general* Markov category. This generalizes the usual notation for $\mathbf{FinStoch}$ from Example 2.5. We can then use the Chapman–Kolmogorov equation in the form (2.8) as a general *formal* notation for composition of morphisms, irrespectively of whether such composition involves any sort of summation in the category under consideration. We can similarly use (2.9) as a general formal notation for the tensor product of morphisms.

Given a morphism $f : A \rightarrow X \otimes Y$, composing with $\mathrm{del}_X \otimes \mathrm{id} : X \otimes Y \rightarrow Y$ implements *marginalization* over the first variable. Just as we can denote the original morphism by $f(x, y|a)$, we can simply write $f(y|a)$ for this marginal, and likewise for the similarly defined other marginal $f(x|a)$. In line with our formal probability notation, the defining equation of marginalization

$$f(y|a) = \sum_x f(x, y|a)$$

thereby acquires a completely formal meaning valid in any Markov category, where the sum over x corresponds to the composition in $(\mathrm{del}_X \otimes \mathrm{id}) \circ f$. The comonoid maps del_X and copy_X do not appear explicitly in this notation: del_X is denoted simply by the empty symbol, indicating no dependence on x ; while using copy_X amounts to using the same variable x in an expression twice. For example, the composition of the tensor product of $f : X \rightarrow Y$ and $g : X \rightarrow Z$ with copy_X as in (2.10) will be the morphism denoted simply by $f(y|x)g(z|x)$.

Despite the apparent utility and generality of this notation, we will leave a complete formalization of it to future work, and will use it in this paper only occasionally and with explicit mention. It may be interesting to note that it is similar to the *abstract index notation* used in differential geometry and general relativity [113].

2.9. Notation. As is commonly done with monoid or comonoid structures in the context of string diagrams, we also draw a bullet with three wires coming out



as a shorthand for either side of the associativity equation (2.2), to be interpreted as copying with three outputs. And similarly for more than three outgoing wires, which is easily seen to be unambiguous by repeated application of associativity.

3. Example: Kleisli categories of monoidal monads

In this section, we give a general construction of Markov categories as Kleisli categories of symmetric monoidal affine monads. Subsequently, we will sketch the application of this construction to the Giry monad and to the Radon monad. This implies that all results from our later sections can be instantiated beyond the discrete case implemented by FinStoch, namely in particular in the category of measurable spaces and measurable Markov kernels (Section 4) and in the category of compact Hausdorff spaces and continuous Markov kernels (Section 5). We expect that an analogous straightforward development will be possible also in other cases, such as for the monad of Scott-continuous probability valuations on the category of (all) topological spaces [61, Section 10] (see also [57,42]).

To begin the general construction, let $\mathcal{D}$ be a symmetric monoidal category. In the examples of interest, $\mathcal{D}$ will typically be cartesian monoidal, but we do not need this assumption. Further, let $T : \mathcal{D} \rightarrow \mathcal{D}$ be a monad with multiplication $\mu : TT \rightarrow T$ and unit $\eta : 1 \rightarrow T$. It is a *symmetric monoidal monad* (e.g. [59,103]) if it comes equipped with morphisms

$$\nabla_{X,Y} : TX \otimes TY \longrightarrow T(X \otimes Y)$$

natural in X and Y , which make T into a lax symmetric monoidal functor with unit $\eta_I : I \rightarrow TI$, and such that μ and η are monoidal transformations. Concretely, the latter means that the diagrams

$$\begin{array}{ccc} TTX \otimes TTY & \xrightarrow{\nabla} & T(TX \otimes TY) \xrightarrow{T\nabla} TT(X \otimes Y) \\ \downarrow \mu \otimes \mu & & \downarrow \mu \\ TX \otimes TY & \xrightarrow{\nabla} & T(X \otimes Y) \end{array} \quad (3.1)$$

$$\begin{array}{ccc} & X \otimes Y & \\ \eta \otimes \eta \swarrow & & \searrow \eta \\ TX \otimes TY & \xrightarrow{\nabla} & T(X \otimes Y) \end{array} \quad (3.2)$$

must commute for all $X, Y \in \mathcal{D}$. We then have the following standard result [59, Corollaire 7], [103, Proposition 1.2.2].

3.1. Proposition. *Under these assumptions, the Kleisli category $\mathbf{Kl}(T)$ becomes symmetric monoidal, with:*

- ▷ the tensor product of objects being the one of $\mathbf{D}$,
- ▷ the tensor product of morphisms represented by $f : A \rightarrow TX$ and $g : B \rightarrow TY$ being represented by the composite

$$A \otimes B \xrightarrow{f \otimes g} TX \otimes TY \xrightarrow{\nabla} T(X \otimes Y).$$

Moreover, the standard inclusion functor $\mathbf{D} \rightarrow \mathbf{Kl}(T)$ is strict symmetric monoidal.

Thanks to the final statement, we can also transport comonoid structures along the inclusion functor: if an object $X \in \mathbf{D}$ carries a distinguished comonoid structure, then so does the corresponding object $X \in \mathbf{Kl}(T)$. Moreover, if the monoidal unit I is terminal in $\mathbf{D}$ and $TI \cong I$ (such a monad is also called *affine*), then I is terminal in $\mathbf{Kl}(T)$ as well. Thus we have the following:

3.2. Corollary. *Let $\mathbf{D}$ be a Markov category and T a symmetric monoidal affine monad on $\mathbf{D}$. Then the Kleisli category $\mathbf{Kl}(T)$ is again a Markov category in a canonical way.*

Concretely, the comonoid multiplication of an object $X \in \mathbf{Kl}(T)$ is represented by the composite morphism

$$X \xrightarrow{\text{copy}_X} X \otimes X \xrightarrow{\eta} T(X \otimes X).$$

In the case where $\mathbf{D}$ is cartesian monoidal, Corollary 3.2 has essentially been given by Golubtsov [54, Theorem 2] and has also been hinted at by Cho and Jacobs [15, p. 6/7]. For example, the submonad of normalized measures of a *measure category* in the sense of [102, Definition 4.2] is one context in which the result applies.

3.3. Example. For $\mathbf{D} := \mathbf{FinSet}$ considered as a cartesian monoidal category, let T be the covariant nonempty powerset monad, whose underlying functor is $TX := 2^X \setminus \{\emptyset\}$. Then T comes equipped with a canonical symmetric monoidal structure $\nabla_{X,Y}$ which takes a subset of X and subset of Y and maps it to their cartesian product, which is a subset of $X \times Y$. Then Corollary 3.2 applies, and the resulting Kleisli category $\mathbf{Kl}(T)$ is exactly the category of finite sets and multivalued functions from Example 2.6.

4. Example: measurable spaces and measurable Markov kernels

In this section, we introduce **Stoch**, the most general category of measurable Markov kernels between measurable spaces, as considered as a running example in the later sections, as well as its more well-behaved subcategory **BorelStoch** consisting of standard

Borel spaces. **Stoch** contains as particular subcategories also many other Markov categories that may be of interest to probabilists and statisticians; see e.g. Proposition 6.1. It has previously been considered as a Markov category in [15, Example 2.5] and with minor variations prior to that in [54, Section 8.1].

Roughly speaking, **Stoch** is the generalization of **FinStoch** from Example 2.5 beyond the finite case, facilitating the treatment of Markov kernels between arbitrary measurable spaces. This category can be considered to be *the* paradigmatic example of a Markov category to which our formalism applies. **Stoch** can be obtained by applying the construction of Section 3 to the *Giry monad*. But let us start with a brief outline of the illustrious history of this category, or rather that part of its history that we know of.

It seems that the category **Stoch** was first considered by Lawvere [79] in 1962, before Kleisli categories had been introduced in general [74]. Lawvere used these ideas to develop aspects of a formal theory of stochastic processes, and seems to have envisioned synthetic probability theory, as partly developed in this paper, more generally. In 1965, and apparently independently of Lawvere's work, the category **Stoch** was also considered by Čencov as the *category of statistical decision rules* [11] in the context of information geometry, and Čencov derived another description of this category [13, Theorem 5.2] reminiscent of what we now know as the Kleisli category construction [69]. In 1966, a closely related category, in fact containing **Stoch** as a full subcategory, was introduced by Morse and Sacksteder [92], again apparently independently of Lawvere and Čencov. The construction of **Stoch** as a Kleisli category was made explicit in 1982 by Giry [48, Section 4], using what we now call the *Giry monad* based only on Lawvere's ideas. A more contemporary expository account can be found e.g. in Panangaden's book on labelled Markov processes [94, Chapter 5].¹² We will follow his account by spelling out directly what the Kleisli category is.

To begin, let **Meas** be the category of measurable spaces (X, Σ_X) with measurable maps as morphisms. **Meas** becomes symmetric monoidal upon equipping it with the usual product of measurable spaces with product σ -algebras,

$$(X, \Sigma_X) \otimes (Y, \Sigma_Y) := (X \times Y, \Sigma_X \otimes \Sigma_Y).$$

The Giry monad functor $P : \mathbf{Meas} \rightarrow \mathbf{Meas}$ assigns to every measurable space (X, Σ_X) the set of probability measures $P(X, \Sigma_X)$, equipped with the smallest σ -algebra which makes the evaluation map

$$P(X, \Sigma_X) \longrightarrow [0, 1], \quad \mu \longmapsto \mu(S)$$

measurable for every $S \in \Sigma_X$. It acts on measurable maps by taking every probability measure to its pushforward, which is easily seen to be a measurable operation itself.

¹² Panangaden works with *subprobability* measures, meaning measures normalized to ≤ 1 . This does not change anything substantially.

We omit the construction of the unit and the multiplication of the Girly monad, referring to [48,2] for the details. Instead, we spell out directly what the Kleisli category $\text{Stoch} := \text{Kl}(P)$ looks like. The objects are again measurable spaces. A morphism $(X, \Sigma_X) \rightarrow (Y, \Sigma_Y)$ is given by a map

$$f : \Sigma_Y \times X \longrightarrow [0, 1], \quad (S, x) \longmapsto f(S|x),$$

having the property that every $f(-|x) : \Sigma_Y \rightarrow [0, 1]$ is a probability measure, and such that $f(S|-) : X \rightarrow [0, 1]$ is measurable for every $S \in \Sigma_Y$. The vertical bar reminds us that we think of f as a specification of probability measures conditional on any given $x \in X$. Composition with an analogous $g : \Sigma_Z \times Y \rightarrow [0, 1]$ is given by a version of the Chapman–Kolmogorov equation (2.8),

$$g \circ f : \Sigma_Z \times X \longrightarrow [0, 1], \quad (g \circ f)(T|x) := \int_{y \in Y} g(T|y) f(dy|x). \quad (4.1)$$

Here, the integral exists since $g(T|-)$ is measurable by assumption, and trivially bounded since it is $[0, 1]$ -valued. The σ -additivity in T follows from the σ -additivity of $g(-|y)$ by the dominated convergence theorem. The measurability in x for fixed T holds by the measurability assumption on f whenever g is a simple function; the definition of the Lebesgue integral then implies the general case, using the fact that the limit of a sequence of measurable functions is again measurable. For the proof that this indeed gives a category, see e.g. [94, Proposition 5.2]. Identity morphisms are given by

$$\text{id}_X(S|x) = \begin{cases} 1 & \text{if } x \in S, \\ 0 & \text{otherwise.} \end{cases}$$

In order to make $\text{Stoch} = \text{Kl}(P)$ into a Markov category [15, Example 2.5], we use Corollary 3.2 and make P into a symmetric monoidal monad via the maps

$$P(X, \Sigma_X) \otimes P(Y, \Sigma_Y) \longrightarrow P((X, \Sigma_X) \otimes (Y, \Sigma_Y)), \quad (\mu, \nu) \longmapsto \mu \otimes \nu. \quad (4.2)$$

As usual in categorical probability, these maps implement the formation of product measures $\mu \otimes \nu$ [41].

4.1. Lemma. *These maps make the Girly monad into an affine symmetric monoidal monad.*

Proof. It is clear that the Girly monad is affine. It is also straightforward to check that the maps (4.2) are natural in X and Y , as well as that all the equations hold that make P into a symmetric monoidal monad. The only problem is that it is not so clear whether (4.2) is a morphism of Meas in the first place, i.e. whether it is measurable. We now focus on proving this.

Since the σ -algebra on $P((X, \Sigma_X) \otimes (Y, \Sigma_Y))$ is the smallest one which makes the evaluation map on every $S \in \Sigma_X \otimes \Sigma_Y$ measurable, it is enough to show that the composite

$$P(X, \Sigma_X) \otimes P(Y, \Sigma_Y) \longrightarrow P((X, \Sigma_X) \otimes (Y, \Sigma_Y)) \xrightarrow{\text{ev}_S} [0, 1]$$

is jointly measurable for each fixed S . Concretely, this map is given by

$$P(X, \Sigma_X) \otimes P(Y, \Sigma_Y) \longrightarrow [0, 1], \quad (\mu, \nu) \longmapsto (\mu \otimes \nu)(S).$$

We prove that it is measurable using a standard application of the π - λ -theorem. Namely, consider the collection of all measurable S which make this map measurable. Since limits of pointwise convergent sequences of measurable maps are again measurable, this set system is closed under countable disjoint unions; and it is clearly closed under complements, so that we have a λ -system. On the other hand, $\Sigma_X \otimes \Sigma_Y$ is itself the σ -algebra generated by the rectangle sets $T \times U$ with $T \in \Sigma_X$ and $U \in \Sigma_Y$, which itself form a π -system. Thus by the π - λ -theorem, it is now enough to prove the claim in the case $S = T \times U$, i.e. to show that the map

$$P(X, \Sigma_X) \otimes P(Y, \Sigma_Y) \longrightarrow [0, 1], \quad (\mu, \nu) \longmapsto \mu(T)\nu(U)$$

is measurable for every $T \in \Sigma_X$ and $U \in \Sigma_Y$. But now this is because this map is equal to the composite

$$P(X, \Sigma_X) \otimes P(Y, \Sigma_Y) \xrightarrow{\text{ev}_T \otimes \text{ev}_U} [0, 1] \otimes [0, 1] \longrightarrow [0, 1], \quad (4.3)$$

where the second map is just multiplication, which is clearly measurable. $\square$

Concretely, the comonoid maps in **Stoch** are given by

$$\text{copy}_{(X, \Sigma_X)} : (\Sigma_X \otimes \Sigma_X) \times X \longrightarrow [0, 1], \quad (S \times T, x) \longmapsto \begin{cases} 1 & \text{if } x \in A \cap B, \\ 0 & \text{otherwise,} \end{cases}$$

for any generating rectangle $A \times B$ in the product σ -algebra. Indeed this corresponds to assigning to every $x \in X$ the measure $\delta_{x,x}$ on $X \times X$.

Here is an observation which we will frequently use when instantiating our definitions and results in **Stoch**:

4.2. Lemma. *Two morphisms $f, g : A \rightarrow X_1 \otimes \dots \otimes X_n$ in **Stoch** are equal if and only if*

$$f(S_1 \times \dots \times S_n | a) = g(S_1 \times \dots \times S_n | a)$$

for all $a \in A$ and $S_i \in \Sigma_{X_i}$.

Proof. Use the fact that the cylinder sets $S_1 \times \dots \times S_n$ form a π -system which generates the product σ -algebra $\Sigma_{X_1} \otimes \dots \otimes \Sigma_{X_n}$, and apply the π - λ -theorem. $\square$

Although we will consider measure-theoretic probability to take place in **Stoch** in the rest of this paper, there is evidence that **Stoch** is not the only or even the best Markov category for this purpose; see Examples 10.4 and 11.18 for some of its counterintuitive and perhaps undesirable features. However, its full subcategory **BorelStoch** $\subseteq$ **Stoch**, consisting of all standard Borel spaces or equivalently all Polish spaces together with measurable maps as morphisms, turns out to be much more well-behaved. Since the Giry monad restricts to Polish spaces, **BorelStoch** also arises as a Kleisli category via Corollary 3.2. As far as we are aware, the only thing left to be desired of **BorelStoch** is that the Kolmogorov extension theorem holds in this Markov category only with respect to countable products [43, Example 3.6].

Finally, we present two other candidate categories which are of potential interest in the context of measure-theoretic probability, based on two related definitions of morphisms given by Sacksteder and Dawid, and which may relate to conditional expectations (see also Conjecture 11.10). Readers who are less interested in these technical details may proceed to Section 5 without any loss.

If (X, Σ_X) is a measurable space, then a σ -ideal is a subcollection $\mathcal{N}_X \subseteq \Sigma_X$ closed under countable unions and intersections with elements of Σ_X . The idea is that $\mathcal{N}_X$ specifies a collection of sets which are considered negligibly small. If $(X, \Sigma_X, \mathcal{N}_X)$ and $(Y, \Sigma_Y, \mathcal{N}_Y)$ are two measurable spaces with σ -ideals, then their *product* is defined by

$$(X, \Sigma_X, \mathcal{N}_X) \otimes (Y, \Sigma_Y, \mathcal{N}_Y) := (X \times Y, \Sigma_X \otimes \Sigma_Y, (\mathcal{N}_X \times \Sigma_Y) \vee (\Sigma_X \times \mathcal{N}_Y)),$$

where the indicated new σ -ideal is the one generated by all sets of the form $S \times T$ for $S \in \mathcal{N}_X$ and $T \in \Sigma_Y$ or $S \in \Sigma_X$ and $T \in \mathcal{N}_Y$. We would like this product to be the object part of a monoidal structure in both of the categories which we define next, although we have not yet been able to define a monoidal structure on the morphisms.

We will say that some property holds *almost everywhere* if the set of points where it does not hold is in the specified σ -ideal.

4.3. Example. Following Sacksteder [101, Section 2], a *statistical operation* $(X, \Sigma_X, \mathcal{N}_X) \rightarrow (Y, \Sigma_Y, \mathcal{N}_Y)$ is an equivalence class of maps

$$f : \Sigma_Y \times X \longrightarrow [0, 1], \quad (S, y) \longmapsto f(S|y),$$

which satisfy the following conditions:

- (a) For given $S \in \Sigma_Y$, the value $f(S|x)$ may not be defined for every x , but is defined almost everywhere;
- (b) $f(Y|x) = 1$ for almost every x ;

- (c) Given a sequence of disjoint sets $(S_n)_{n \in \mathbb{N}}$ in Σ_X , we have $f(\bigcup_n S_n|x) = \sum_n f(S_n|x)$ for almost every x ;
- (d) For every $S \in \mathcal{N}_Y$, we have $f(S|x) = 0$ for almost every x .

Although this does not seem to have been considered by Sacksteder, it is natural for two such maps f_1 and f_2 to be considered equivalent, $f_1 \sim f_2$, if for every $S \in \Sigma_Y$ the equation $f_1(S|x) = f_2(S|x)$ holds for almost every x .

We define composition of such statistical operations using the already familiar Chapman–Kolmogorov equation (4.1). Some straightforward measure-theoretic arguments show that this integral indeed makes sense, that the resulting composite again satisfies Properties (a) to (d), and that this composition operation respects the equivalence. Since the associativity of composition and the existence of identities are obvious, we can define the category of statistical operations in Sacksteder’s sense, using equivalence classes of statistical operations as morphisms.

In order to make this into a monoidal category, for every $f : A \rightarrow X$ and $g : B \rightarrow Y$ we would like to define their tensor product $f \otimes g : A \otimes B \rightarrow X \otimes Y$. On rectangles $S \times T$ for $S \in \Sigma_X$ and $T \in \Sigma_Y$, this tensor product should be a new statistical operation satisfying

$$(f \otimes g)(S \times T|a, b) := f(S|a)g(T|b)$$

whenever the right-hand side is defined. However, since neither $f(-|a)$ nor $g(-|b)$ is σ -additive in general, it is not clear to us whether this definition can be extended in any meaningful way from measurable rectangles to the product σ -algebra. In other words, it is not clear to us whether Sacksteder’s category of statistical operations can be made into a symmetric monoidal category, let alone into a Markov category.

For $(X, \Sigma_X, \mathcal{N}_X)$ a measurable space equipped with a σ -ideal, we write $L^\infty(X, \Sigma_X, \mathcal{N}_X)$, or more concisely $L^\infty(X)$, for the set of equivalence classes of bounded measurable functions $f : X \rightarrow \mathbb{R}$, where $f_1 \sim f_2$ if and only if $f_1(x) = f_2(x)$ for almost every x . Declaring $f \geq 0$ to hold if $f(x) \geq 0$ for almost all x makes $L^\infty(X)$ into an ordered vector space.

4.4. Example. Following Dawid [25, Section 3], a *statistical operation* $(X, \Sigma_X, \mathcal{N}_X) \rightarrow (Y, \Sigma_Y, \mathcal{N}_Y)$ is a map

$$F : L^\infty(Y) \rightarrow L^\infty(X)$$

having the following properties:

- (a) F is $\mathbb{R}$ -linear;
- (b) If $g \geq 0$ for $g \in L^\infty(Y)$, then also $F(g) \geq 0$ in $L^\infty(X)$;
- (c) $F(1) = 1$;

- (d) If $(g_n)_{n \in \mathbb{N}}$ is a monotonically increasing sequence in $L^\infty(Y)$ whose supremum exists in $L^\infty(Y)$, then $F(\sup_n g_n) = \sup_n F(g_n)$.

The idea is that F encodes the operation of pulling back a random variable along the statistical operation. It is clear that these statistical operations can be composed, resulting in the category of statistical operations in Dawid's sense.

As in the previous example, we run into a problem upon trying to define the tensor product of morphisms. If $F : L^\infty(X) \rightarrow L^\infty(A)$ and $G : L^\infty(Y) \rightarrow L^\infty(B)$ are two operators representing statistical operations $A \rightarrow X$ and $B \rightarrow Y$, then their tensor product $F \otimes G : L^\infty(A \times B) \rightarrow L^\infty(X \times Y)$ should be an operator which satisfies

$$(F \otimes G)(f \otimes g) := F(f) G(g)$$

for all $f \in L^\infty(X)$ and $g \in L^\infty(Y)$, where $f \otimes g \in L^\infty(X \times Y)$ is the product function given by $(f \otimes g)(x, y) := f(x)g(y)$, and similarly for the right-hand side.¹³ Upon trying to extend this definition to all $h \in L^\infty(X \times Y)$, we run into the problem that h may not have a representative as the supremum of an increasing sequence of functions of the form $\sum_{i=1}^n f_i \otimes g_i$. Hence we again do not know whether this category of statistical operations can be made into a symmetric monoidal category, let alone into a Markov category.

A possible way around this problem may be to restrict attention to those measurable spaces equipped with σ -ideals $(X, \Sigma_X, \mathcal{N}_X)$ which are such that $\Sigma_X / \mathcal{N}_X$ is a complete Boolean algebra, and then strengthen condition (d) by replacing sequences by arbitrary directed sets (Scott continuity). Then it would be enough to write every $h \in L^\infty(X \times Y)$ as the directed supremum of a set of functions of the form $\sum_{i=1}^n f_i \otimes g_i$.

Due to the problems with the definition of the tensor product of morphisms in the previous two examples, we have not yet been able to instantiate the definitions and results of Sections 10 to 16 in these cases. We also do not know whether either of these categories arises as a Kleisli category via Corollary 3.2, nor do we know if and how these two categories are related.

5. Example: compact Hausdorff spaces and continuous Markov kernels

In measure theory, it is often useful to work with probability measures on Borel σ -algebras of suitable topological spaces. We here investigate one particularly clean way to do so, based on the *Radon monad*. It involves $\mathbf{CHaus}$, the category of compact Hausdorff spaces and continuous maps, as perhaps the simplest nontrivial category of not necessarily finite topological spaces. We equip it with the cartesian monoidal structure.

In more detail, the Radon monad $R : \mathbf{CHaus} \rightarrow \mathbf{CHaus}$, as introduced by Świrszcz [111], and studied further by Furber and Jacobs [46, 45], is the canonical probability monad on

¹³ It is easy to see that the product $f \otimes g \in L^\infty(X \times Y)$ does not depend on the choice of representatives.

compact Hausdorff spaces. Its underlying functor assigns to every $X \in \mathbf{CHaus}$ the set of Radon probability measures on X , equipped with the coarsest topology which makes the integration maps

$$RX \longrightarrow \mathbb{R}, \quad \mu \longmapsto \int f \, d\mu$$

continuous for every continuous $f : X \rightarrow \mathbb{R}$. As shown in [46], the Kleisli category $\mathbf{Kl}(R)$ is opposite equivalent to the category of unital commutative C^* -algebras with positive unital linear maps as the morphisms.

Concretely, a morphism from X to Y in $\mathbf{Kl}(R)$ is represented by a continuous map $f : X \rightarrow RY$, and composition is again given by the Chapman–Kolmogorov equation in the form (4.1). In order to show that $\mathbf{Kl}(R)$ is also a Markov category, by Corollary 3.2 it is enough to equip R with the structure of a symmetric monoidal monad. We again want to do this by taking product measures,

$$RX \times RY \longrightarrow R(X \times Y), \quad (\mu, \nu) \longmapsto \mu \otimes \nu, \quad (5.1)$$

but we now run into the problem that the Borel σ -algebra on $X \otimes Y$ is generally much larger than the product of the Borel σ -algebras. We therefore define the product measure $\mu \otimes \nu$ via the Riesz representation theorem by putting

$$\int_{X \times Y} f \, d(\mu \otimes \nu) := \int_X \int_Y f \, d\mu \, d\nu, \quad (5.2)$$

leaving the necessary verifications of linearity and positivity to the reader. We now note that:

5.1. Lemma. *These maps equip R with the structure of a symmetric monoidal monad.*

Proof. As in the proof of Lemma 4.1, the nontrivial part is to show that (5.1) is actually a morphism, i.e. that it is continuous, and our argument will be structurally similar.

To see this, it is enough to show that for any continuous $f : X \times Y \rightarrow \mathbb{R}$, the composite

$$RX \times RY \longrightarrow R(X \times Y) \xrightarrow{\int f \, d(-)} \mathbb{R}$$

is continuous, which is by definition given by

$$(\mu, \nu) \longmapsto \int_X \int_Y f \, d\mu \, d\nu.$$

If f is of the form $f(x, y) = g(x)h(y)$ for continuous $g : X \rightarrow \mathbb{R}$ and $h : Y \rightarrow \mathbb{R}$, then it is easy to see that this holds by an argument analogous to the one around (4.3). Moreover,

the continuity clearly also holds for any linear combination of such functions; and since these linear combinations are also closed under multiplication and clearly separate points, the Stone–Weierstraß theorem implies that they are dense with respect to the supremum norm.

Now let an arbitrary f and $\varepsilon > 0$ be given, and suppose that we want to find a neighbourhood U of (μ, ν) in $RX \times RY$ such that $|\int_X \int_Y f d\mu' d\nu' - \int_X \int_Y f d\mu d\nu| < \varepsilon$ for all $(\mu', \nu') \in U$. We first choose an approximation $\|f - \sum_{i=1}^n g_i \otimes h_i\|_\infty < \varepsilon/3$ for suitable $n \in \mathbb{N}$ and continuous $g_i : X \rightarrow \mathbb{R}$ and $h_i : Y \rightarrow \mathbb{R}$, which is possible by the previous paragraph. We next choose a neighbourhood U such that the desired inequality holds for every $g_i \otimes h_i$ in place of f , and with quality of approximation $\varepsilon/3n$ in place of ε . This gives, for every $(\mu', \nu') \in U$,

$$\begin{aligned} & \left| \int_X \int_Y f d\mu' d\nu' - \int_X \int_Y f d\mu d\nu \right| \\ & \leq \left| \int_X \int_Y \left(f - \sum_{i=1}^n g_i \otimes h_i \right) d\mu' d\nu' \right| \\ & \quad + \left| \sum_{i=1}^n \int_X \int_Y (g_i \otimes h_i) d\mu' d\nu' - \sum_{i=1}^n \int_X \int_Y (g_i \otimes h_i) d\mu d\nu \right| \\ & \quad + \left| \int_X \int_Y \left(f - \sum_{i=1}^n g_i \otimes h_i \right) d\mu d\nu \right| \\ & \leq \varepsilon/3 + n \cdot \varepsilon/3n + \varepsilon/3 = \varepsilon, \end{aligned}$$

as was to be shown. $\square$

We therefore have that $\mathbf{Kl}(R)$ is a Markov category per Corollary 3.2. The copy maps are again represented by $x \mapsto \delta_{(x,x)}$.

6. Example: Gaussian probability theory

We now consider a Markov category describing Gaussian probability theory. This involves formulas for the composition of Gaussian conditionals reproducing those of Lauritzen and Jensen [78, Section 4.4]. As our previous two examples, it also has the flavour of a Kleisli category, although we conjecture that it does not actually arise in that way. A similar category was also considered by Golubtsov in [49, § 6] and [54, Section 8.2], who in addition claimed to have proven that it is not a Kleisli category.

So let **Gauss** be the monoidal category whose monoid of objects is $(\mathbb{N}, +)$, and whose morphisms $n \rightarrow m$ are tuples (M, C, s) with $M \in \mathbb{R}^{m \times n}$, $C \in \mathbb{R}^{m \times m}$ positive semidefinite, and $s \in \mathbb{R}^m$; we think of this data as defining the conditional distribution of a random variable $Y \in \mathbb{R}^m$ in terms of $X \in \mathbb{R}^n$ as

$$Y := MX + \xi, \quad (6.1)$$

where ξ is Gaussian noise independent of X with mean $\mathbf{E}[\xi] := s$ and covariance matrix

$$\mathbf{Var}[\xi] = \mathbf{E}[\xi\xi^T] - \mathbf{E}[\xi]\mathbf{E}[\xi]^T = \mathbf{E}[(\xi - s)(\xi - s)^T] := C.$$

If C is invertible, then the distribution of Y has a density with respect to the Lebesgue measure on $\mathbb{R}^m$ which is given by

$$\frac{1}{\sqrt{(2\pi)^m \det(C)}} \exp\left(-\frac{1}{2}(y - Mx - s)^T C^{-1}(y - Mx - s)\right).$$

Before formally defining the composition of two morphisms

$$(M, C, s) : \ell \rightarrow m, \quad (N, D, t) : m \rightarrow n,$$

let us consider what it should represent. If $Y = MX + \xi$ and $Z = NY + \eta$, then we get

$$Z = NMX + N\xi + \eta.$$

When composing two morphisms in a category of Markov kernels, we assume that the noise added in each of them is independent of the other. Hence we expect

$$\mathbf{E}[N\xi + \eta] = N \mathbf{E}[\xi] + \mathbf{E}[\eta] = Ns + t,$$

and

$$\mathbf{Var}[N\xi + \eta] = N \mathbf{Var}[\xi]N^T + \mathbf{Var}[\eta] = NCN^T + D.$$

Therefore what we want to define is the composition

$$(N, D, t) \circ (M, C, s) := (NM, NCN^T + D, Ns + t),$$

where it is straightforward to check that this is indeed associative and has the obvious identity morphisms $(1_n, 0, 0)$. This composition operation is our version of the “direct combination” equations given at [78, Section 4.4], specialized to case where no discrete variables are present.¹⁴ The tensor product of morphisms should correspond to acting

¹⁴ It should not be difficult to combine **FinStoch** and **Gauss** into one Markov category which models both discrete variables and Gaussian ones as in [78], but we have not yet done so.

on two random variables at the same time. This means that we want to take direct sums of all the data describing two arbitrary morphisms,

$$(M, C, s) \otimes (N, D, t) := (M \oplus N, C \oplus D, s \oplus t),$$

since for independent random variables $\xi \in \mathbb{R}^m$ and $\eta \in \mathbb{R}^n$, the expectation and variance of $(\xi, \eta) \in \mathbb{R}^{m+n}$ are given by

$$\mathbf{E}[(\xi, \eta)] = (\mathbf{E}[\xi], \mathbf{E}[\eta]), \quad \mathbf{Var}[(\xi, \eta)] = \begin{pmatrix} \mathbf{Var}[\xi] & 0 \\ 0 & \mathbf{Var}[\eta] \end{pmatrix}.$$

It is also straightforward to check that this indeed defines a (strict) symmetric monoidal category with the obvious symmetry isomorphisms. It may be interesting to note that forgetting the covariance matrix of every morphism, $(M, C, s) \mapsto (M, s)$, implements a strong monoidal functor from **Gauss** to the category of affine maps between Euclidean spaces.

Since the only morphism $n \rightarrow 0$ is the trivial one $(0, 0, 0)$, the monoidal unit $0 \in \mathbb{N}$ is terminal, so that **Gauss** is indeed semicartesian. Moreover, every object n has a canonical comultiplication given by

$$\text{copy}_n := \left(\begin{pmatrix} 1_n \\ 1_n \end{pmatrix}, 0, 0 \right),$$

which simply amounts to copying the value of a variable $X \in \mathbb{R}^n$, resulting in $(X, X) \in \mathbb{R}^{n+n}$. The necessary requirements to be a Markov category are again straightforward to check.

We briefly investigate the relation between **Gauss** and **Stoch**. Unfortunately this is somewhat interdependent with the upcoming Definition 10.14, since the following statement involves the notion of Markov functor introduced there.

6.1. Proposition. *There is a faithful Markov functor $\mathbf{Gauss} \rightarrow \mathbf{Stoch}$ which takes $n \mapsto \mathbb{R}^n$, where $\mathbb{R}^n$ carries the Borel σ -algebra.*

Proof. We have already described how to interpret a morphism $(M, C, s) : n \rightarrow m$ as a Markov kernel taking an $\mathbb{R}^n$ -valued variable to an $\mathbb{R}^m$ -valued variable. It is straightforward to verify that this defines a Markov functor $\mathbf{Gauss} \rightarrow \mathbf{Stoch}$, and that this functor is faithful. $\square$

6.2. Remark. As pointed out to us by Sam Staton, another Markov category with a similar flavour and significance for probability theory has been considered in [108]. In more explicit categorical terms, the objects are lists of natural numbers $(n_1, \dots, n_k)$, representing a disjoint union of measurable spaces $[0, 1]^{n_1} + \dots + [0, 1]^{n_k}$, and the morphisms are those Markov kernels that are definable just using beta and Bernoulli distributions.

The paper [108] then gives a description of this category in purely syntactic terms; for example, the comultiplications $\text{copy}_{(n)} : [0, 1]^n \rightarrow [0, 1]^{2n}$ are given syntactically by

$$(p_1, \dots, p_n) \mid x : 2n \vdash x(p_1, \dots, p_n, p_1, \dots, p_n).$$

7. Example: diagram categories and stochastic processes

In this section and the subsequent two, we present classes of examples of Markov categories arising from more general categorical constructions. Unfortunately, the construction of this section already requires the notion of deterministic morphism from Definition 10.1 and the subcategory of deterministic morphisms $\mathbf{C}_{\text{det}} \subseteq \mathbf{C}$ from Remark 10.13.

Let $\mathbf{C}$ be a Markov category and $\mathbf{D}$ an arbitrary small category. We then introduce another Markov category $\text{Fun}(\mathbf{D}, \mathbf{C})$. Its objects are defined to be functors $\mathbf{D} \rightarrow \mathbf{C}_{\text{det}}$, which we think of as diagrams with shape $\mathbf{D}$ of deterministic morphisms in $\mathbf{C}$, and its morphisms are natural transformations between these functors (with not necessarily deterministic components). We use notation $X = (X_d)_{d \in \mathbf{D}}$ for such functors, emphasizing their interpretation as diagrams.

7.1. Proposition. *Under these assumptions, $\text{Fun}(\mathbf{D}, \mathbf{C})$ is again a Markov category with respect to the pointwise monoidal structure: for $X, Y \in \text{Fun}(\mathbf{D}, \mathbf{C})$, we put*

$$(X \otimes Y)_d := X_d \otimes Y_d,$$

with the obvious action on morphisms of $\mathbf{D}$, and similarly for the tensor product of morphisms. The comultiplications are also defined pointwise,

$$(\text{copy}_X)_d := \text{copy}_{X_d} : X_d \longrightarrow X_d \otimes X_d. \quad (7.1)$$

The proof is straightforward; the assumption that all morphisms in the diagrams (X_d) are deterministic is required in order to make the comultiplications (7.1) into natural transformations.

As an interesting and highly relevant example, we now illustrate how Proposition 7.1 lets us instantiate the theory of Markov categories in the context of stochastic processes. Here, we take $\mathbf{D} := (\mathbb{Z}, \geq)$, the category whose objects are integers, with exactly one morphism $n \rightarrow m$ if and only if $n \geq m$ and no morphism otherwise. For example with $\mathbf{C} = \text{BorelStoch}$, the functors $\mathbf{D} \rightarrow \mathbf{C}_{\text{det}}$ can be identified with infinite diagrams of measurable spaces

$$\dots \longrightarrow X_{n+1} \longrightarrow X_n \longrightarrow X_{n-1} \longrightarrow \dots$$

where all the maps involved are measurable functions (Example 10.5). We think of each component X_n as the space of joint values of a stochastic process over all times $t \leq n$, where the map $X_{n+1} \rightarrow X_n$ amounts to forgetting the value of the process at time $n+1$. A

morphism $I \rightarrow X$ in $\text{Fun}(\mathbf{D}, \mathbf{C})$ then assigns a probability measure to each of the measurable spaces X_n in such a way that the projections $X_{n+1} \rightarrow X_n$ are measure-preserving; thus such a morphism exactly turns X into a stochastic process. More generally, a morphism $f : A \rightarrow X$ may be interpreted as a stochastic process with control: A_n is the space of values of control parameters over all times $t \leq n$, and the component $f_n : A_n \rightarrow X_n$ assigns a joint distribution over all values of the stochastic process as a function of these control parameters. The required commutativity of the diagram

$$\begin{array}{ccc} A_{n+1} & \longrightarrow & A_n \\ \downarrow f_{n+1} & & \downarrow f_n \\ X_{n+1} & \longrightarrow & X_n \end{array}$$

then ensures that these joint distributions are compatible in exactly the way that one would expect: given the control parameters up to time $n + 1$, the marginal of the associated joint distribution up to time n coincides with the joint distribution associated to the control parameters up to time n .

It therefore follows that all of our definitions and results can be applied at the level of stochastic processes.¹⁵ For example, the definition of statistic from Definition 14.2 amounts to the following: a statistic consists of another sequence of spaces $(T_n)_{n \in \mathbb{Z}}$ together with a sequence of deterministic maps $s_n : X_n \rightarrow T_n$ such that the diagrams

$$\begin{array}{ccc} X_{n+1} & \longrightarrow & X_n \\ \downarrow s_{n+1} & & \downarrow s_n \\ T_{n+1} & \longrightarrow & T_n \end{array}$$

commute. This states that for every given value of X_n , the values of the statistic T_n can be computed from the value of the process up to time $n + 1$ by first computing the values of the extended statistic T_{n+1} and then projecting, or by marginalizing the process first to time up to n and then applying the statistic there. Lauritzen's notion of *transitivity* of a sequence of statistics [76, Definition 2.1] is closely related.

Thus we can apply the upcoming results of Sections 14 to 16 to the relevant functor category and thereby obtain results about this type of statistic living in time.

7.2. Remark. There are many conceivable variations on this basic idea. The extension from discrete to continuous time is fairly obvious. We can also consider *stationary* stochastic processes as follows. Let now $\mathbf{D}$ be the category consisting of one single object, such that its monoid of endomorphisms is $(\mathbb{Z}, +)$. Then the objects of $\text{Fun}(\mathbf{D}, \mathbf{BorelStoch})$ are measurable spaces equipped with an action of $\mathbb{Z}$, which makes them into dynamical

¹⁵ Golubtsov's mention of "dynamical nondeterministic information transformers [...] and information interactions evolving in time" [54, p. 64] suggests that he may already have been aware of this possibility.

systems. The morphisms $I \rightarrow X$ are then easily seen to be in canonical bijection with stationary measures, turning X into a measure-preserving dynamical system.

7.3. Remark. The developments of this section are not sufficiently expressive to be turned into a formal theory of stochastic processes, because the latter would clearly require more structure: the theory of Markov categories itself does not provide any way to talk about the temporal aspects. Nevertheless, we believe that it may be possible to develop a formal theory of stochastic processes by equipping Markov categories with extra structure, such as an action of a functor $\mathbf{C} \rightarrow \mathbf{C}$ which implements time translation, or more generally a group action on $\mathbf{C}$.

8. Example: hypergraph categories

Let $\mathbf{D}$ be a hypergraph category [10,72,34,36]. This means in particular that each object $X \in \mathbf{D}$ comes equipped with a distinguished comultiplication $\text{copy}_X : X \rightarrow X \otimes X$ and counit morphism $\text{del}_X : X \rightarrow I$ satisfying (2.4). Now let $\mathbf{C}$ be the subcategory consisting of all morphisms which preserve the counit. In other words, $f : X \rightarrow Y$ is in $\mathbf{C}$ if and only if $\text{del}_Y \circ f = \text{del}_X$. Then it is not hard to see that $\mathbf{C}$ carries the structure of a Markov category.

8.1. Example. A good example is $\mathbf{FinRel}$, the category of finite sets and relations, with its usual monoidal structure given by the cartesian product and the canonical hypergraph structure where the distinguished relation $X^n \rightarrow X^m$ relates a tuple $(x_1, \dots, x_n)$ to a tuple $(x'_1, \dots, x'_m)$ if and only if $x_1 = \dots = x_n = x'_1 = \dots = x'_m$, implementing both “copying” as a morphism $X \rightarrow X \otimes X$ and “testing for equality” as a morphism $X \otimes X \rightarrow I$.

Restricting $\mathbf{FinRel}$ to the counit-preserving morphisms recovers exactly $\mathbf{FinSetMulti}$ (Example 2.6).

8.2. Example. A closely related example is the hypergraph category $\mathbf{Mat}(R)$ for a semiring $(R, +, \cdot)$. Its objects are finite sets, and morphisms $X \rightarrow Y$ are matrices $(f_{xy})_{x \in X, y \in Y}$ with $f_{xy} \in R$, which compose by matrix multiplication. The monoidal structure is given by cartesian product of sets and tensor product of matrices. If R is the Boolean semiring $R = \{0, 1\}$ with $1 + 1 := 1$, then $\mathbf{Mat}(R) \cong \mathbf{FinRel}$, so that this generalizes the previous example.

Restricting to the counit-preserving morphisms results in a Markov category, namely the category of “stochastic matrices” with entries in R . In particular for $R = \mathbb{R}_+$, we obtain the category of stochastic matrices, which is equivalent to $\mathbf{FinStoch}$ (Example 2.5). Taking $R := ([0, 1], \max, \min)$ or $R := ([0, 1], \max, \cdot)$ recovers Golubtsov’s two categories of *fuzzy information transformers* [54, Section 8.5], modulo the minor difference that our objects are merely finite instead of arbitrary sets.

While the idea of imposing normalization in this way by restricting to the subcategory of counit-preserving morphisms is not new and can be done in somewhat greater general-

ity (see e.g. [15, Section 7]), we believe that applying this procedure in the particular case of hypergraph categories, such as the ones of [34], has the potential to produce further examples of Markov categories in which our results will have interesting instantiations.

9. Example: categories of commutative comonoids

Let $\mathcal{D}$ be a symmetric monoidal category, and let $\mathcal{C}$ be the category of commutative comonoids in $\mathcal{D}$ with morphisms those maps which preserve the counits. Hence an object in $\mathcal{C}$ is by definition a triple (X, μ_X, e_X) with $\mu_X : X \rightarrow X \otimes X$ and $e_X : X \rightarrow I$ in $\mathcal{D}$ satisfying the usual equations, while a $\mathcal{C}$ -morphism $(X, \mu_X, e_X) \rightarrow (Y, \mu_Y, e_Y)$ is a $\mathcal{D}$ -morphism $f : X \rightarrow Y$ such that $e_Y = e_X f$. The usual tensor product of commutative comonoids makes $\mathcal{C}$ into a symmetric monoidal category, where the symmetry is also the one inherited from $\mathcal{D}$. Every object $(X, \mu_X, e_X) \in \mathcal{C}$ now has a canonical commutative comonoid structure not only in $\mathcal{D}$, but even in $\mathcal{C}$, given by the structure maps $\text{copy}_X := \mu_X$ and $\text{del}_X := e_X$. We have therefore sketched the proof of:

9.1. Proposition. *If $\mathcal{D}$ is a symmetric monoidal category, then the category of commutative comonoids in $\mathcal{D}$ with counit-preserving morphisms is a Markov category.*

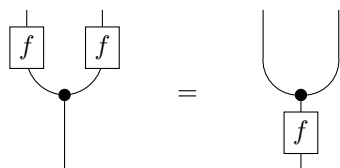
One use of Markov categories of this type is that they can provide counterexamples to plausible-sounding conjectures, as for example in Remark 10.10 and Example 11.33.

10. Deterministic morphisms and a strictification theorem

We now return to the development of the general theory, frequently illustrating it in some of the previous examples.

The reader may have noticed that Definition 2.1 postulates the naturality of del but not of copy , for which it would take the form (10.1). Indeed, the reason is that this naturality does not hold in the examples of interest to us, such as FinStoch from Example 2.5, or more generally Stoch , the example of measurable Markov kernels between measurable spaces from Section 4. The simple intuitive reason is that applying a Markov kernel independently to two copies of a variable is different than applying the map first and then copying. In fact, those morphisms for which the naturality holds are of a very special kind:

10.1. Definition. A morphism $f : X \rightarrow Y$ in a Markov category is *deterministic* if it respects the comultiplication,


(10.1)

This definition goes back to the characterization of functions as comonoid-preserving relations given by Carboni and Walters [9, Lemma 2.5] and has also been used in the present form by Fong [33, Lemma 5.2]. Deterministic morphisms have also been investigated as *copyable* morphisms in the program semantics literature [112,44]. In the specific context of Markov kernels, our definition may first have been used in [20, Section 3.3], although there also are strong similarities with Golubtsov's earlier work [54, Eq. (2)].

In terms of our notion of conditional independence to be introduced in Definition 12.12, this can also be understood as saying that the morphisms on the right-hand side displays the conditional independence $Y \perp Y \parallel X$. Intuitively, we can phrase this as saying that the random variable f is independent of itself, which is a standard equivalent way to talk about deterministic events.

We denote the collection of deterministic morphisms by $\mathbf{C}_{\text{det}} \subseteq \mathbf{C}$. As we will show in Remark 10.13, $\mathbf{C}_{\text{det}}$ is a symmetric monoidal subcategory of $\mathbf{C}$.

10.2. Remark. The commutative comonoid axioms of Definition 2.1 imply that the structure maps $\text{del}_X : X \rightarrow I$ and $\text{copy}_X : X \rightarrow X \otimes X$ are themselves deterministic.

10.3. Example. It is easy to see that in $\mathbf{FinStoch}$, the deterministic maps are exactly the usual ones, i.e. those stochastic matrices f where each entry f_{xy} is either 0 or 1. It follows that the deterministic morphisms form a symmetric monoidal subcategory which is precisely the image of the inclusion $\mathbf{FinSet} \rightarrow \mathbf{FinStoch}$.

10.4. Example. Let us consider deterministic morphisms in $\mathbf{Stoch}$, the category of measurable spaces and Markov kernels (Section 4). For $f : (X, \Sigma_X) \rightarrow (Y, \Sigma_Y)$, the left-hand side of the determinism equation is represented by the map

$$(\Sigma_Y \otimes \Sigma_Y) \times X \longrightarrow [0, 1], \quad (S \times T, x) \longmapsto f(S|x)f(T|x),$$

where we have specified the map only on generating rectangles $S \times T \in (\Sigma_Y \otimes \Sigma_Y)$, which is enough. Similarly, the right-hand side of the determinism equation is

$$(\Sigma_Y \otimes \Sigma_Y) \times X \longrightarrow [0, 1], \quad (S \times T, x) \longmapsto f(S \cap T|x).$$

So if the left-hand side and right-hand side are equal, then we must have $f(S|x)^2 = f(S|x)$ for all $S \in \Sigma_Y$, which implies that every measure $f(-|x) : \Sigma_Y \rightarrow [0, 1]$ is actually $\{0, 1\}$ -valued. Conversely, Lemma 4.2 and additivity of f in the first argument show that this is enough. Hence the deterministic morphisms in $\mathbf{Stoch}$ are precisely those Markov kernels which land in $\{0, 1\}$ -valued measures.

In general, a $\{0, 1\}$ -valued measure does *not* need to be a delta measure; a standard example is given by the countable-cocountable σ -algebra on an uncountable set Y , with a $\{0, 1\}$ -valued measure which assigns measure 1 to precisely the cocountable sets. So although the deterministic morphisms are again precisely those Markov kernels which do not involve any randomness, the subcategory of deterministic morphisms does not

coincide with **Meas**, the category of measurable spaces: the induced functor from **Meas** to the deterministic morphisms in **Stoch** is not full. And it is not even faithful either, although for a more stupid reason: if Σ_Y does not separate the points of Y , then there will be different measurable maps $1 \rightarrow Y$ whose induced delta measures $\Sigma_Y \rightarrow \{0, 1\}$ are the same. As is well-known, both pathologies can be avoided by imposing suitable restrictions on the measurable spaces involved. Namely if Σ_Y is countably generated and separates the points of Y , then the deterministic morphisms $(X, \Sigma_X) \rightarrow (Y, \Sigma_Y)$ are exactly the measurable maps, as can be seen as follows. The map $y \mapsto \delta_y$ identifies the elements of Y with the $\{0, 1\}$ -valued probability measures $\Sigma_Y \rightarrow [0, 1]$ by [99, p. 14], and the measurability condition on $f(S|-) : X \rightarrow [0, 1]$ for $S \in \Sigma_Y$ then becomes precisely the requirement that

$$\{x \in X \mid f(S|x) = 1\} \in \Sigma_X,$$

which amounts to the condition that the representing map $X \rightarrow Y$ must be measurable. Hence the functor from **Meas** to deterministic morphisms in **Stoch** is fully faithful on morphisms into (Y, Σ_Y) as soon as Σ_Y is countably generated and separating, but in general neither full nor faithful. We take this as evidence for the hypothesis that **Stoch** is not actually the “best” Markov category for measure-theoretic probability. (See also Examples 4.3 and 4.4.)

10.5. Example. The situation is much better in the full subcategory $\mathbf{BorelStoch} \subseteq \mathbf{Stoch}$. Thinking of a standard Borel space as a Polish space, noting that every probability measure on a Polish space has a support implies that a $\{0, 1\}$ -valued measure is indeed a delta measure. Furthermore using the fact that the relevant Borel σ -algebras separate points, it follows that the subcategory of deterministic morphisms is isomorphic to the category of standard Borel spaces and measurable maps.

10.6. Remark. Consider more generally the setting of Corollary 3.2. There, it is easy to see that the image of every deterministic morphism in **D** is also deterministic in $\mathbf{Kl}(T)$. But as we have seen in Example 10.4, the functor $\mathbf{D} \rightarrow \mathbf{Kl}(T)$ need not be faithful, and it may also not be full on deterministic morphisms. In other words, deterministic morphisms and morphisms of **D** must be carefully distinguished.

10.7. Example. In **Gauss**, a morphism $(M, C, s) : m \rightarrow n$ is deterministic if and only if the morphism

$$\left(\begin{pmatrix} 1_n \\ 1_n \end{pmatrix}, 0, 0 \right) \circ (M, C, s) = \left(\begin{pmatrix} M \\ M \end{pmatrix}, \begin{pmatrix} C & C \\ C & C \end{pmatrix}, \begin{pmatrix} s \\ s \end{pmatrix} \right)$$

is equal to the morphism

$$((M, C, s) \otimes (M, C, s)) \circ \left(\begin{pmatrix} 1_m \\ 1_m \end{pmatrix}, 0, 0 \right) = \left(\begin{pmatrix} M \\ M \end{pmatrix}, \begin{pmatrix} C & 0 \\ 0 & C \end{pmatrix}, \begin{pmatrix} s \\ s \end{pmatrix} \right).$$

Therefore (M, C, s) is deterministic if and only if $C = 0$, which indeed means if and only if there is no randomness involved.

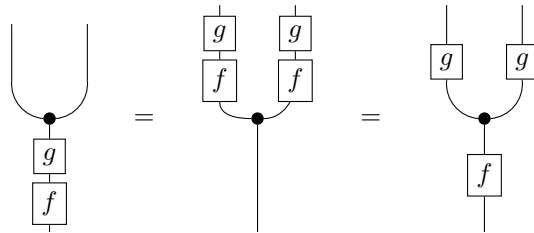
10.8. Example. In a category of commutative comonoids and counit-preserving morphisms (Proposition 9.1), the deterministic morphisms are precisely the comonoid homomorphisms.

Let us return to the general theory. Deterministic morphisms in a Markov category have the following in common with epimorphisms in any category:

10.9. Lemma. *Let $f : X \rightarrow Y$ and $g : Y \rightarrow Z$. Then:*

- (a) *If f and g are deterministic, then so is gf .*
- (b) *If gf is deterministic and f is a deterministic epimorphism, then g is deterministic as well.*
- (c) *If f and g are inverse isomorphisms and if one of them is deterministic, then so is the other.*

Proof. We show (b), leaving the even simpler proofs of (a) and (c) to the reader. We compute



and now use the epimorphism property to cancel f . $\square$

We dedicate the remainder of this section to formal categorical considerations. These address some important technical issues which turn out to be surprisingly subtle. Readers who are primarily interested in our treatment of conditional independence and statistics instead can safely move on to Section 11.

The following observation is the main reason for why categorical subtleties arise.

10.10. Remark. There are Markov categories $\mathbf{C}$ which contain isomorphisms that are not deterministic.

We can build concrete examples of this phenomenon based on the construction of Markov categories $\mathbf{C}$ as categories of comonoids in symmetric monoidal categories $\mathbf{D}$ via Proposition 9.1. The idea is the same object $X \in \mathbf{D}$ may support two comonoid structures with the same counit $e_X : X \rightarrow I$ but different comultiplications

$$\mu_X, \mu'_X : X \rightarrow X \otimes X,$$

resulting in a non-deterministic isomorphism

$$\text{id}_X : (X, \mu_X, e_X) \rightarrow (X, \mu'_X, e_X).$$

Here, it may or may not be the case that the two comonoids (X, μ_X, e_X) and (X, μ'_X, e_X) are isomorphic in some other way.

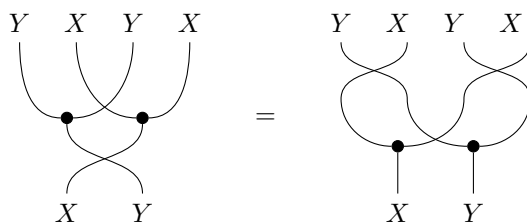
More concretely, $\mathbf{D}$ may be the category of finite-dimensional vector spaces over a field k equipped with the usual tensor product of vector spaces. Then since this $\mathbf{D}$ is monoidally equivalent to its own opposite, we can identify $\mathbf{C}$ with the opposite of the category of finite-dimensional commutative k -algebras together with k -linear unital maps. If we restrict to the k -algebras of the form k^n (with componentwise operations), then our $\mathbf{D}$ becomes equivalent to the category of “stochastic matrices” with entries in k , where a stochastic matrix is $(f_{xy})_{x \in X, y \in Y} \in k^{X \times Y}$ with $\sum_y f_{xy} = 1$ as in Example 2.5. Equivalently, this $\mathbf{D}$ is the full subcategory of the Kleisli category of the k -affine space monad on $\mathbf{Set}$ with finite sets as objects. But without the restriction to k -algebras of this form, the category is significantly larger, and contains for example algebras like $k[X]/\langle X^2 \rangle$. This algebra is isomorphic to k^2 in $\mathbf{C}$, even in a number of different ways, but none of these isomorphisms is deterministic because $k[X]/\langle X^2 \rangle$ and k^2 are not isomorphic as k -algebras.

10.11. Remark. We expect that the remainder of this section will also apply, after some straightforward modifications, in the context of *hypergraph categories* [72,34,36], and perhaps more generally to symmetric monoidal categories *supplied* with extra structure [35].

10.12. Lemma. *The subcategory of deterministic morphisms $\mathbf{C}_{\text{det}} \subseteq \mathbf{C}$ is closed under tensor product and contain all structure maps, meaning the associators, unitors, del , copy , and swap .*

See the proof of [35, Theorem 3.12] for an independently obtained similar result in a more general setup for symmetric monoidal categories with extra “bells and whistles”.

Proof. The first claim is a direct consequence of (2.4). That the structure maps are deterministic is easiest to see for every del_X and copy_X (Remark 10.2). For $\text{swap}_{X,Y} : X \otimes Y \rightarrow Y \otimes X$, we only need (2.4) and the naturality of the symmetry isomorphisms,



The unitors are deterministic thanks to (2.7) applied with one of the two tensor factors being I together with copy_I being equal to the unitor $I \cong I \otimes I$. Finally, the problem is trickiest for an associator $(X \otimes Y) \otimes Z \xrightarrow{\cong} X \otimes (Y \otimes Z)$. For this case, we denote the composite vertical isomorphism $(X \otimes Y) \otimes (X \otimes Y) \rightarrow (X \otimes X) \otimes (Y \otimes Y)$ of (2.7) by $\text{gr}_{X,Y}$. Writing “;” instead of “ $\otimes$ ” for brevity, we consider the diagram

$$\begin{array}{ccc}
 (X; Y); Z & \xrightarrow{\text{copy}_{(X; Y); Z}} & ((X; Y); Z); ((X; Y); Z) \\
 \parallel & & \downarrow \text{gr}_{(X; Y); Z} \\
 (X; Y); Z & \xrightarrow{\text{copy}_{X; Y}; \text{id}} & ((X; Y); (X; Y)); Z \xrightarrow{\text{id}; \text{copy}_Z} ((X; Y); (X; Y)); (Z; Z) \\
 \parallel & & \downarrow \text{gr}_{X; Y}; \text{id} \\
 (X; Y); Z & \xrightarrow{(\text{id}; \text{copy}_Y); \text{id}} (X; (Y; Y)); Z \xrightarrow{(\text{copy}_X; \text{id}); \text{id}} ((X; X); (Y; Y)); Z \xrightarrow{\text{id}; \text{copy}_Z} ((X; X); (Y; Y)); (Z; Z) & \downarrow \text{gr}_{X; Y}; \text{id} \\
 \cong \downarrow & \cong \downarrow & \downarrow \cong \\
 X; (Y; Z) & \xrightarrow{\text{id}; (\text{copy}_Y; \text{id})} X; ((Y; Y); Z) \xrightarrow{\text{copy}_X; \text{id}} (X; X); ((Y; Y); Z) \xrightarrow{\text{id}; (\text{id}; \text{copy}_Z)} (X; X); ((Y; Y); (Z; Z)) & \downarrow \cong \\
 \parallel & \parallel & \parallel \\
 X; (Y; Z) & \xrightarrow{(\text{id}; \text{copy}_Y); \text{id}} X; ((Y; Y); Z) \xrightarrow{\text{id}; (\text{id}; \text{copy}_Z)} X; ((Y; Y); (Z; Z)) \xrightarrow{\text{copy}_X; \text{id}} (X; X); ((Y; Y); (Z; Z)) & \downarrow \text{id}; \text{gr}_{Y; Z}^{-1} \\
 \parallel & \parallel & \parallel \\
 X; (Y; Z) & \xrightarrow{\text{id}; \text{copy}_{Y; Z}} X; ((Y; Z); (Y; Z)) \xrightarrow{\text{copy}_X; \text{id}} (X; X); ((Y; Z); (Y; Z)) & \downarrow \text{id}; \text{gr}_{Y; Z}^{-1} \\
 \parallel & \parallel & \downarrow \text{gr}_{X; (Y; Z)} \\
 X; (Y; Z) & \xrightarrow{\text{copy}_{X; (Y; Z)}} & (X; (Y; Z)); (X; (Y; Z))
 \end{array}$$

Each square commutes either by standard properties of symmetric monoidal categories, including naturality of the associator, or as an instance of (2.7). Note that the bottom three rows of the diagram are somewhat similar to the top three. Commutativity of this diagram establishes the claim: the left vertical composite is obviously just the associator, while the right vertical composite is the tensor product of the associator with itself by the coherence theorem for symmetric monoidal categories, where applying this requires one to distinguish the two factors of X on the right, and likewise for Y and Z , noting that their order is the same at the bottom as it is on top. $\square$

10.13. Remark. Since we already know that the deterministic morphisms are closed under composition and include all structure maps, Lemma 10.12 therefore concludes that they form a symmetric monoidal subcategory $\mathcal{C}_{\text{det}} \subseteq \mathcal{C}$. Using the well-known fact that having

a natural comonoid structure on every object of a symmetric monoidal category implies that the monoidal structure is cartesian (e.g. [63]), we see that $\mathbf{C}_{\text{det}}$ is cartesian monoidal. We will make judicious use of this observation below. It provides the basis for relating our Definition 2.1 to Golubtsov's formulation in terms of cartesian monoidal structure [54].

Note also that every monoidal product $X \otimes Y$ in $\mathbf{C}$, when equipped with the two projections

$$\text{id}_X \otimes \text{del}_Y : X \otimes Y \longrightarrow X, \quad \text{del}_X \otimes \text{id}_Y : X \otimes Y \longrightarrow Y,$$

always satisfies the *existence* part of the universal property of a product by (2.10). The uniqueness part hinges on the naturality of *copy*.

10.14. Definition. The collection of Markov categories makes up a 2-category *Markov*, where:

- ▷ objects are Markov categories;
- ▷ 1-morphisms $F : \mathbf{C} \rightarrow \mathbf{D}$ are *Markov functors*, by which we mean strong symmetric monoidal functors which preserve the comonoids in the sense that the diagram

$$\begin{array}{ccc} & FX & \\ \text{copy}_{FX} \swarrow & & \searrow F(\text{copy}_X) \\ FX \otimes FX & \xrightarrow{\cong} & F(X \otimes X) \end{array} \quad (10.2)$$

commutes for every $X \in \mathbf{C}$, where the horizontal arrow is the relevant structure morphism of F ;

- ▷ 2-morphisms $t : F \Rightarrow G$ for $F, G : \mathbf{C} \rightarrow \mathbf{D}$ are monoidal natural transformations with deterministic components.

A Markov functor which is faithful is also called a *Markov embedding*.

For both 1-morphisms and 2-morphisms, one might also expect the obvious compatibility conditions with the counits to be required. These hold automatically due to the terminality of the monoidal unit, and we therefore do not spell them out. Another relevant observation is that the monoidal structure isomorphisms $FX \otimes FY \xrightarrow{\cong} F(X \otimes Y)$ are automatically deterministic [35, Theorem 4.6].

10.15. Lemma. *Let $F : \mathbf{C} \rightarrow \mathbf{D}$ be a comonoid functor. Then F takes deterministic morphisms to deterministic morphisms, giving a functor $F_{\text{det}} : \mathbf{C}_{\text{det}} \rightarrow \mathbf{D}_{\text{det}}$. Moreover, F_{det} preserves finite products.*

Proof. Straightforward. $\square$

We say that $F : \mathcal{C} \rightarrow \mathcal{D}$ is a *comonoid equivalence* if it is an equivalence in the 2-category **Markov**. The following result provides a convenient way of recognizing comonoid equivalences, analogous to the standard fact that a functor is an equivalence of categories if and only if it is fully faithful and essentially surjective.

10.16. Proposition. *A symmetric monoidal functor $F : \mathcal{C} \rightarrow \mathcal{D}$ is a comonoid equivalence if and only if it is an equivalence of categories, and for every object $Y \in \mathcal{D}$ there is $X \in \mathcal{C}$ and an isomorphism $i : Y \xrightarrow{\cong} FX$ which is deterministic, meaning that the diagram*

$$\begin{array}{ccc} Y & \xrightarrow[\cong]{i} & FX \\ \text{copy}_Y \downarrow & & \downarrow \text{copy}_{FX} \\ Y \otimes Y & \xrightarrow[\cong]{i \otimes i} & FX \otimes FX \end{array}$$

commutes.

Proof. If F is an equivalence in **Markov** with inverse equivalence $G : \mathcal{D} \rightarrow \mathcal{C}$, then we can take $X := GY$, and the square commutes because the given natural isomorphism $FG \cong \text{id}_{\mathcal{D}}$ is required to have deterministic components.

Conversely, choose for every object $Y \in \mathcal{D}$ an object $X \in \mathcal{C}$ together with $i : Y \xrightarrow{\cong} FX$ which makes the square commute. Then it is a standard fact that these choices determine an equivalence $g : \mathcal{D} \rightarrow \mathcal{C}$ sending every $Y \in \mathcal{D}$ to the corresponding $X \in \mathcal{C}$, and being right adjoint to F via $\varepsilon : FG \cong \text{id}_{\mathcal{D}}$ and $\eta : \text{id}_{\mathcal{C}} \cong GF$ [86, Theorem IV.4.1]. Moreover, G has a unique monoidal structure which makes it into a monoidal equivalence weakly inverse to F [100, 4.4.2.1]. Now the natural isomorphism $\varepsilon : FG \cong \text{id}_{\mathcal{D}}$ has deterministic components by construction. We show that same holds true for $\eta : \text{id}_{\mathcal{C}} \cong GF$. Since F is faithful, it is enough to prove commutativity of the relevant diagrams after application of F . In other words, we only need to show that

$$\begin{array}{ccc} FX & \xrightarrow{F\eta_X} & FGFX \\ \text{copy}_{FX} \downarrow & & \downarrow \text{copy}_{FGFX} \\ FX \otimes FX & \xrightarrow{F\eta_X \otimes F\eta_X} & FGFX \otimes FGFX \\ \cong \downarrow & & \downarrow \cong \\ F(X \otimes X) & \xrightarrow{F(\eta_X \otimes \eta_X)} & F(GFX \otimes GFX) \end{array}$$

commutes for every $X \in \mathcal{C}$. Since the lower square commutes by naturality, it remains to be shown that $F\eta_X$ is deterministic. But this is because $F\eta_X$ is inverse to ε_{FX} by the triangle identities for adjoints, and we already know that ε_{FX} is deterministic.

We still need to prove that G preserves the comonoids. We first show this for objects of the form $FX \in \mathcal{D}$, where we have

$$\begin{array}{ccccccc}
 GFX & \xlongequal{\quad} & GFX & \xrightarrow{\eta_X^{-1}} & X & \xrightarrow{\eta_X} & GFX \\
 \downarrow G(\text{copy}_{FX}) & & \downarrow GF(\text{copy}_X) & & \downarrow \text{copy}_X & & \downarrow \text{copy}_{GFX} \\
 G(FX \otimes FX) & \xrightarrow{G(\cong)} & GF(X \otimes X) & \xrightarrow{\eta_{X \otimes X}^{-1}} & X \otimes X & \xrightarrow{\eta_X \otimes \eta_X} & GFX \otimes GFX
 \end{array}$$

where the top composite is id_{GFX} as desired, and the bottom composite is the desired structure isomorphism $G(FX \otimes FX) \cong GFX \otimes GFX$ thanks to the monoidality of η . The left square commutes since F is a comonoid functor, the middle square by naturality of η^{-1} , and the right one since η_X is deterministic. The general case reduces to this one via $i : Y \xrightarrow{\cong} FX$ and the diagram

$$\begin{array}{ccccccc}
 GY & \xrightarrow{Gi} & GFX & \xlongequal{\quad} & GFX & \xrightarrow{G(i^{-1})} & GY \\
 \downarrow G(\text{copy}_Y) & & \downarrow G(\text{copy}_{FX}) & & \downarrow \text{copy}_{GFX} & & \downarrow \text{copy}_{GY} \\
 G(Y \otimes Y) & \xrightarrow{G(i \otimes i)} & G(FX \otimes FX) & \xrightarrow{\cong} & GFX \otimes GFX & \xrightarrow{G(i^{-1}) \otimes G(i^{-1})} & GY \otimes GY
 \end{array}$$

where the top composite is simply the desired id_{GY} , and the bottom composite is the desired structure isomorphism $G(Y \otimes Y) \cong GY \otimes GY$. The left square is the G -image of a commuting square by the assumption that i is deterministic, the middle square commutes as per the above, and the right square because $G(i^{-1})$ is deterministic by virtue of being equal to $G\varepsilon_Y$, which is the inverse of the deterministic morphism η_{GY} . $\square$

The following theorem is analogous to an existing strictification theorem for hypergraph categories due to Fong and Spivak [36, Theorem 1.1], whereas our proof has the additional features that it uses Mac Lane's strictification theorem as a mere black box. It has inspired Fong and Spivak to prove a more general strictification theorem for symmetric monoidal categories with extra structure [35, Proposition 3.24 and Corollary 4.9].

Let us say that a Markov category is *strict* if its underlying monoidal category is strict.

10.17. Theorem. *Every Markov category is comonoid equivalent to a strict one.*

Proof. Consider the subcategory of deterministic morphisms $\mathbf{C}_{\text{det}} \subseteq \mathbf{C}$, which is cartesian monoidal, and choose a strict symmetric monoidal category $\mathbf{C}'_{\text{det}}$ with a symmetric monoidal equivalence $F_{\text{det}} : \mathbf{C}'_{\text{det}} \xrightarrow{\cong} \mathbf{C}_{\text{det}}$. Now let $\mathbf{C}'$ be the monoidal category having the same objects as $\mathbf{C}'_{\text{det}}$, but hom-sets $\mathbf{C}'(X, Y) := \mathbf{C}(F_{\text{det}}X, F_{\text{det}}Y)$. If we consider $\mathbf{C}'_{\text{det}}$ to be a subcategory of $\mathbf{C}'$ via F_{det} , and write F for the obvious functor $\mathbf{C}' \rightarrow \mathbf{C}$, then we have a diagram of categories

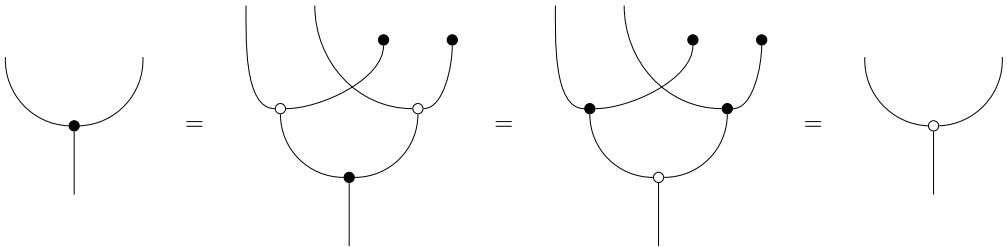
$$\begin{array}{ccc}
 \mathbf{C}'_{\text{det}} & \xrightarrow{F_{\text{det}}} & \mathbf{C}_{\text{det}} \\
 \text{in} & & \text{in} \\
 \mathbf{C}' & \xrightarrow{F} & \mathbf{C}
 \end{array}$$

which commutes on the nose. Since $\mathbf{C}'_{\text{det}}$ is cartesian monoidal, $\mathbf{C}'$ inherits canonical comonoids satisfying the compatibility with the monoidal structure of (2.7), and the monoidal unit $I \in \mathbf{C}'$ is terminal by construction. Hence $\mathbf{C}'$ is also a Markov category.

Since F_{det} is symmetric monoidal and F merely extends it to the non-deterministic morphisms, also F comes equipped with a symmetric monoidal structure, and it is a symmetric monoidal equivalence by virtue of being an equivalence and symmetric monoidal. The essential surjectivity criterion of Proposition 10.16 now holds because F_{det} is an equivalence. $\square$

This result guarantees us that applying the graphical language for symmetric monoidal categories [70], as we do throughout the paper, indeed gives results that apply to all Markov categories (as the graphical language depicts only the strict case).

10.18. Remark (cf. [64, Proposition 4.3]). If $\mathbf{C}$ is a Markov category and $X \in \mathbf{C}$, then a version of the Eckmann–Hilton argument shows that X carries a unique deterministic comonoid structure. For if we have any other deterministic comultiplication on X , drawn as a white dot, then we get



where the second equation is by the determinism assumption as well as (2.4), and we have used the assumption that I is terminal, which means that the two counits coincide.

This property is not particularly surprising, since the subcategory $\mathbf{C}_{\text{det}} \subseteq \mathbf{C}$ is cartesian monoidal (Remark 10.13). So we mention this mainly as preparation for the stronger observation made in the upcoming Remark 11.29.

10.19. Remark. The previous arguments indicate that it may be useful to think of the subcategory $\mathbf{C}_{\text{det}} \subseteq \mathbf{C}$ as the primary object which specifies a Markov category, together with some way of equipping this category with additional morphisms. This leads us to suspect that there could be a simple alternative definition of Markov categories as categories with finite products, modelling $\mathbf{C}_{\text{det}}$, equipped with a suitable kind of *arrow* [67], modelling $\mathbf{C}$. Similarly and essentially equivalently [67, Theorem 6.1], the inclusion func-

tor $\mathbf{C}_{\text{det}} \rightarrow \mathbf{C}$ can also be seen as a *commutative Freyd category* [84].¹⁶ In fact, it may be possible to generalize many of our methods and results to all semicartesian commutative Freyd categories. In our language, this amounts to starting with a pair of categories $(\mathbf{C}_{\text{det}}, \mathbf{C})$ where $\mathbf{C}_{\text{det}}$ is cartesian monoidal, $\mathbf{C}$ is symmetric semicartesian monoidal, and we are given a strong monoidal functor $\mathbf{C}_{\text{det}} \rightarrow \mathbf{C}$, which in particular equips $\mathbf{C}$ with affine comonoids, where the comultiplications are given as the $\mathbf{C}$ -images of the diagonals $X \rightarrow X \otimes X$ defined in $\mathbf{C}_{\text{det}}$. In this setting, deterministic morphisms and stochastic morphisms live in two separate categories, a setup proposed earlier by Golubtsov [54, Sections 2 and 3]¹⁷ with minor variations. For cartesian monoidal $\mathbf{D}$, the situation $\mathbf{D} \rightarrow \mathbf{Kl}(\mathbf{T})$ described in Corollary 3.2 gives a paradigmatic example.

There are advantages and disadvantages to such an approach. The main advantage is that one can work with the inclusion $\mathbf{Meas} \rightarrow \mathbf{Stoch}$, and consider the deterministic morphisms to be the morphisms of $\mathbf{Meas}$, thereby getting around the problem described in Example 10.4. A seemingly greater disadvantage is that one can no longer state and prove results such as Proposition 13.15, stating that retracts are almost surely deterministic.

11. Further candidate axioms for Markov categories

Traditionally, probability and statistics take place in some souped-up version of $\mathbf{FinStoch}$, namely some suitable category of measurable spaces and Markov kernels (Sections 4 and 5). Such a category has many additional properties which sometimes play an important role. In this section, we describe some of these properties, thinking of them as potential additional axioms that can make a Markov category into a category that “looks more like” a category of Markov kernels. However, our present investigations seem too preliminary to settle on any particular set of axioms, so that we will not make a definite choice. None of these axioms will play a central role in the rest of this paper, although we will have occasional use for them in the upcoming sections (where we make explicit mention when one of them is assumed).

Conditionals and conditioning. The first special property of probability theory is the possibility to do conditioning. The standard properties of conditional distributions are captured in the following definition, which is originally due to Golubtsov [54, Section 7.1] and has been rediscovered more recently by Cho and Jacobs [15, Section 3], who called it *admitting disintegration*.

11.1. Definition. Let $\mathbf{C}$ be a Markov category. We say that $\mathbf{C}$ *has conditional distributions* if for every morphism $\psi : I \rightarrow X \otimes Y$, there is $\psi|_X : X \rightarrow Y$ such that

¹⁶ This has been pointed out to us by Sam Staton.

¹⁷ See also Golubtsov’s [50, Definition 7] for where the same definition has already been put forward earlier in the context of a specific category.

$$(11.1)$$

In terms of Notation 2.8, this equation looks exactly like the familiar defining equation of conditional distributions in the discrete case, which is the *chain rule*

$$\psi(x, y) = \psi_{|X}(y|x) \psi(x). \quad (11.2)$$

The difference is that Notation 2.8 lets us interpret this equation not only in the discrete case, but in *any* Markov category, including in categories like **Stoch** or **BorelStoch** concerned with measure-theoretic probability. Note that in this case, this equation is *not* an equation about densities, and x and y are *not* elements of the measurable spaces involved, but rather mere formal variables which encode the wiring structure of (11.1) in symbolic form. The correct interpretation of this equation in measure-theoretic probability is the upcoming (11.3).

11.2. Example. In **FinStoch**, we can interpret the formal variables x and y as actual elements. It follows that **FinStoch** has conditional distributions given by

$$\psi_{|X}(y|x) = \frac{\psi(x, y)}{\psi(x)},$$

in case that $x \in X$ is such that $\psi(x) > 0$, which is the usual defining equation for conditional probabilities. In case that $\psi(x) = 0$, we choose any set of values $\psi_{|X}(y|x) \in [0, 1]$ satisfying the required normalization condition $\sum_y \psi_{|X}(y|x) = 1$. Now indeed the relevant equation (11.2) holds by definition, and it can be interpreted either formally in the sense of Notation 2.8 as a different way of writing (11.4), or now equivalently as an actual equation involving real numbers and multiplication.

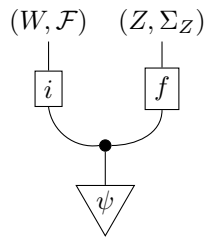
11.3. Example. What are conditional distributions in **Stoch**? For a given probability measure ψ on a product σ -algebra $\Sigma_X \otimes \Sigma_Y$, a conditional distribution is a Markov kernel $\psi_{|X} : (X, \Sigma_X) \rightarrow (Y, \Sigma_Y)$ satisfying the equation

$$\psi(S \times T) = \int_{x \in S} \psi_{|X}(T|x) \psi(dx) \quad (11.3)$$

for all $S \in \Sigma_X$ and $T \in \Sigma_Y$, since this is enough to prove the equation (11.1) by Lemma 4.2. Such a $\psi_{|X}$ is sometimes called a *product regular conditional probability*,

and is known to exist whenever (Y, Σ_Y) is standard Borel, but not in general [31]. Thus **BorelStoch** has conditional distributions, but **Stoch** does not.

A more commonly studied concept in probability and statistics are *regular conditional distributions*, defined as follows [75, Definition 7.28]. Let (W, Σ_W, ψ) be a probability space and $f : (W, \Sigma_W) \rightarrow (Z, \Sigma_Z)$ a measurable function (or “random variable”) taking values in some other measurable space (Z, Σ_Z) . Suppose that $\mathcal{F} \subseteq \Sigma_W$ is a sub- σ -algebra. Then the identity map $i : (W, \Sigma_W) \rightarrow (W, \mathcal{F})$ is measurable, and therefore also a (deterministic) morphism in **Stoch**. Then we can consider the joint probability measure on $(W, \mathcal{F})$ and (Z, Σ_Z) given by



A *regular conditional distribution* in the traditional sense is then the same as a conditional distribution in our sense, instantiated in **Stoch** on this particular joint distribution. Again, it is a standard fact that such regular conditional distributions exist whenever (Z, Σ_Z) is standard Borel [75, Theorem 8.37], [39, 452N], but not in general [7, Exercise 33.13].

Conversely, a generic conditional distribution in our sense can be constructed as a regular conditional probability by putting

$$W := X \times Y, \quad Z := Y,$$

with the obvious σ -algebras, as well as taking f to be the projection $X \times Y \rightarrow Y$ and $\mathcal{F}$ to be given by the sub- σ -algebra $\Sigma_X \subseteq \Sigma_X \otimes \Sigma_Y$. Thus conditional distributions in our sense and regular conditional distributions in the usual sense are mutual special cases of each other.

In conclusion, **Stoch** does not have conditional distributions in general, but **BorelStoch** does. Furthermore, the existence of conditional distributions in our sense is equivalent to the existence of regular conditional distributions in the usual sense.

11.4. Example. The Kleisli category of the Radon monad $\mathbf{Kl}(R)$ (Section 5) does not have conditional distributions, for the simple reason that requiring the conditional $X \rightarrow Y$ to be represented by a continuous function $X \rightarrow RY$ is too much to ask for: conditionals are rarely continuous.

However, Definition 11.1 is not yet the definition that seems most natural to us, and we will next present a more general definition which turns out to be more useful for our purposes. For motivation, consider a morphism $\psi : I \rightarrow X \otimes Y \otimes Z$, which we may think of as a joint distribution of three random variables. Then it ought to be possible to obtain a conditional $\psi_{|X \otimes Y} : X \otimes Y \rightarrow Z$ by first conditioning on X and then conditioning on Y . However, the second step is *not* a form of conditioning in the sense of Definition 11.1, since only conditioning on morphisms with domain I is allowed, while we would like to condition $\psi_{|X} : X \rightarrow Y \otimes Z$ on Y . Using the following improved definition, we will prove this iterated conditioning property in Lemma 11.11.

11.5. Definition. Let $\mathbf{C}$ be a Markov category. We say that $\mathbf{C}$ has *conditionals* if for every morphism $f : A \rightarrow X \otimes Y$, there is $f_{|X} : X \otimes A \rightarrow Y$ such that

The diagram shows an equality between two expressions. On the left, a box labeled f has two input wires labeled X and Y at the top, and one output wire labeled A at the bottom. On the right, a more complex diagram is shown. It features a box labeled $f_{|X}$ at the top with two input wires labeled X and Y at the top. The Y wire goes down into a box labeled f . The X wire goes down, loops around the left, and then goes into the top of the f box. The output of the f box is a wire labeled A at the bottom. There are also two small black dots on the wires: one on the Y wire just above the f box, and one on the A wire just below the f box.

$=$

(11.4)

This property specializes to the property of Definition 11.1 by setting $A = I$.

In terms of Notation 2.8, equation (11.4) becomes a conditional version of the chain rule,

$$f(x, y|a) = f_{|X}(y|x, a)f(x|a),$$

which again is familiar from discrete probability, but now becomes valid in any Markov category. If $\mathbf{C}$ has conditionals, then we similarly get a conditional $f_{|Y}(x|y, a)$ by suitable permutations of the wires. This results in *Bayes' theorem* in the generalized form

$$f_{|X}(y|x, a)f(x|a) = f_{|Y}(x|y, a)f(y|a),$$

where the special case $A = I$ has been treated by Cho and Jacobs [15], and appears also in similar form and somewhat implicitly in earlier work of Golubtsov [54, Theorem 6].¹⁸

In typical examples, permitting A to be general in the definition of conditionals amounts to a certain regularity condition on the formation of conditionals: if the joint distribution of X and Y depends suitably nicely on A so as to form a morphism in $\mathbf{C}$, then there also is a conditional $f_{|X}$ which depends on A suitably nicely so as to form a morphism in $\mathbf{C}$.

11.6. Example. **FinStoch** has conditionals. Similar to Example 11.2, for a given morphism $f(x, y|a)$ we form

$$f_{|X}(y|x, a) := \frac{f(x, y|a)}{f(x|a)}, \quad (11.5)$$

in case that $x \in X$ and $a \in A$ are such that $f(x|a) > 0$, which again is essentially the usual defining equation for conditional probabilities, and again arbitrary values satisfying the normalization $\sum_y f_{|X}(y|x, a) = 1$ otherwise.

11.7. Example. Continuing from Example 11.3, it was recently proven by Bogachev and Malofeev that **BorelStoch** has conditionals [8, Theorem 3.5].¹⁹ In other words, a joint distribution which depends measurably on another parameter $a \in A$ has a conditional whose dependence on $a \in A$ is still measurable. This is obvious when the σ -algebra Σ_A is discrete, since then this extra measurability requirement is trivial.

11.8. Example. We now investigate whether **Gauss** (Section 6) has conditionals. A morphism $A \rightarrow X \otimes Y$ can conveniently be represented in the block form

$$\left(\begin{pmatrix} M \\ N \end{pmatrix}, \begin{pmatrix} C_{\xi\xi} & C_{\xi\eta} \\ C_{\eta\xi} & C_{\eta\eta} \end{pmatrix}, \begin{pmatrix} s \\ t \end{pmatrix} \right),$$

meaning that we have

$$X = MA + \xi, \quad Y = NA + \eta,$$

where now ξ and η are potentially *correlated* random variables, with the specified expectations and covariance matrices. Using well-known results on conditionals of multivariate Gaussians [30, Proposition 3.13], we find that the conditional distribution of Y given X can itself be written in the form

$$Y = C_{\eta\xi}C_{\xi\xi}^-X + (N - C_{\eta\xi}C_{\xi\xi}^-M)A + \tau,$$

¹⁸ See also [50, Theorem 4] for an even earlier reference where Golubtsov's considerations already appear in the context of one particular category.

¹⁹ We thank Kenta Cho for pointing this out to us.

where τ is normal with expectation $\mathbf{E}[\tau] := t - C_{\eta\xi}C_{\xi\xi}^-s$ and variance

$$\mathbf{Var}[\tau] := C_{\eta\eta} - C_{\eta\xi}C_{\xi\xi}^-C_{\xi\eta},$$

and $C_{\xi\xi}^-$ is the Moore–Penrose pseudoinverse of $C_{\xi\xi}$. This matrix satisfies the relevant equations

$$C_{\xi\xi}^-C_{\xi\xi}C_{\xi\xi}^- = C_{\xi\xi}^-, \quad C_{\eta\xi}C_{\xi\xi}^-C_{\xi\xi} = C_{\eta\xi}, \quad C_{\xi\xi}C_{\xi\xi}^-C_{\xi\eta} = C_{\xi\eta},$$

where the proof of the latter two employs standard properties of positive semidefinite matrices.²⁰ The positive semidefiniteness of the above expression for $\mathbf{Var}[\tau]$ follows from the fact that it is a matrix of the form $v^T C v$, where C is the original covariance matrix in the above block form and

$$v = \begin{pmatrix} C_{\xi\xi}^-C_{\xi\eta} \\ -1 \end{pmatrix}.$$

The dependence of Y on X and A therefore corresponds to the morphism $X \otimes A \rightarrow Y$ represented by the triple

$$\left((C_{\eta\xi}C_{\xi\xi}^- \quad N - C_{\eta\xi}C_{\xi\xi}^-M), C_{\eta\eta} - C_{\eta\xi}C_{\xi\xi}^-C_{\xi\eta}, t - C_{\eta\xi}C_{\xi\xi}^-s \right),$$

reproducing formulas due to Lauritzen and Jensen [78, Section 4.5], which generalize Golubtsov’s earlier [51, Theorem 4]. A straightforward but tedious computation using the Markov category structure introduced above shows that this morphism is indeed the desired conditional in the sense of Definition 11.5. Therefore Gauss has conditionals.

We now present some open problems before developing the theory of conditionals further.

11.9. Problem. Suppose that $\mathbf{C}$ is a Markov category which has conditionals, and $\mathbf{D}$ a small category. Then does the functor category $\mathbf{Fun}(\mathbf{D}, \mathbf{C})$ from Proposition 7.1 have conditionals as well?

This is not obvious, because if we try to form the conditional of a morphism of diagrams by doing so separately at each object of $\mathbf{D}$, then there is no guarantee that we again obtain a morphism of diagrams. A possible way around might be to relax the notion of morphism of diagrams in some way with respect to almost sure equality (Section 13).

²⁰ They are simple to prove in a basis where $C_{\xi\xi}$ is diagonal: if there are zeros on the diagonal of $C_{\xi\xi}$, then positive semidefiniteness of C implies that the relevant rows and columns of $C_{\eta\xi}$ and $C_{\xi\eta}$ must vanish as well and hence can be ignored, reducing the problem to the case where $C_{\xi\xi}$ is invertible.

In measure-theoretic probability, conditioning is also often formalized in terms of *conditional expectations*. We suspect that conditional expectations indeed arise from Definition 11.5 as well as follows.

11.10. Conjecture. *There is a Markov category such that:*

- ▷ *Objects are measurable spaces (X, Σ_X) , or some variant thereof such as measure algebras [38].*
- ▷ *Morphisms $(X, \Sigma_X) \rightarrow (Y, \Sigma_Y)$ are some version of positive unital linear maps between spaces of bounded measurable functions in the other direction, $\mathcal{L}^\infty(Y, \Sigma_Y) \rightarrow \mathcal{L}^\infty(X, \Sigma_X)$, taking bounded directed suprema to bounded directed suprema.*
- ▷ *The monoidal structure is given by some variant of the usual product of measurable spaces.*
- ▷ *The comultiplications $X \rightarrow X \otimes X$ are implemented by restriction to the diagonal, $\mathcal{L}^\infty(X \times X, \Sigma_X \otimes \Sigma_X) \rightarrow \mathcal{L}^\infty(X, \Sigma_X)$.*
- ▷ *Conditionals exist in the sense of Definition 11.5, and are given by conditional expectations.*

One candidate for such a Markov category may be the category described in Example 4.4.

Returning to the general theory, sometimes we may want to condition on more than one output. The following guarantees that this can be done either all at once or step by step:

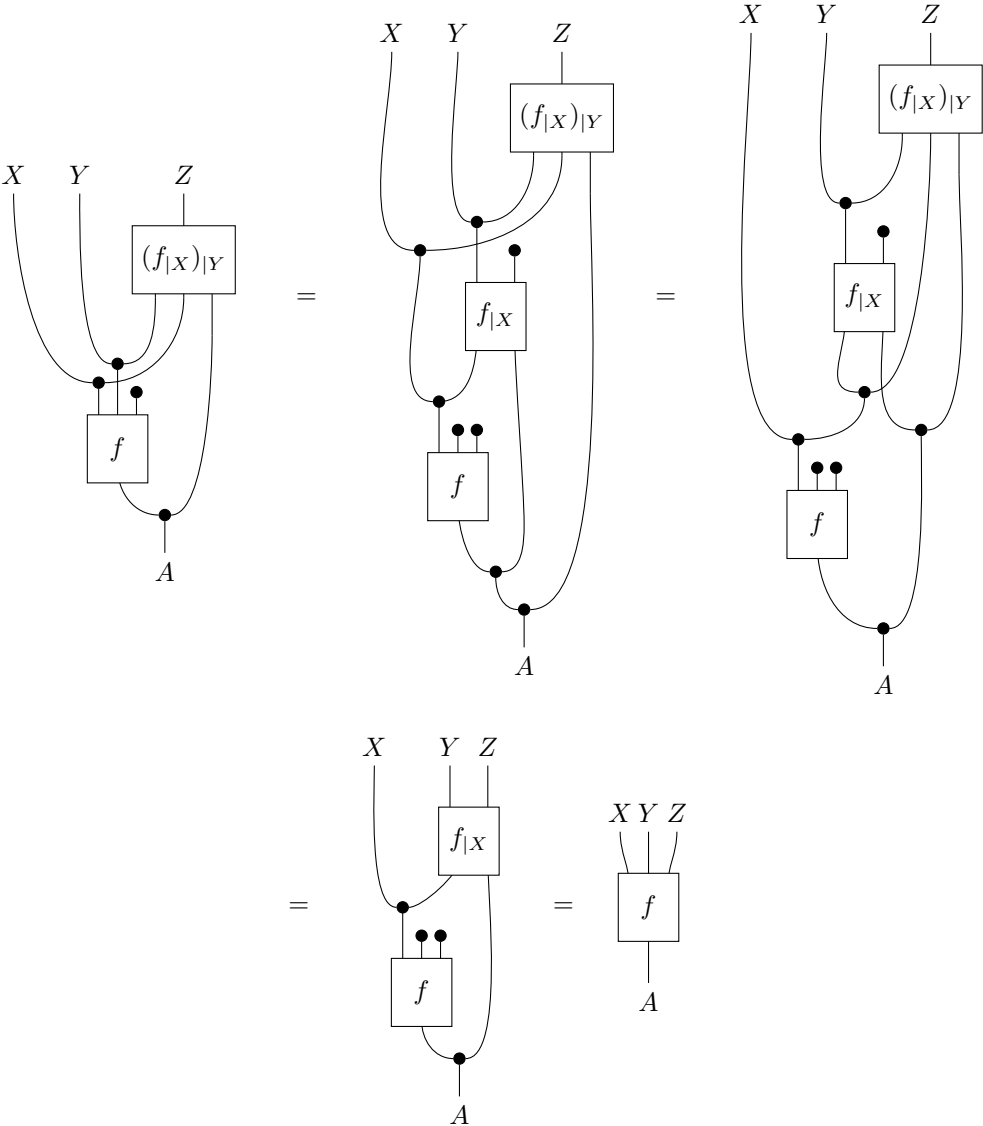
11.11. Lemma. *Suppose that $\mathbf{C}$ has conditionals. Then for any morphism $f : A \rightarrow X \otimes Y \otimes Z$, we have*

$$\begin{array}{c} Z \\ \downarrow \\ \boxed{f_{|X \otimes Y}} \\ \downarrow \downarrow \downarrow \\ X \quad Y \quad A \end{array} = \begin{array}{c} Z \\ \downarrow \\ \boxed{(f_{|X})_{|Y}} \\ \downarrow \downarrow \downarrow \\ X \quad Y \quad A \end{array}$$

in the sense that every double conditional $(f_{|X})_{|Y}$ as on the right is also a conditional $f_{|X \otimes Y}$ as on the left (with first two inputs swapped).

Note that this result can also be applied iteratively in order to compute a conditional $f_{|X_1 \otimes \dots \otimes X_n}$ of a morphism $f : A \rightarrow X_1 \otimes \dots \otimes X_n \otimes Y$. So far we have not found any formulation and proof of a converse, meaning a way to say that every conditional $f_{|X \otimes Y}$ is also a double conditional $(f_{|X})_{|Y}$ (with first two inputs swapped).

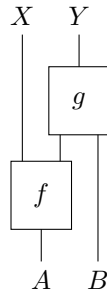
Proof. We start with the chosen $(f|_X)|_Y$ and compute



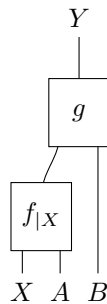
which shows indeed that the given $(f|_X)|_Y$ is also a conditional $f|_{X \otimes Y}$, modulo swapping the first two inputs. $\square$

Conditioning also enjoys a certain compositionality property:

11.12. Lemma. *If $f : A \rightarrow X \otimes W$ and $g : W \otimes B \rightarrow Y$ are arbitrary morphisms and $f|_X : X \otimes A \rightarrow W$ is a corresponding conditional, then the composite*



has a conditional given by



Proof. Straightforward string diagram computation. $\square$

11.13. Remark. As a special case, we obtain that conditioning commutes with marginalization as follows. Using Notation 2.8, any $f : A \rightarrow X \otimes Y \otimes Z$ is such that the marginal $f|_X(y|xa)$ of the conditional $f|_X(yz|xa)$ is also a conditional $f(y|xa)|_X$ for the marginal $f(xy|a)$. This follows from Lemma 11.12 upon taking $B := I$ and $g : Y \otimes Z \rightarrow Y$ to be the projection.

11.14. Remark. Besides the previous two lemmas, we have not been able to identify any further potentially useful categorical properties of conditioning. Our original goal in this work had been to axiomatize conditioning as an operation on a category satisfying certain naturality conditions. We have abandoned this due to the lack of further candidate axioms for conditioning which would hold e.g. in $\mathbf{FinStoch}$, and the well-known problem that conditionals are generally non-unique in cases where zero probabilities occur (Proposition 11.15), leading us to suspect that a reasonably coherent choice of conditionals is not possible. We have been all the more surprised to realize that conditioning is also not a very relevant concept for most of the developments of this paper.

Within our framework, it is impossible to guarantee the uniqueness of conditionals except in uninteresting degenerate cases, because of the following observation.

11.15. Proposition. *Suppose that $\mathcal{C}$ has conditionals. Then conditionals in $\mathcal{C}$ are unique if and only if any two parallel morphisms in $\mathcal{C}$ are equal.*

Proof. The “if” direction is trivial, so that we focus on the “only if” part. Assuming the uniqueness of conditionals, we first prove that for every $Y \in \mathbf{C}$, we have an equality of morphisms $Y \otimes Y \rightarrow Y$ given by

$$\begin{array}{c} | \\ \bullet \\ | \end{array} = \begin{array}{c} | \\ \bullet \\ | \end{array}$$

Indeed, a straightforward computation using the comonoid axioms (2.2) and (2.3) shows that both sides are conditionals for the comultiplication

$$\begin{array}{c} \cup \\ \bullet \\ | \end{array}$$

Now for any two parallel morphisms $f, g : X \rightarrow Y$, we get

$$\begin{array}{c} | \\ \boxed{f} \\ | \end{array} = \begin{array}{c} | \quad \bullet \\ \boxed{f} \quad \boxed{g} \\ \cup \\ \bullet \\ | \end{array} = \begin{array}{c} \bullet \quad | \\ \boxed{f} \quad \boxed{g} \\ \cup \\ \bullet \\ | \end{array} = \begin{array}{c} | \\ \boxed{g} \\ | \end{array}$$

as was to be shown. $\square$

11.16. Remark. The equality of parallel morphisms, $f = g$ for $f, g : X \rightarrow Y$, means that $\mathbf{C}$ is equivalent to a preordered set: the category structure is uniquely determined by writing $X \leq Y$ if and only if there is a (necessarily unique) morphism $X \rightarrow Y$. Then the comonoid structures of Definition 2.1 show that the tensor product of objects must be a meet operation in the sense of lattice theory, making $\mathbf{C}$ into a meet-semilattice with neutral element $I \in \mathbf{C}$. Conversely, upon considering every meet-semilattice as a monoidal category, it also is a Markov category which has unique conditionals.

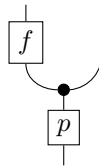
The existence of conditionals implies a stronger disintegration property, which we will relate to the standard notion of disintegration in measure-theoretic probability in Example 11.18.

11.17. Proposition (Disintegration). Suppose that $\mathbf{C}$ has conditionals. Then for every $p : A \rightarrow X$ and $f : X \rightarrow Y$, there is $s : A \otimes Y \rightarrow X$ such that

$$(11.6)$$

For $A = I$ and deterministic f , this equation implies the almost sure equality $fs =_{f p\text{-a.s.}} \text{id}_Y$ (Section 13) by composing with f on the right output.

Proof. Define s to be a conditional of the morphism



on Y . This means that s satisfies the equation

which immediately implies the claim. $\square$

11.18. Example. In *Stoch*, setting $A = I$ in Proposition 11.17 and taking f to be given by a measure-preserving function of measurable spaces $f : X \rightarrow Y$ specializes the statement to the usual notion of disintegration of a probability measure p on X with respect to a

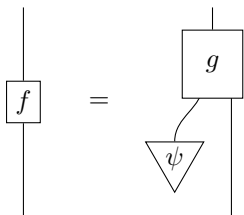
measure-preserving function $f : X \rightarrow Y$.²¹ Indeed under these assumptions, (11.6) states that

$$p(f^{-1}(T) \cap S) \stackrel{!}{=} \int_{y \in T} s(S|y) (fp)(dy) = \int_{y \in T} s(S|y) p(f^{-1}(dy))$$

for all $S \in \Sigma_X$ and $T \in \Sigma_Y$. This is exactly the usual notion of *regular conditional probability*. It is well-known that regular conditional probabilities do not exist in general [28, p. 624]. We can therefore conclude, by the contrapositive of Proposition 11.17, that *Stoch* does not have conditionals, as we had already noted in Example 11.3.

Randomness pushback. We move on to our second candidate axiom. It is inspired by the treatment of random mappings in [69, Example 2.6(3)], and a similar definition has been given as [22, Definition 8]. For *FinStoch*, it is well-known that randomness can be “pushed back”, in the sense that every Markov kernel can be replaced by a deterministic map with one extra input into which one feeds a random value with a fixed distribution.

11.19. Definition (*Randomness pushback*). We say that $\mathbb{C}$ has *randomness pushback* if for every morphism $f : X \rightarrow Y$ in $\mathbb{C}$, there are $A \in \mathbb{C}$, deterministic $g : A \otimes X \rightarrow Y$ and $\psi : I \rightarrow A$ such that


(11.7)

While this is an interesting property, we will have no use for it in this paper. In certain other contexts, it may be useful to apply it repeatedly. This happens for example in the theory of Bayesian networks, where one can use it to prove the equivalence between Bayesian networks with stochastic dependencies, and networks with purely functional (i.e. deterministic) dependencies plus extra “exogenous” root variables which supply the randomness [29]. Our new formal perspective should have the potential to shed new light on technical problems with measure-theoretic aspects in these contexts, such as [40, Conjecture 4.5].

As the following examples show more concretely, for $Y = X$ randomness pushback essentially specializes to the problem of representing a stochastic dynamical system in terms of a random choice of deterministic dynamics [71, Theorem 1.1].

²¹ For example, we may consider a sub- σ -algebra Σ'_X and consider the measurable map $f := \text{id}_X : (X, \Sigma_X) \rightarrow (X, \Sigma'_X)$, thereby recovering the notion of disintegration with respect to a sub- σ -algebra.

11.20. Example. **FinStoch** has randomness pushback. For given f , take $A := Y^X$ to be the finite set of functions from X to Y , and define $\psi : I \rightarrow A$ by declaring the probability weight of any function $\eta \in Y^X$ to be given by

$$\psi(\eta) := \prod_{x \in X} f(\eta(x)|x),$$

so that all the individual values $\eta(x)$ are stochastically independent and have the same distribution as $f(y|x)$.²² Using the deterministic function $g : Y^X \times X \rightarrow Y$ given by evaluation,

$$g(\eta, x) := \eta(x),$$

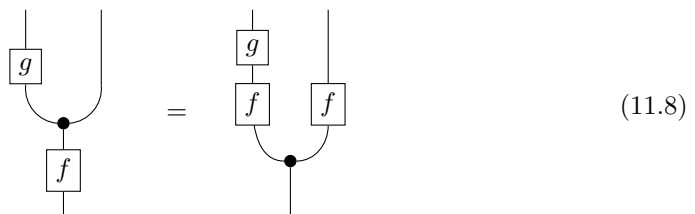
then clearly results in (11.7).

We do not know whether **Stoch** has randomness pushback in general. One obstacle in generalizing the above argument from **FinStoch** to **Stoch** is the fact that **Meas** is not cartesian closed [1].

11.21. Example. We now show that **Gauss** also has randomness pushback. Any morphism $(M, C, s) : n \rightarrow m$, representing $Y = MX + \xi$, can be obtained by plugging in the morphism $\psi : 0 \rightarrow m$ with parameters $(0, C, s)$ into the addition map $+: m + m \rightarrow m$, which is the deterministic morphism represented by $((1 \ 1), 0, 0)$.

Positivity. Our third candidate axiom is a bit more mysterious. We will show that it holds in our running examples, and then make crucial use of it to prove our version of the Fisher–Neyman factorization theorem (Theorem 14.5).

11.22. Definition. We say that **C** is *positive* if whenever $f : X \rightarrow Y$ and $g : Y \rightarrow Z$ are such that gf is deterministic, then also



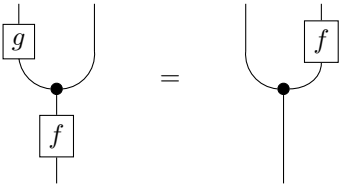
$$(11.8)$$

In terms of the upcoming Definition 12.12, this equation can be understood as saying that the morphism shown on the left-hand side displays the conditional independence $Y \perp Z \parallel X$. Intuitively, this equation means the following: if a deterministic process (first

²² Any other joint distribution with the same marginals on the Y -factors of Y^X will work just as well.

output) has a random intermediate result (second output), then the same distribution over both results can also be obtained by computing the intermediate result independently of the deterministic process.

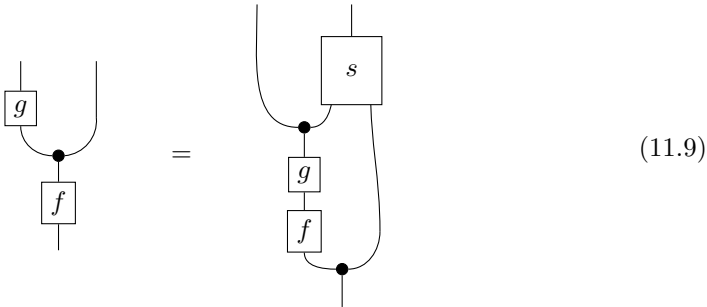
11.23. Remark. A relevant special case which comes up frequently is with $gf = \text{id}$, in which case we get



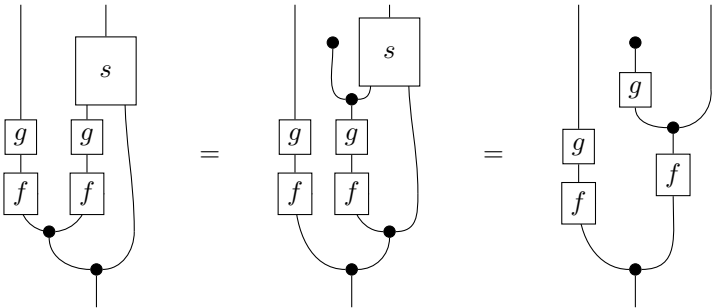
The following result implies that many of our example Markov categories are indeed positive, in particular **FinStoch** and **Gauss**.

11.24. Lemma. *If $\mathcal{C}$ has conditionals, then $\mathcal{C}$ is positive.*

Proof. By Proposition 11.17, we have s such that



The determinism assumption evaluates this further to



where the second step is another use of (11.9). This implies the claim. $\square$

The term “positive” in Definition 11.22 is motivated by the crucial role of the non-negativity of probabilities in the following arguments.

11.25. Example. *Stoch* is positive, as we now show. By the characterization of deterministic morphisms in Example 10.4, the assumption is that

$$\int_{y \in Y} g(S|y) f(dy|x) \in \{0, 1\} \quad (11.10)$$

for all $S \in \Sigma_X$ and $x \in X$. By Lemma 4.2, what we want to prove is the equation

$$\int_{y \in T} g(S|y) f(dy|x) = f(T|y) \int_{y \in Y} g(S|y) f(dy|x) \quad (11.11)$$

for all $S \in \Sigma_Z$ and $T \in \Sigma_Y$ and $x \in X$.

It is easy to see that if the claimed equation holds for S , then it also holds for the complementary event $Z \setminus S$. Thus it is enough to consider the case where (11.10) is zero. Since an integral of a nonnegative function vanishes if and only if the integrand vanishes almost surely, we can conclude that the integrand $y \mapsto g(S|y)$ vanishes $f(-|x)$ -almost surely on all of Y . Therefore both sides of the claimed equation (11.11) vanish as well.

11.26. Remark. It is easy to see that if $C \rightarrow C'$ is a Markov embedding and C' is positive, then so is C . For example since *Stoch* is positive, the positivity of *Gauss* follows alternatively from the Markov embedding $\text{Gauss} \rightarrow \text{Stoch}$ of Proposition 6.1.

We now present a negative example.

11.27. Example. Let $\text{FinStoch}_\pm$ be defined just as *FinStoch*, but dropping the nonnegativity of the matrix entries, so that a morphism $f : X \rightarrow Y$ is a matrix $(f(y|x))_{x \in X, y \in Y}$ of arbitrary real numbers with $\sum_y f(y|x) = 1$ for all $x \in X$. In other words, negative probabilities are allowed. $\text{FinStoch}_\pm$ is easily seen to be a Markov category in the same way as *FinStoch* is. We can also obtain $\text{FinStoch}_\pm$ as a Markov category from the hypergraph category $\text{Mat}(\mathbb{R})$ via Section 8.

$\text{FinStoch}_\pm$ is not positive: a straightforward but slightly tedious computation shows that

$$f = \begin{pmatrix} 2 & 2 \\ -1 & 0 \\ 0 & -1 \end{pmatrix}, \quad g = \begin{pmatrix} \frac{1}{2} & 0 & 1 \\ \frac{1}{2} & 1 & 0 \end{pmatrix},$$

satisfy $gf = \text{id}$, but the equation of Remark 11.23 does not hold.

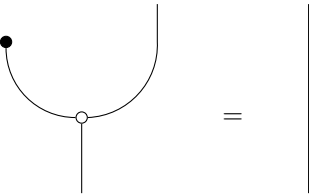
In particular, the contrapositive of Lemma 11.24 implies that $\text{FinStoch}_\pm$ does not have conditionals. This can also be seen directly by noting that the desired equation for conditional distributions, $\psi(x, y) = \psi|_X(y|x) \psi(x)$, has vanishing right-hand side

whenever $\sum_y \psi(x, y) = 0$, although the left-hand side does not need to vanish in this case.

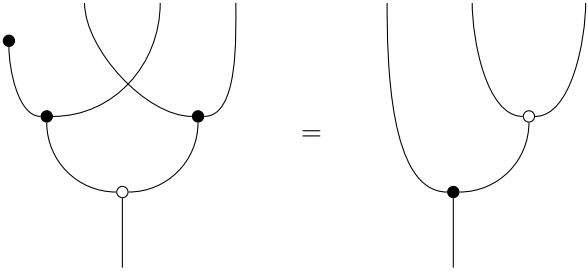
Let us return to the general theory.

11.28. Remark. If $\mathbf{C}$ is positive, then it is easy to see that every isomorphism is deterministic: just take g to be the inverse of the isomorphism f , and compose the equation of Remark 11.23 with $f \otimes \text{id}$. This is a property which fails in general Markov categories (Section 9).

11.29. Remark. If $\mathbf{C}$ is positive, then we also get a stronger version of Remark 10.18: every comonoid structure on every object $X \in \mathbf{C}$ is equal to the distinguished one. Namely, we have $gf = \text{id}$ if we take f to be the other comultiplication and g to be the marginalization over the second tensor factor,



Now applying the positivity condition of Remark 11.23 results in the equation



Marginalizing over the third output now shows that the two comultiplications must be equal.

While we have not worked out the details, we therefore hope that it may be possible to define positive Markov categories equivalently as semicartesian symmetric monoidal categories together with a symmetric monoidal equivalence to their category of internal comonoids. This comes close to describing them entirely in 2-monadic terms.

These considerations make it seem plausible to us that some version of the positivity condition could also be of interest in the study of hypergraph categories [72,34,36].

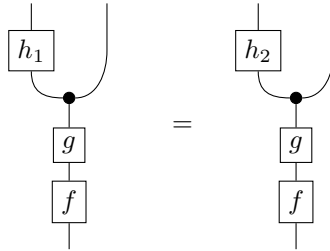
Using positivity, we can strengthen Lemma 10.9(b).

11.30. Lemma. Suppose that $\mathcal{C}$ is positive. Let $f : X \rightarrow Y$ and $g : Y \rightarrow Z$ in $\mathcal{C}$. If gf is deterministic and f is an epimorphism, then g is deterministic as well.

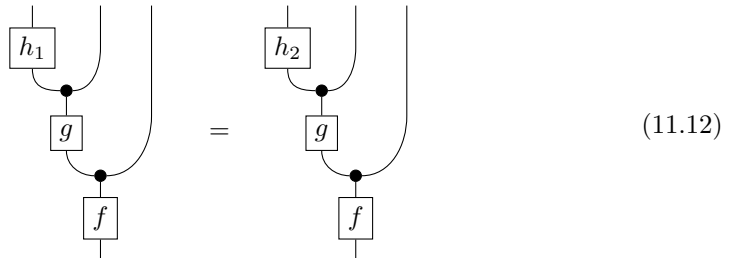
Proof. This is by the same computation as in the proof of Lemma 10.9, now using the positivity axiom for the second step. $\square$

We finally consider another property which will be of use in some of our later proofs. As we show in Remark 11.36, this axiom is closely related to the notion of *equality strengthening* introduced by Cho and Jacobs [15, Definition 5.7] (Remark 11.36).

11.31. Definition. We say that $\mathcal{C}$ is *causal* if the following condition holds: whenever



holds for any morphisms as indicated, then also



The name of this condition indicates that we think of it as a causality property: with time going up, we assume that the choice of using h_1 or h_2 at a late time has no influence on how its outcome is correlated with what happened at an earlier time; and then if causality holds, we can conclude that there is no correlation with what happened at an even earlier time either.

11.32. Remark. One may wonder how our axiom relates to the naturality condition on the counit (2.5), which is sometimes also called *causality* [19,14]. So which one of the two actually expresses causality, and correspondingly deserves to carry the name *causality axiom*? Our perspective is that they *both* capture disjoint aspects of the causality of information flow. It is an open question, suggested to us by Thomas C. Fraser, whether

there is an interesting causality axiom or axiom schema which comprises both as special cases, and which could be argued to capture all aspects of causality.

It may also be the case that our Definition 11.22 and Definition 11.31 are instances of a more comprehensive axiom comprising both as special cases. We have not yet found any such condition either.

The implication postulated by the causality axiom holds automatically in many cases: for example if f is deterministic; if g is deterministic; if f is an epimorphism; if $gf = g$; or if $\mathbf{C}$ has conditionals (Proposition 11.34). It is correspondingly nontrivial to find examples where our causality condition fails. Here is one.

11.33. Example. Applying Proposition 9.1 with $\mathbf{D} := \mathbf{Set}^{\text{op}}$ shows that the opposite of the category of commutative monoids with unit-preserving functions is a Markov category. Using additive notation for the monoid operations, Definition 11.31 now states the following: whenever f, g and h_1, h_2 are unit-preserving maps between commutative monoids, and if the equation

$$f(g(h_1(x) + y)) = f(g(h_2(x) + y)) \quad (11.13)$$

holds for all x and y , then also the stronger equation

$$f(g(h_1(x) + y) + z) = f(g(h_2(x) + y) + z) \quad (11.14)$$

holds for all x, y and z . To see that this is not true in general, let us specialize to the monoid $(\mathbb{N}, +)$ and all maps of type $(\mathbb{N}, +) \rightarrow (\mathbb{N}, +)$, and consider more concretely the maps given by

$$f(x) := \max(x - 1, 0), \quad g(x) := \min(x, 1), \quad h_1(x) := x, \quad h_2(x) := 0,$$

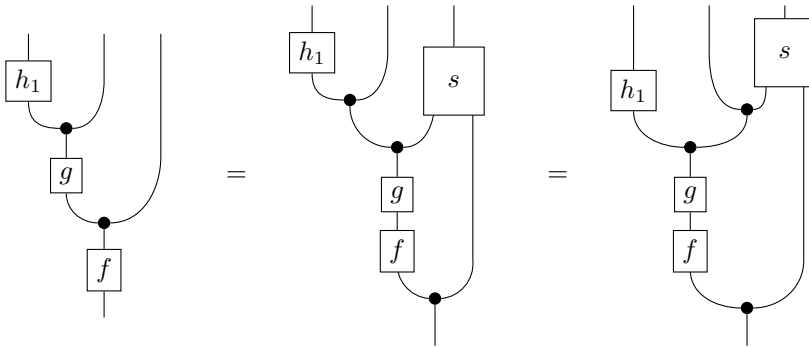
which clearly all preserve the unit 0. Then $fg = 0$, so that the assumption (11.13) holds trivially. However, the conclusion (11.14) fails with $x = z = 1$ and $y = 0$. We therefore have found a Markov category which is not causal.

Many uses of the causality condition use it only in a weaker form, where the middle output in (11.12) is marginalized over; see e.g. the proof of Proposition 13.9.

We now show that causality is intimately related to the existence of conditionals.

11.34. Proposition. *If $\mathbf{C}$ has conditionals, then $\mathbf{C}$ is causal.*

Proof. In terms of s as in the proof of Lemma 11.24,



Since h_2 can be replaced by h_2 on the right-hand side by assumption, the same holds true on the left-hand side. $\square$

It follows in particular that **FinStoch** and **Gauss** are causal. The next example shows in particular that causality does not imply the existence of conditionals, and is therefore strictly weaker.

11.35. Example. **Stoch** is causal despite not having conditionals (Example 11.18). Indeed, with $f : A \rightarrow X$ and $g : X \rightarrow Y$ and $h_1, h_2 : Y \rightarrow Z$, the assumption is that

$$\int_{x \in X} \int_{y \in T} h_1(U|y) g(dy|x) f(dx|a) = \int_{x \in X} \int_{y \in T} h_2(U|y) g(dy|x) f(dx|a) \quad (11.15)$$

for all $T \in \Sigma_Y$ and $U \in \Sigma_Z$ and $a \in A$. What we need to prove is that this equation implies a generalization where the domain of integration of x is a general $S \in \Sigma_X$,

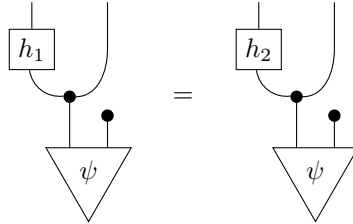
$$\int_{x \in S} \int_{y \in T} h_1(U|y) g(dy|x) f(dx|a) = \int_{x \in S} \int_{y \in T} h_2(U|y) g(dy|x) f(dx|a). \quad (11.16)$$

Writing the assumption (11.15) in the form

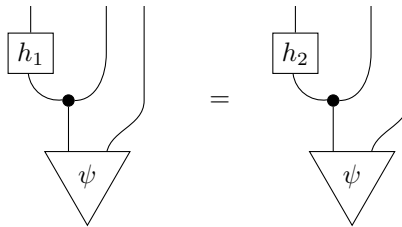
$$\int_{y \in T} h_1(U|y) \int_{x \in X} g(dy|x) f(dx|a) = \int_{y \in T} h_2(U|y) \int_{x \in X} g(dy|x) f(dx|a)$$

results in two functions on Y whose integrals over every $T \in \Sigma_Y$ are equal when integrated against the measure $\int_{x \in X} g(dy|x) f(dx|a)$. Hence the functions $y \mapsto h_1(U|y)$ and $y \mapsto h_2(U|y)$ are equal almost surely with respect to this measure. Therefore they are also equal almost everywhere with respect to any submeasure $\int_{x \in S} g(dy|x) f(dx|a)$ for arbitrary $S \in \Sigma_X$, which is (11.16). Hence **Stoch** is causal.

11.36. Remark. As pointed out to us by Kenta Cho, our causality condition is closely related to the notion of *equality strengthening* considered by Cho and Jacobs [15, Definition 5.7]. Indeed, equality strengthening means that every equation of the form



implies the stronger equation



Now if $\mathbf{C}$ is causal in the sense of Definition 11.31, then it also admits equality strengthening, as one can see readily by taking f to be ψ and g to be the marginalization map which discards the right output. Furthermore, if one generalizes equality strengthening by replacing the ψ above by an arbitrary morphism with nontrivial input, then it is straightforward to prove that causality and generalized equality strengthening are equivalent.

There is also an existing proof showing that the existence of conditionals implies equality strengthening in the original sense [15, Proposition 5.8], which now also follows from Proposition 11.34. Similarly, the proof that **Stoch** has equality strengthening given by Cho and Jacobs [15, Proposition 5.9] now follows from Example 11.35. On the other hand, Kenta Cho has informed us that equality strengthening is enough for the proof of our Proposition 13.9(a).

12. Conditional independence and the semigraphoid properties

Conditional independence is a central concept in probability, for example due to its importance for statistics [24] and its basic role in the definition of graphical model [77]. Generalizing conditional independence to arbitrary Markov categories is therefore a natural part of developing synthetic probability theory. Among other things, a formal treatment of conditional independence at our level of generality enables the interpretation of graphical models in *any* Markov category (although we will not elaborate on this further).

Traditionally, conditional independence is closely tied to conditioning. For example in **FinStoch**, a distribution $\psi : I \rightarrow X \otimes W \otimes Y$ is considered to display the conditional independence of X and Y given W if the conditional $\psi|_W(xy|w)$ satisfies the factorization

$$\psi|_W(xy|w) = \psi|_W(x|w) \psi|_W(y|w),$$

or a number of equivalent reformulations of this equation, such as $\psi(xyw)\psi(w) = \psi(xw)\psi(yw)$. In the context of Markov categories with conditional distributions, conditional independence has been defined by Cho and Jacobs [15], who proved that some of the usual equivalences hold at this level of generality, where they can be derived in terms of string diagrams. In [15, Proposition 6.10], Cho and Jacobs also showed that their generalized notion of conditional independence satisfies the well-known semigraphoid properties [96], [110, Section 2.2.2]. In a somewhat different technical setup, similar definitions had also been made earlier by Coecke and Spekkens [21, Section 5.1], who also proved the equivalence of various definitions [21, Proposition 5.3] under strong assumptions on the existence of their version of conditionals.

Here, we will show that it is still meaningful to introduce and work with conditional independence in the absence of conditionals. However, now the single notion of conditional independence splits into several distinct ones. For example, one may want to say that for a morphism $X \rightarrow Y \otimes Z$, the output Y is independent of the input X given the output Z (as in the definition of sufficient statistic, Definition 14.3). Here, the standard notion of conditional independence does not apply directly, since X does not come with any distribution; it therefore seems pertinent to distinguish this notion of conditional independence from the corresponding one for distributions $I \rightarrow X \otimes Y \otimes Z$. We believe that such distinctions may be able to shed further light on Dawid's analysis of the role of conditional independence in statistics [24], where distinct notions of conditional independence seem to be coming up.

In the following, we introduce the relevant definitions of conditional independence and study some of their properties, proving versions of the semigraphoid properties as well as certain compositionality properties, including conditional products. Our various definitions of conditional independence can be unified by formulating them in terms of disconnecting the relevant string diagrams (Remark 12.21). An interesting application, which we mostly leave to future work, should be to instantiate these definitions in **Stoch**, the category of measurable spaces and Markov kernels of Section 4, and to compare it in more detail with existing notions of conditional independence in measure-theoretic probability (Remark 12.4). One difficulty here is that **Stoch** does not have conditionals (Example 11.3). Instantiating our definitions and results in this case should give a theory of conditional independence for Markov kernels between products of arbitrary measurable spaces, which may be new (since conditional independence is usually primarily considered for joint distributions of three random variables, or equivalently for a triple of σ -algebras on a probability space).

Concretely, we start by introducing two simple notions of conditional independence, one denoted $X \perp Y | W$ for “distributions” $I \rightarrow X \otimes W \otimes Y$, and one denoted $X \perp Y || A$ for “processes” $A \rightarrow X \otimes Y$, and study some of their properties. Subsequently, we will show how to combine both notions into one more general notion of conditional independence for morphisms $A \rightarrow X \otimes W \otimes Y$.

12.1. Definition. Let $\mathbf{C}$ be a Markov category. A morphism $\psi : I \rightarrow X \otimes W \otimes Y$ in $\mathbf{C}$ displays the conditional independence $X \perp Y | W$ if there are $f : W \rightarrow X$, $g : W \rightarrow Y$ and $\phi : I \rightarrow W$ such that

The diagram shows an equality between two morphisms. On the left, a triangle labeled ψ has three input wires from the top. On the right, a morphism is shown where a triangle labeled ϕ has one input wire from the top. Two curved lines, labeled f and g , originate from the output of the ϕ triangle and connect to the inputs of two boxes labeled f and g respectively. The output of the f box is connected to the first output wire of the ψ triangle, and the output of the g box is connected to the third output wire of the ψ triangle. The middle output wire of the ψ triangle is connected to the output of the ϕ triangle. The equation is labeled (12.1) on the right.

12.2. Remark. In terms of Notation 2.8, marginalizing over the first and third output shows that ϕ must be equal to the marginal $\psi(w)$. Marginalizing over the third output only then shows that f must be a conditional $\psi(xw)_{|W}$ in the sense of Definition 11.5. Similarly, g must be a conditional $\psi(yw)_{|W}$.

Our definition essentially coincides with the characterization given by Cho and Jacobs in [15, Proposition 6.9(2)], who have shown that it is equivalent to the factorization $\psi_{|W}(xy|w) = \psi_{|W}(x|w)\psi_{|W}(y|w)$, in terms of Notation 2.8, if conditionals exist. In the case of $\mathbf{FinStoch}$, this implies in particular that Definition 12.1 is equivalent to the usual definition of conditional independence for probability distributions on products of finite sets.

Note that our definition applies regardless of whether $\mathbf{C}$ has conditionals or not. And of course, it also applies when X , W or Y are themselves tensor products of other objects. The latter recovers the fact that conditional independence as usually conceived applies to *sets* of random variables [110].

12.3. Remark. As a simple special case going back to Manes [88, Section 5] and Golubtsov [54, Section 7.1], for $W = I$ we get the obvious notion of plain *independence*: a morphism $\psi : I \rightarrow X \otimes Y$ displays the independence $X \perp Y$ if and only if

The diagram shows an equality between two morphisms. On the left, a triangle labeled ψ has two input wires from the top. On the right, two triangles labeled ψ are shown side-by-side. The first triangle has one input wire from the top and a black dot on its top boundary. The second triangle has one input wire from the top and a black dot on its top boundary. The equation is labeled with an equals sign between the two sides.

which, in terms of Notation 2.8, becomes exactly the usual factorization condition $\psi(x, y) = \psi(x)\psi(y)$.

12.4. Remark. It is an interesting problem to work out what Definition 12.1 instantiates to in *Stoch*. The following is based on discussion with Alex Simpson.

In *Stoch*, our definition directly coincides with the one given by Dawid and Studený via [27, 2.2.2 & Eq. (3)]; while their definition uses regular conditional probabilities in place of f and g , our f and g must also automatically be conditionals as noted in Remark 12.2.

A more commonly used definition of conditional independence in measure-theoretic probability is arguably the definition of conditional independence for a triple of σ -algebras on a probability space in terms of conditional expectation [17, Section 7.3]. Applying this definition in the case of a measure ψ on a product σ -algebra $\Sigma_X \otimes \Sigma_W \otimes \Sigma_Z$, using the three constituting factors, also gives a notion of conditional independence for $\psi : I \rightarrow X \otimes W \otimes Y$ in *Stoch*, in the form

$$\mathbb{E}[1_{S \cap T} | \Sigma_W] = \mathbb{E}[1_S | \Sigma_W] \mathbb{E}[1_T | \Sigma_W] \quad (12.2)$$

for all $S \in \Sigma_X$ and $T \in \Sigma_Y$. (The general definition involving a triple of σ -algebras is not really more general, since one can always work with the third power of the original space, equipped with the product of the three original σ -algebras, and equipping this product measurable space with the pushforward probability measure along the diagonal inclusion.) While we currently do not see how to arrive at the more commonly used definition (12.2) within our framework, we suspect that a Markov category as in Conjecture 11.10 would do exactly that.

Returning to the general theory, we can now derive versions of the four semigraphoid properties, which we state in terms of Notation 2.8.

12.5. Lemma. *The following hold:*

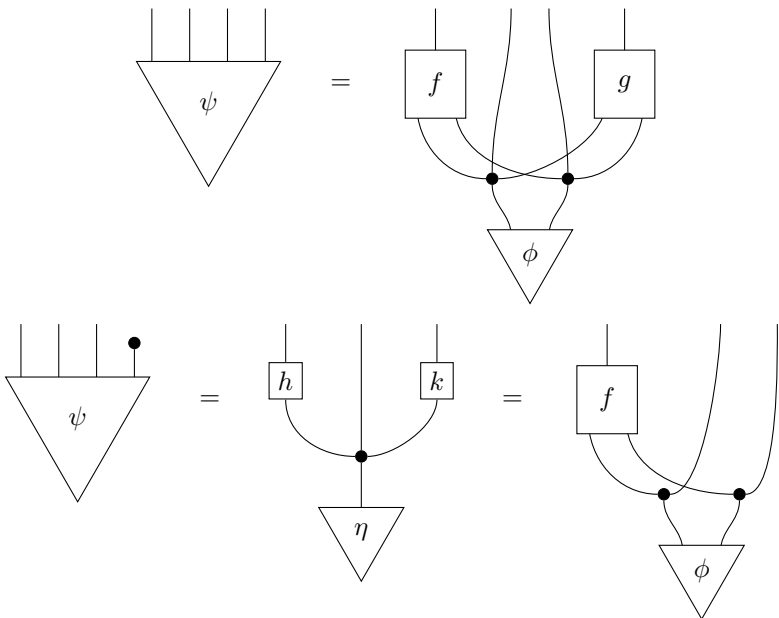
- (a) *Symmetry: if $\psi(x, w, y)$ displays $X \perp Y | W$, then $\psi(y, w, x)$ displays $Y \perp X | W$.*
- (b) *Decomposition: if $\psi(x_1, x_2, w, y)$ displays $(X_1 \otimes X_2) \perp Y | W$, then the marginal $\psi(x_1, w, y)$ displays $X_1 \perp Y | W$.*
- (c) *Contraction: if $\psi(x, w, z, y)$ displays $X \perp Y | (W \otimes Z)$ and the marginal $\psi(x, w, z)$ displays $X \perp Z | W$, then ψ also displays $X \perp (Z \otimes Y) | W$.*

Moreover if $\mathbf{C}$ has conditionals, then also:

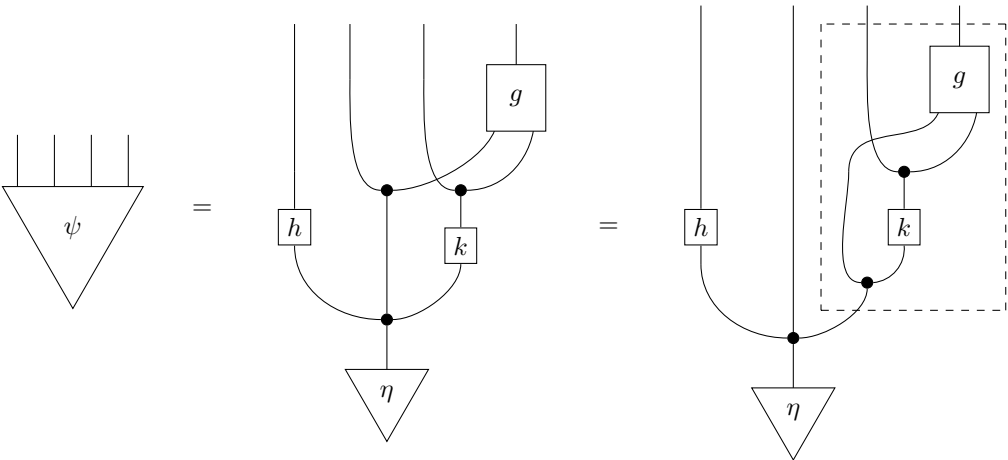
- (d) *Weak union: if $\psi(x_1, x_2, w, y)$ displays $(X_1 \otimes X_2) \perp Y | W$, then it also displays $X_1 \perp Y | (W \otimes X_2)$.*

12.6. Remark. On first look, these statements may sound more complicated than the usual semigraphoid properties, since they state explicitly *which morphism* displays the conditional independence, where this morphism is taken to be the given joint distribution of all variables under consideration. But this is implicitly contained also in the standard way of phrasing the semigraphoid properties, and our formulation merely makes this more explicit.

Proof. The first two statements are trivial. Concerning (c), we have the assumptions

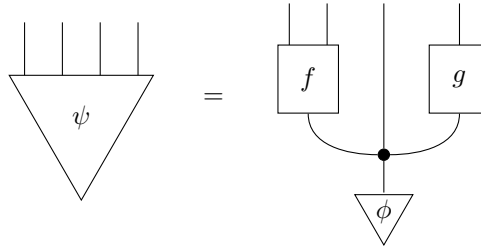


for suitable additional morphisms as indicated, where the final equation follows by marginalizing the first one. Substituting this third equation into the first one gives

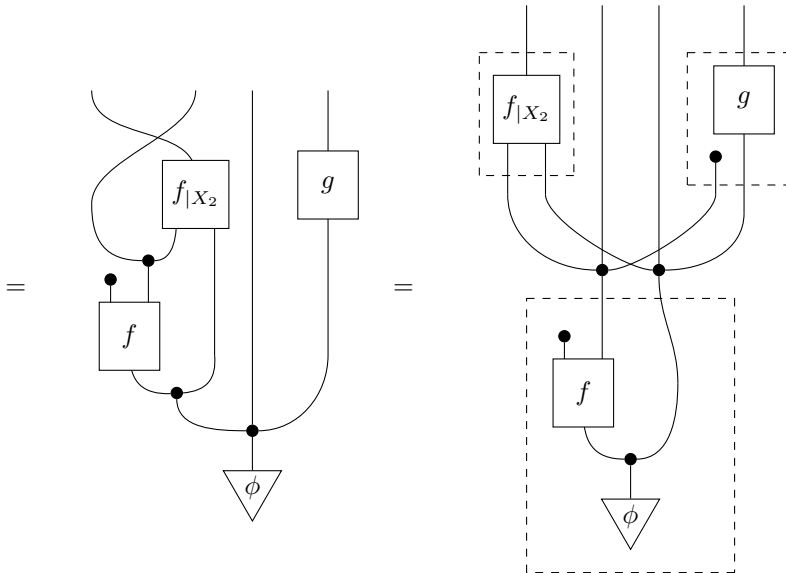


which proves the claim.

We now show (d). By the assumption that $\psi(x_1, x_2, w, y)$ displays $(X_1 \otimes X_2) \perp Y \mid W$, we have $\phi : I \rightarrow W$ and $f : W \rightarrow X_1 \otimes X_2$ and $g : W \rightarrow Y$ such that



By conditioning f on the second output, we can also write this in the form



where now the dashed boxes indicate the desired conditional independence factorization. $\square$

12.7. Problem. Beyond the semigraphoid properties of Lemma 12.5, does our general notion of conditional independence also satisfy the additional implication between conditional independences in probability found by Studený [109]?

An interesting compositionality property of joint distributions is the possibility of constructing *conditional products*, as first defined by Dawid and Studený [27] and generalized in a sheaf-theory-like categorical framework in [32]. We now show how to construct conditional products in any Markov category which has conditionals, and we prove that they

satisfy the axioms of Dawid and Studený. To this end, we work with the conceptually clearer but equivalent framework of *gleaves on a lattice* [32, Section 4.1].

12.8. Definition. Suppose that $\mathcal{C}$ has conditionals. Then for any $\psi : I \rightarrow X \otimes W$ and $\phi : I \rightarrow W \otimes Y$ with equal marginals on W ,

$$\triangle_{\pi} := \triangle_{\psi} = \triangle_{\phi}$$

the *conditional product* of ψ and ϕ is defined by

$$\triangle_{\psi \otimes_W \phi} := \text{Diagram with } \psi|_W, \phi|_W, \text{ and } \pi \quad (12.3)$$

The calculation

$$\text{Diagram 1} = \text{Diagram 2} = \text{Diagram 3} \quad (12.4)$$

shows that the conditional product does not depend on the particular choice of conditional $\psi|_W$, and similarly for $\phi|_W$ (compare Proposition 13.7). Hence $\psi \otimes_W \phi$ is well-defined.

The conditional product forms part of our story on conditional independence as follows.

12.9. Proposition. If $\mathcal{C}$ has conditionals, then the conditional product (12.3) is the unique morphism $I \rightarrow X \otimes W \otimes Y$ which marginalizes to both ψ and ϕ and displays $X \perp Y | W$.

Proof. The fact that $\psi \otimes_W \phi$ displays $X \perp Y | W$ is immediate from the definition. The fact that it marginalizes to ψ and ϕ is clear by the definition of conditionals (Defini-

tion 11.5). Conversely, if $\eta : I \rightarrow X \otimes W \otimes Y$ has these properties, then consider a factorization (12.1). By Remark 12.2, the f and g making up that factorization must be the corresponding conditionals of the marginals of η , i.e. of ψ and ϕ . Hence η is exactly the conditional product. $\square$

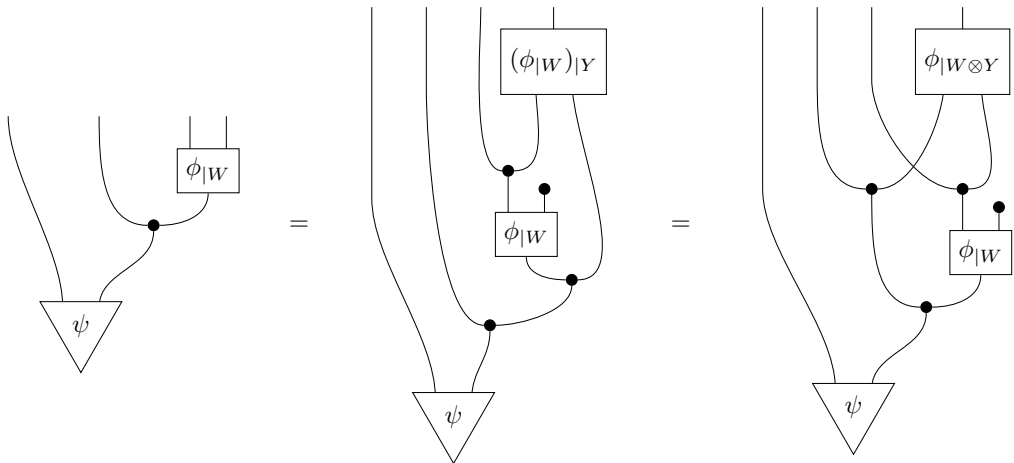
Still assuming that $\mathbf{C}$ has conditionals, we now explain how marginalization and conditional products define a gleaf [32] on $\mathcal{P}_{\text{fin}}(O)$, the lattice of finite subsets of any set of objects $O \subseteq \text{Obj}(\mathbf{C})$. To every $\mathcal{A} \in \mathcal{P}_{\text{fin}}(O)$ we assign the set of morphisms²³ $I \rightarrow \bigotimes_{X \in \mathcal{A}} X$. For any inclusion $\mathcal{A} \subseteq \mathcal{B}$, marginalization defines a restriction map

$$c\left(I, \bigotimes_{X \in \mathcal{B}} X\right) \rightarrow c\left(I, \bigotimes_{X \in \mathcal{A}} X\right),$$

and it is easy to see that this results in a presheaf on $\mathcal{P}_{\text{fin}}(O)$. The above conditional products equip this presheaf with a gluing operation: in [32, Definition 4.1.1], property (a) holds because a conditional product marginalizes to the two given original morphisms. Condition (c) holds because of the commutativity of conditioning with marginalization described in Remark 11.13. Condition (b) amounts to the claim that for $\psi : I \rightarrow X \otimes W$ and $\phi : I \rightarrow W \otimes Y \otimes Z$, the conditional product can be computed in two steps,

$$\psi \otimes_W \phi = (\psi \otimes_W \phi|_W) \otimes_{W \otimes Y} \phi, \quad (12.5)$$

where we have again used Notation 2.8. The proof of this property is more involved: using the alternative representation for the conditional product given on the right of (12.4), we compute



²³ Strictly speaking, we need to make a choice of order for this finite tensor product, but the particular choice is irrelevant. One way to do this is to fix any total order on O and order the objects in $\mathcal{A}$ accordingly.

where the second step consists of reordering wires and applying²⁴ Lemma 11.11. Since this is exactly the desired (12.5), we conclude that we indeed have a gleaf, or equivalently conditional products satisfying the axioms of Dawid and Studený. We can therefore apply the theory of gleaves as developed in [32]. One consequence is that conditional products enjoy an associativity property [32, Proposition 4.3.4].

12.10. Remark. This associativity property of conditional products implies in particular that marginalizing a conditional product in the form (12.3) over W defines the composition in a symmetric monoidal category which we call the *category of couplings*, where:

- ▷ Objects are pairs (X, π_X) where $\pi_X : I \rightarrow X$.
- ▷ The hom-set between (X, π_X) and (Y, π_Y) is the set of all $\psi : I \rightarrow X \otimes Y$ with marginals $\psi(x) = \pi_X$ and $\psi(y) = \pi_Y$.
- ▷ Composition is given by the conditional product (12.3) marginalized over the middle output.
- ▷ The identity morphism of (X, π_X) is given by $\text{copy}_X \circ \pi_X$.

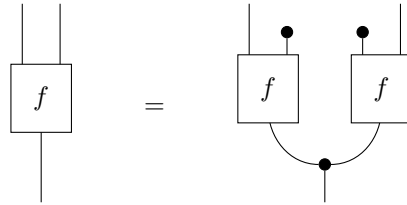
It is also straightforward to verify directly that this data defines a category, and that it has a symmetric monoidal structure defined in the obvious way. However, we believe that the compositional structure of conditional products is more accurately captured by the concept of gleaf as described above. The main reason is that this keeps track of the joint distribution of all variables involved, not requiring one to apply marginalization during composition; it also unifies the sequential product and the monoidal product that we will consider more explicitly in Proposition 13.9 into one kind of gluing.

This ends our discussion of conditional independence via Definition 12.1. However, just as with the notion of conditional in Definition 11.5, we expect the concept of conditional independence to be meaningful also for morphisms whose domain is not the monoidal unit. In order to lead up to this, we briefly consider a notion of conditional independence $X \perp Y \parallel A$ for morphisms $A \rightarrow X \otimes Y$ closely related to the original definition of Cho and Jacobs [15, Definition 6.6]; and also to the earlier conditions CI_{2L} and CI_{2R} of Coecke and Spekkens [21, Section 5.1], which takes place in a different categorical setup.

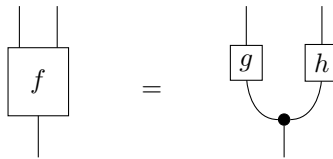
12.11. Lemma. *Given a morphism $f : A \rightarrow X \otimes Y$, the following are equivalent:*

²⁴ This step shows that we really need the existence of conditionals in the sense of Definition 11.5, whereas conditional distributions alone in the sense of Definition 11.1 are not enough for our argument to go through.

(a) $f(x, y|a) = f(x|a)f(y|a)$ in terms of Notation 2.8, meaning that



(b) There are $g : A \rightarrow X$ and $h : A \rightarrow Y$ such that



Proof. While one direction is trivial, the other is by straightforward marginalization. $\square$

12.12. Definition. Let $\mathbf{C}$ be a Markov category. If a morphism $f : A \rightarrow X \otimes Y$ satisfies these equivalent conditions, then we say that f displays the conditional independence $X \perp Y \parallel A$.

Here, we use the double bar “ $\parallel$ ” notation in order to indicate independence conditional on an input, to be distinguished from independence conditional on an output as in Definition 12.1.

We now have at least the simplest two of the semigraphoid properties, in terms of Notation 2.8.

12.13. Lemma. The following hold:

- (a) *Symmetry:* if $f(x, y|a)$ displays $X \perp Y \parallel A$, then $f(y, x|a)$ displays $Y \perp X \parallel A$.
- (b) *Decomposition:* if $f(x_1, x_2, y|a)$ displays $(X_1 \otimes X_2) \perp Y \parallel A$, then the marginal $f(x_1, y|a)$ displays $X_1 \perp Y \parallel A$.

Proof. Straightforward. $\square$

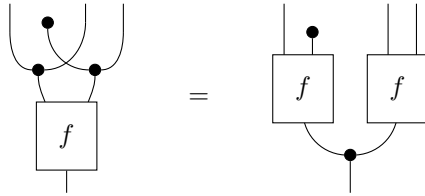
This notion of conditional independence also interacts well with positivity. The following property generalizes the well-known fact that if a probability measure on a product σ -algebra has a deterministic marginal, then it is a product measure.

12.14. Proposition. Let $\mathbf{C}$ be positive (Definition 11.22). Then if a morphism $f : A \rightarrow X \otimes Y$ is such that its marginal $A \rightarrow X$ is deterministic, then f displays $X \perp Y \parallel A$.

Proof. We apply the positivity property of Definition 11.22 with f in place of f and the marginalization morphism



in place of g . Then the resulting (11.8) instantiates to

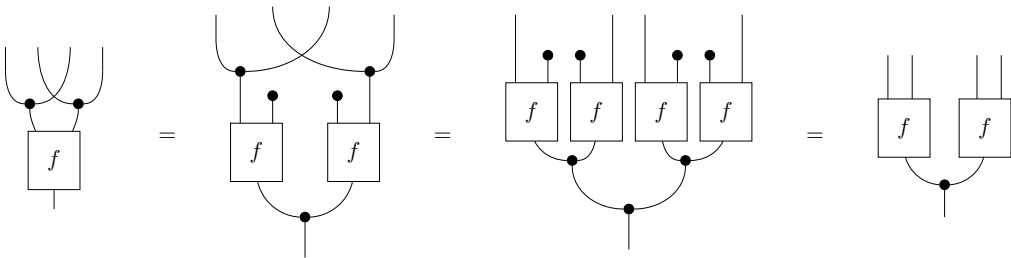


Marginalizing over the second output now gives the desired factorization. $\square$

12.15. Corollary. Let $\mathcal{C}$ be positive. Then for a morphism $f : A \rightarrow X \otimes Y$, the following are equivalent:

- (a) f is deterministic.
- (b) Both marginals of f are deterministic.

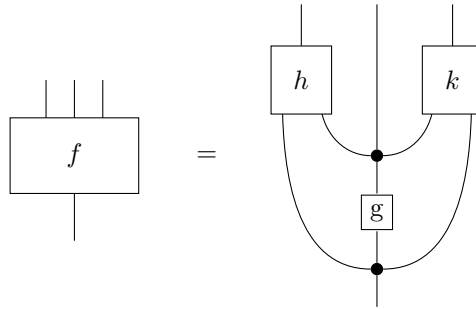
Proof. Assuming (a), the claim (b) is clear since the marginalization map itself is deterministic as well. The implication from (b) to (a) is a straightforward string diagram computation,



where the first and third step are by Proposition 12.14 and the second one is by assumption. $\square$

We now introduce our more involved combined notion of conditional independence, which has not been considered before.

12.16. Definition. A morphism $f : A \rightarrow X \otimes W \otimes Y$ displays the conditional independence $X \perp Y \mid W \parallel A$ if there are $g : A \rightarrow W$, $h : A \otimes W \rightarrow X$ and $k : W \otimes A \rightarrow Y$ such that



It is straightforward to see that this specializes to Definition 12.12 for $W = I$ and to Definition 12.1 for $A = I$.

We now have some more complicated but general rules for reasoning about this kind of conditional independence. Again these are versions of the usual semigraphoid properties, phrased in terms of Notation 2.8.

12.17. Proposition. *The following hold:*

- (a) *Symmetry: if $f(x, w, y|a)$ displays $X \perp Y | W || A$, then $f(y, w, x|a)$ displays $Y \perp X | W || A$.*
- (b) *Decomposition: if $f(x_1, x_2, w, y|a)$ displays $(X_1 \otimes X_2) \perp Y | W || A$, then the marginal $f(x_1, w, y|a)$ displays $X_1 \perp Y | W || A$.*
- (c) *Contraction: if $f(x, w, z, y|a)$ displays $X \perp Y | (W \otimes Z) || A$ and the marginal $f(x, w, z|a)$ displays $X \perp Z | W || A$, then f also displays $X \perp (Z \otimes Y) | W || A$.*

Moreover if $\mathcal{C}$ has conditionals, then also:

- (d) *Weak union: if $f(x_1, x_2, w, y|a)$ displays $(X_1 \otimes X_2) \perp Y | W || A$, then it also displays $X_1 \perp Y | (W \otimes X_2) || A$.*

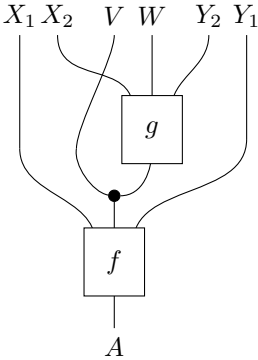
Proof. Again the first two implications are trivial. The proof of (c) and (d) are simple but somewhat tedious generalizations of the corresponding proofs for Lemma 12.5, which we leave to the reader. $\square$

Moreover, conditional independence enjoys certain compositionality properties, formulated as follows in terms of Notation 2.8.

12.18. Proposition. *The following hold:*

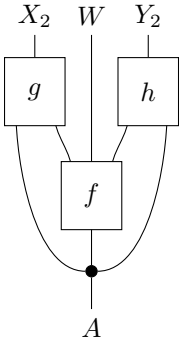
- (a) *If a morphism $f(x_1, w_1, y_1|a_1)$ displays $X_1 \perp Y_1 | W_1 || A_1$ and $g(x_2, w_2, y_2|a_2)$ displays $X_2 \perp Y_2 | W_2 || A_2$, then $(f \otimes g)(x_1, x_2, w_1, w_2, y_1, y_2|a_1, a_2)$ displays $(X_1 \otimes X_2) \perp (Y_1 \otimes Y_2) | (W_1 \otimes W_2) || (A_1 \otimes A_2)$.*

- (b) If a morphism $f(x_1, v, y_1|a)$ displays $X_1 \perp Y_1 | V || A$ and $g(x_2, w, y_2|v)$ displays $X_2 \perp Y_2 | W || V$, then also the composite morphism



displays $(X_1 \otimes X_2) \perp (Y_2 \otimes Y_1) | (V \otimes W) || A$.

- (c) If a morphism $f(x_1, w, y_2|a)$ displays $X_1 \perp Y_1 | W || A$ and $g : A \otimes X_1 \rightarrow X_2$ and $h : Y_1 \otimes A \rightarrow Y_2$ are arbitrary, then also the composite morphism

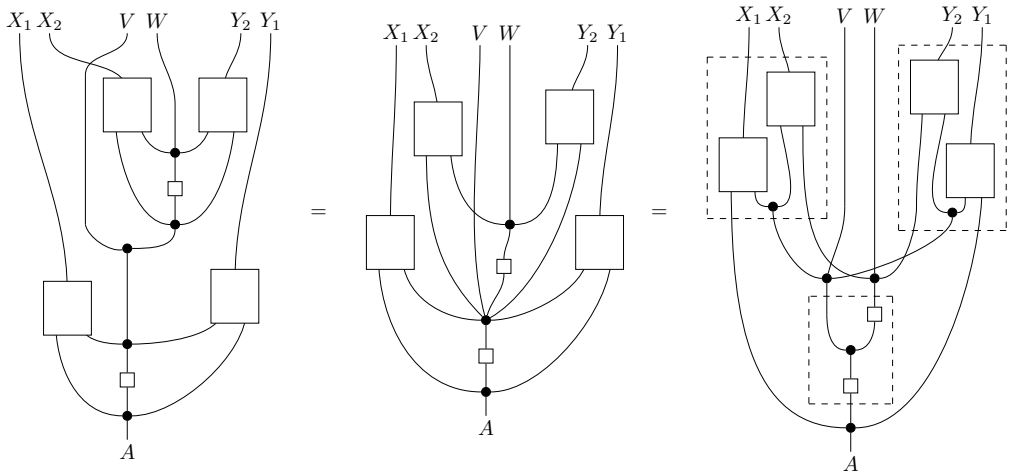


displays $X_2 \perp Y_2 | W || A$.

- (d) If $\mathcal{C}$ has conditionals and a morphism $f(x, v, w, y|a)$ displays $X \perp Y | (V \otimes W) || A$, then there is a conditional $f_{|V}(x, w, y|v, a)$ which displays $X \perp Y | W || (V \otimes A)$.

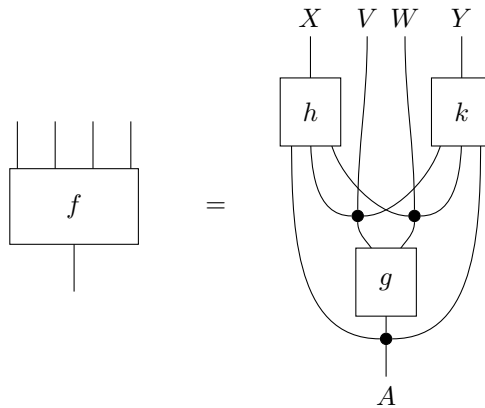
While property (b) is somewhat reminiscent of contraction, property (d) is similar to weak union.

Proof. (a) and (c) are straightforward. For (b), note that the assumptions imply that the composite morphism has the form

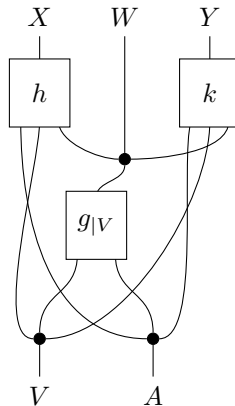


where we leave the relevant morphisms unnamed. The dashed boxes indicate the claimed conditional independence decomposition.

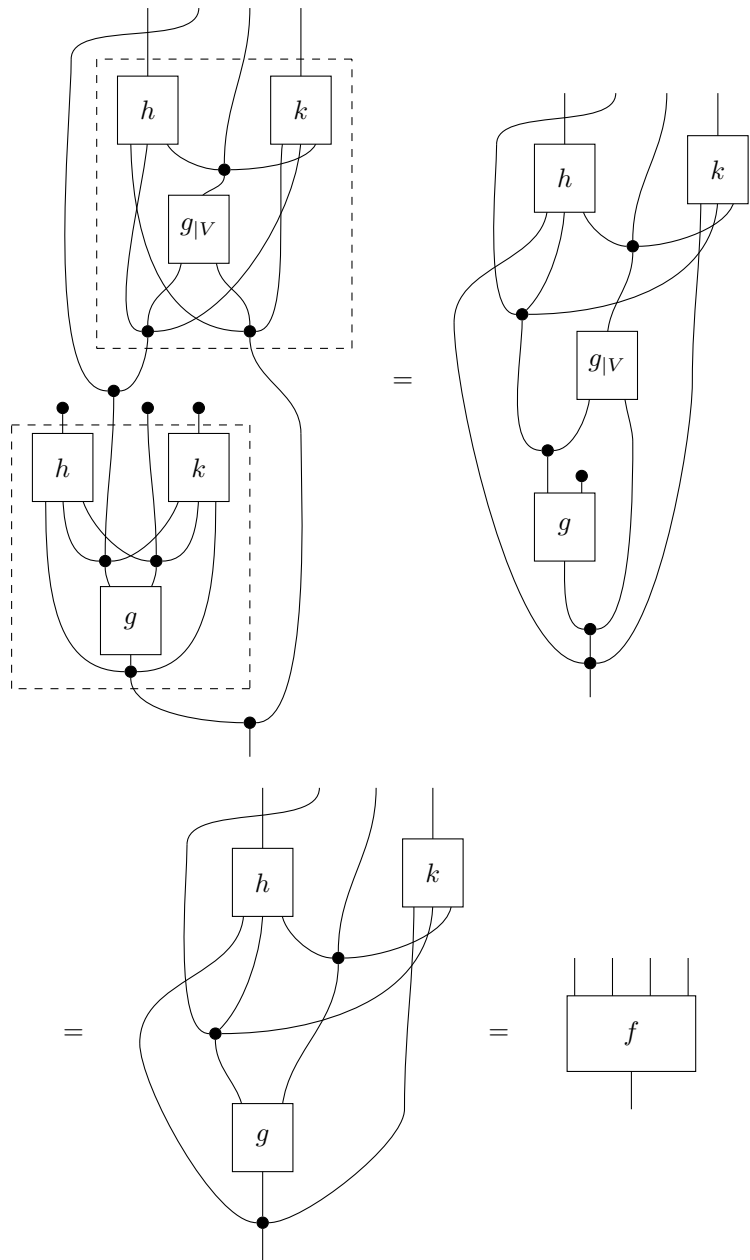
Concerning (d), the assumption implies that f has the form



We construct a new morphism which involves a conditional $g_{|V}$,



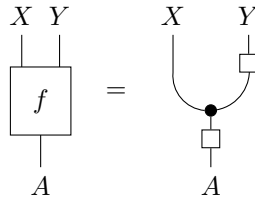
which clearly displays $X \perp Y \mid W \parallel (V \otimes A)$. Furthermore, we verify that this morphism is a conditional $f_{|V}$,



which finishes the proof. $\square$

There is one more notion of conditional independence, which will be of relevance to us in Section 14. We do not yet know how it is related to the one of Definition 12.16.

12.19. Definition. Given a morphism $f : A \rightarrow X \otimes Y$, we say that f displays the conditional independence $A \perp\!\!\!\perp Y \mid X$ if it is of the form



Here, the diagram on the right represents a Markov process, from A to X to Y . This notion of conditional independence appears in the definition of sufficient statistic (Definition 14.3).

For $A = I$, one would expect that this conditional independence always holds. This is indeed the case as soon as $\mathcal{C}$ has conditional distributions, since the conditional independence equation for $I \perp\!\!\!\perp Y \mid X$ amounts to precisely the existence of a conditional for the given morphism $I \rightarrow X \otimes Y$.

Again we have versions of the semigraphoid properties in terms of Notation 2.8, except for symmetry which no longer can even be stated.

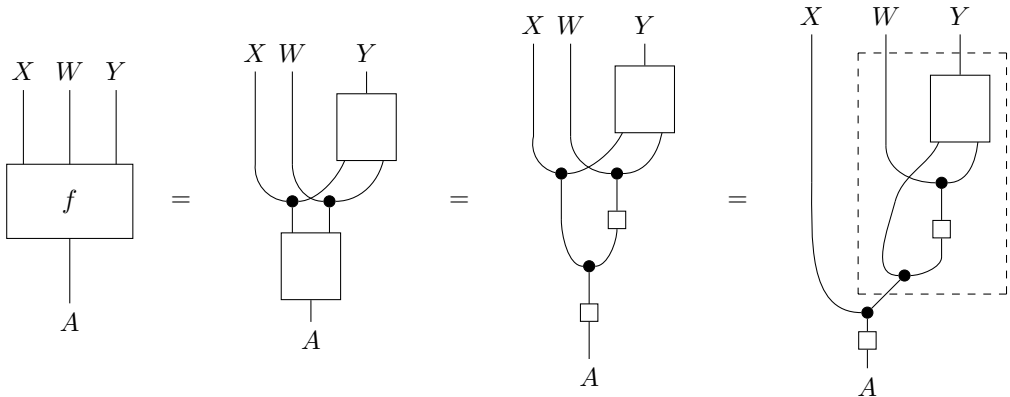
12.20. Proposition. *The following hold:*

- (a) *Right decomposition: if $f(x, y_1, y_2|a)$ displays $A \perp\!\!\!\perp (Y_1 \otimes Y_2) \mid X$, then the marginal $f(x, y_1|a)$ displays $A \perp\!\!\!\perp Y_1 \mid X$.*
- (b) *Contraction: if $f(x, w, y|a)$ displays $A \perp\!\!\!\perp Y \mid (W \otimes X)$ and the marginal $f(x, w|a)$ displays $A \perp\!\!\!\perp W \mid X$, then f also displays $A \perp\!\!\!\perp (W \otimes Y) \mid X$.*

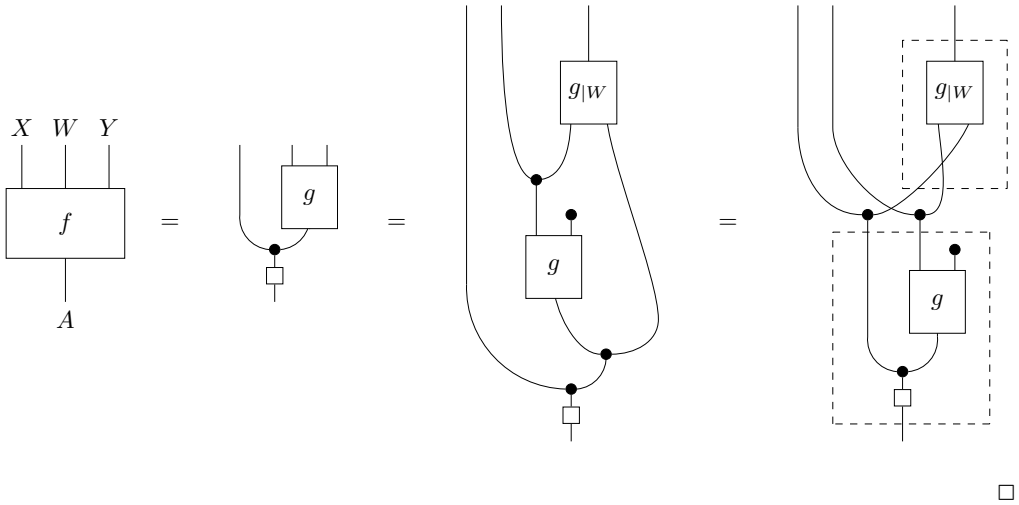
Moreover if $\mathcal{C}$ has conditionals, then also:

- (c) *Weak union: if $f(x, w, y|a)$ displays $A \perp\!\!\!\perp (W \otimes Y) \mid X$, then it also displays $A \perp\!\!\!\perp Y \mid (W \otimes X)$.*

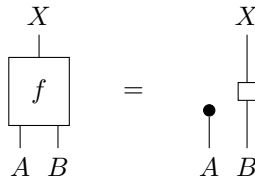
Proof. The proof of (a) is straightforward. For (b), we compute



where the assumptions are used in the first two steps. For (c), we get



There may be one more notion of conditional independence along the lines of [21, Properties CI_{1L} , CI_{1R}] and [15, Proposition 6.9(3)], where a morphism $f : A \otimes B \rightarrow X$ displays this conditional independence of A and X given B if it can be written in a form which simply discards A ,



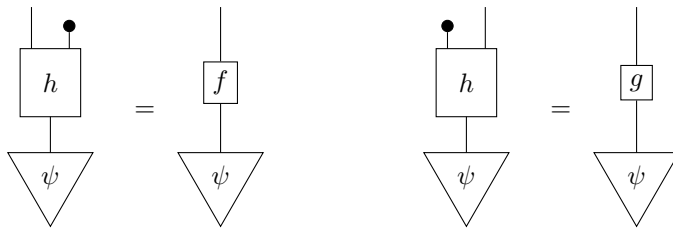
We will not study this notion of conditional independence further in this paper.

12.21. Remark. As has been pointed out to us by Rob Spekkens, all of our definitions of conditional independence have a central feature in common, namely that upon removing all wires from the diagram which refer to objects on which we condition, the diagram disconnects into parts which then become conditionally independent. We hope that this observation may lead to a more unified and more systematic perspective on conditional independence in our setup as a theory of “conditional connectedness”, providing a general understanding of conditional independence as a form of separation similar to separation in undirected graphical models [77], simpler and more intuitive than the complicated *d-separation* criterion for conditional independence in Bayesian networks [95].

Let us end this section with some comments on the relation to other existing axiomatic systems for conditional independence relations. We have already done so throughout this

section for the most well-known such system: the semigraphoid axioms. However, others with a more explicitly categorical flavour have been proposed.

Franz's notion of independent morphisms. Franz [37], and later Gerhold, Lachs and Schürmann [47], have developed a categorical framework for independence which unifies and generalizes several notions of independence that had previously been considered, including ordinary stochastic independence in probability theory and free independence in free probability. However, the basic setup is quite different, and involves working with a category of random variables, meaning measure-preserving maps between probability spaces, instead of a category of Markov kernels. In our setup, involving a Markov category $\mathbf{C}$ whose morphisms we think of as Markov kernels between measurable spaces, we can *define* a probability space as a pair (X, ψ) consisting of an object $X \in \mathbf{C}$ and a morphism $\psi : I \rightarrow X$. A *random variable* or *measure-preserving map* from (X, ψ) to (Y, ϕ) is then a deterministic morphism $f : X \rightarrow Y$ such that $f \circ \psi = \phi$. In other words, we consider the comma category $I/\mathbf{C}_{\text{det}}$ on $\mathbf{C}$, where $\mathbf{C}_{\text{det}} \subseteq \mathbf{C}$ is the subcategory of deterministic morphisms, as our category of probability spaces.²⁵ This category is then again semicartesian, so that Franz's notion of independence applies. Instantiating his definition then gives the following, for (X, ψ) a probability space in our sense: deterministic morphisms $f : X \rightarrow Y$ and $g : X \rightarrow Z$ are independent if and only if there is deterministic $h : X \rightarrow Y \otimes Z$ satisfying



Due to the cartesianess of the monoidal subcategory of deterministic morphisms (Remark 10.13), we can conclude that h is necessarily equal to the tupling $(f, g) : X \rightarrow Y \otimes Z$, the unique deterministic morphism which projects to f and g . Therefore f and g are independent in Franz's sense if and only if $(f, g) \circ \psi$ displays the independence $Y \perp Z$ in the sense of Remark 12.3. Note that $(f, g) \circ \psi : I \rightarrow Y \otimes Z$ can be interpreted as the joint distribution of the random variables f and g .

In summary, Franz's definition keeps track explicitly of the sample space on which random variables are defined (the X in our discussion above), while our notion of conditional independence of Definition 12.1 does not. And there is no need to, as independence is a property only of the *joint distribution* of two random variables, regardless of the underlying sample space. Apart from this difference, Franz's notion of independence, when

²⁵ See Remark 13.14 for a more refined approach.

instantiated in categories of probability spaces which are constructed from Markov categories, is essentially equivalent to the one of Remark 12.3.

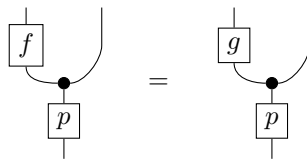
Simpson’s axioms for conditional independence. Simpson [107] has proposed rather general axioms for independence and conditional independence in a categorical framework. When applied to probabilistic independence, one needs to work with categories of probability spaces, just as in Franz’s approach. The key difference is that Simpson does not define what (conditional) independence means in such a category, but rather proposes *axioms* which any reasonable notion of conditional independence should satisfy. We have not yet undertaken a more detailed comparison with Simpson’s approach.

13. Almost surely

A notion that appears in almost every theorem of probability is *almost surely*. In the context of Markov categories, an elegant axiomatization of almost sure equality has been given by Cho and Jacobs [15, Definition 5.1]. Here, we generalize their definition and prove further properties. We also put forward the idea that many other concepts can be relativized with respect to almost surely, and illustrate this by introducing an “almost surely” variant of deterministic morphisms (Definition 10.1). This concept then will be shown to interact well with almost sure equality. As an application, we construct the category of probability spaces internal to any Markov category $\mathcal{C}$.

The following definition specializes to one given by Cho and Jacobs [15, Definition 5.1] in the case $\Theta = I$. Our choice of notation $p : \Theta \rightarrow X$ anticipates statistical models (Definition 14.1), indicating that almost sure equality is particularly relevant in a context where Θ is a parameter space for a statistical model living on a sample space X .

13.1. Definition. Given any morphism $p : \Theta \rightarrow X$, we say that $f, g : X \rightarrow Y$ are *p-a.s. equal* and write $f =_{p\text{-a.s.}} g$ if



13.2. Example. In FinStoch , we have $f =_{p\text{-a.s.}} g$ if and only if $f(y|x) = g(y|x)$ for all $x \in X$ for which there is $a \in A$ with $p(x|a) > 0$.

13.3. Example. In Stoch , we have $f =_{p\text{-a.s.}} g$ if and only if

$$\int_S f(T|x) p(dx|a) = \int_S g(T|x) p(dx|a) \quad (13.1)$$

for all $a \in A$ and $S \in \Sigma_X$ and $T \in \Sigma_Y$. This is equivalent to the functions $f(T|-, g(T|-) : X \rightarrow [0, 1]$ being almost surely equal with respect to the probability measure $p(-|a)$ on X for every parameter value a , generalizing [15, Proposition 5.3].

If f and g are the images in **Stoch** of morphisms in **Meas**, which by abuse of notation we also denote by $f, g : X \rightarrow Y$, then this condition is equivalent to

$$p(S \cap f^{-1}(T)|a) = p(S \cap g^{-1}(T)|a)$$

for all $a \in A$, $S \in \Sigma_X$ and $T \in \Sigma_Y$. Using the additivity of the probability measure $p(-|a)$, this is readily seen to be equivalent to the condition that

$$p(f^{-1}(T) \setminus g^{-1}(T)|a) = 0 = p(g^{-1}(T) \setminus f^{-1}(T)|a)$$

for all $a \in A$ and $T \in \Sigma_Y$. Expressed in terms of symmetric difference Δ , we can also write this as

$$p(f^{-1}(T) \Delta g^{-1}(T)|a) = 0,$$

which is a known condition²⁶ equivalent to almost sure equality in the standard sense for countably generated σ -algebras [38, 343F], but known to be different in general: Fremlin [38, 343J] presents an intuitive example of a measurable and measure-preserving map f from a measure space (X, Σ_X, p) to itself such that although $f(x) \neq x$ for all $x \in X$ and Σ_X separates the points of X , the map f is p -almost surely equal to id_X in our sense.

Generally speaking, our working hypothesis is that (13.1) is the “correct” notion of almost sure equality in **Stoch**. This hypothesis will be supported by our results of the following sections, where we find that using (13.1) as almost sure equality recovers standard concepts, such as bounded completeness (Example 15.4).

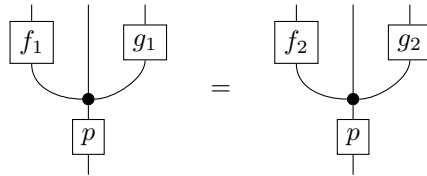
Returning to the general theory, here is a basic property of a.s. equality which follows directly from the definition.

13.4. Lemma. *Let $p : \Theta \rightarrow X$. If $f =_{p\text{-a.s.}} g$ for $f, g : X \rightarrow Y$ and $h : Y \rightarrow Z$, then also $hf =_{p\text{-a.s.}} hg$.*

Furthermore, an occasionally useful fact is that if $g =_{p\text{-a.s.}} f$, then the defining equation still holds with more than one copy operation taking place before the application of f or g . This is a special case of the following:

13.5. Lemma. *Let $p : \Theta \rightarrow X$. If $f_1, f_2 : X \rightarrow Y$ and $g_1, g_2 : X \rightarrow Z$ are such that $f_1 =_{p\text{-a.s.}} f_2$ and $g_1 =_{p\text{-a.s.}} g_2$, then also*

²⁶ It had been introduced to us by Robert Furber and seems to be due to Fremlin.



Proof. Straightforward application of the assumptions and the associativity and commutativity of the comultiplication. $\square$

A similar (and easy to prove) observation is that if $f_1, f_2 : X \rightarrow Y$ and $g_1, g_2 : W \rightarrow Z$ are such that $f_1 =_{p\text{-a.s.}} f_2$ for $p : \Theta \rightarrow X$ and $g_1 =_{q\text{-a.s.}} g_2$ for $q : \Phi \rightarrow W$, then also $(f_1 \otimes g_1) =_{p \otimes q} (f_2 \otimes g_2)$.

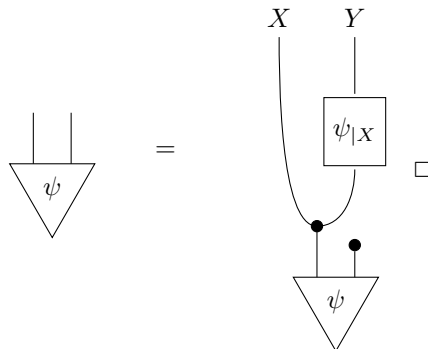
13.6. Remark. Almost sure equality can obviously always be weakened by precomposition: if $f =_{p\text{-a.s.}} g$, and if the codomain of q matches the domain of p , then also $f =_{pq\text{-a.s.}} g$. The causality axiom (Definition 11.31) provides conversely a way to decompose a relativizing morphism: marginalizing its conclusion over the middle output shows that if $f =_{pq\text{-a.s.}} g$, whenever this makes sense, then also $fp =_{q\text{-a.s.}} gp$. But regardless of whether the causality axiom holds or not, $f =_{p\text{-a.s.}} g$ clearly implies $fp = gp$.

We now consider conditionals in the sense of Definition 11.5, and note that they are almost surely unique [15, p. 22].

13.7. Proposition. *Given a morphism $\psi : I \rightarrow X \otimes Y$, the conditional $\psi|_X : X \rightarrow Y$ is unique $\psi(x)$ -a.s. (if it exists).*

Here, $\psi(x)$ is the marginal morphism $I \rightarrow X$ of the original $\psi(x, y)$ in terms of Notation 2.8. The uniqueness statement means that any two such conditionals are $\psi(x)$ -a.s. equal.

Proof. This is an immediate consequence of the defining equation



A related point is that almost sure equality lets us define categories of probability spaces in $\mathbf{C}$. A *probability space* in a Markov category $\mathbf{C}$ is a pair (X, ψ) where $X \in \mathbf{C}$ is an object and $\psi : I \rightarrow X$ is a morphism playing the role of a distribution on X . If (X, ψ) and (Y, ϕ) are probability spaces, then a morphism $f : X \rightarrow Y$ is *measure-preserving* if $\phi = f\psi$. In the following definition, we impose the causality condition as an apparently necessary hypothesis for our well-definedness argument below.

13.8. Definition. Suppose that $\mathbf{C}$ is causal (Definition 11.31). Then the *category of probability spaces and Markov kernels* $\text{ProbStoch}(\mathbf{C})$ has as objects the probability spaces (X, ψ) in $\mathbf{C}$, and as morphisms $(X, \psi) \rightarrow (Y, \phi)$ those maps $f : X \rightarrow Y$ which satisfy $f\psi = \phi$, modulo ψ -a.s. equality, with composition inherited from $\mathbf{C}$.

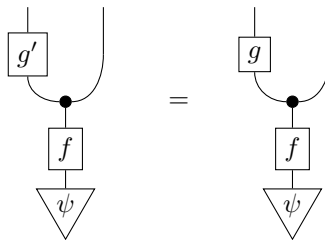
In categories like FinStoch and Stoch , the condition $f\psi = \phi$ says exactly that f should be measure-preserving in the usual sense.

13.9. Proposition. Suppose that $\mathbf{C}$ is causal.

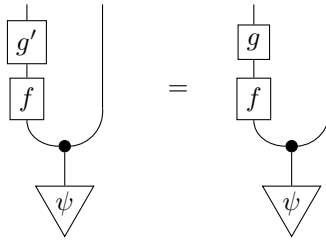
- (a) Composition in $\text{ProbStoch}(\mathbf{C})$ is well-defined, and $\text{ProbStoch}(\mathbf{C})$ inherits the symmetric monoidal structure from $\mathbf{C}$.
- (b) If $\mathbf{C}$ has conditionals, then $\text{ProbStoch}(\mathbf{C})$ is isomorphic to the category of couplings described in Remark 12.10.

For (b), recall that the existence of conditionals implies causality (Proposition 11.34), so that only the existence of conditionals needs to be verified when applying this to a given Markov category $\mathbf{C}$. We do not know whether the well-definedness of composition would still hold without causality.

Proof. (a) Concerning the well-definedness, let $f, f' : X \rightarrow Y$ and $g : Y \rightarrow Z$ be representative morphisms between probability spaces (X, ψ) , (Y, ϕ) and (Z, π) , and suppose that $f =_{\psi\text{-a.s.}} f'$. Then we get $gf =_{\psi\text{-a.s.}} gf'$ by Lemma 13.4. Well-definedness in g follows from the causality property of Definition 11.31: the assumption $g =_{\phi\text{-a.s.}} g'$ means that



and therefore applying causality and marginalizing over the middle output results in the desired equation



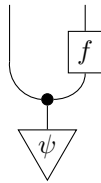
Hence composition is well-defined.

The monoidal structure on $\mathbf{ProbStoch}(\mathbf{C})$ is given by

$$(X, \psi) \otimes (Y, \phi) := (X \otimes Y, I \xrightarrow{\cong} I \otimes I \xrightarrow{\psi \otimes \phi} X \otimes Y),$$

and similarly on morphisms. With the structure isomorphisms also inherited from $\mathbf{C}$ in the obvious way, it is straightforward to check that we indeed get a symmetric monoidal category.

- (b) We first establish the bijection between the hom-sets: to a given morphism $f : (X, \psi) \rightarrow (Y, \phi)$ represented by $f \in \mathbf{C}(X, Y)$, we assign the joint $I \rightarrow X \otimes Y$ given by [54, Section 7.1]



Clearly f itself is a conditional of this distribution with respect to the first output, and this assignment establishes a bijection between ψ -a.s. equivalence classes of morphisms $f : X \rightarrow Y$ satisfying $\phi = f\psi$ and couplings, i.e. morphisms $I \rightarrow X \otimes Y$ with marginals ψ and ϕ . Moreover, together with the formula (12.4) for the conditional product, it is easy to see that this bijection respects composition. $\square$

For example, (a) allows us to construct $\mathbf{ProbStoch}(\mathbf{Stoch})$, the category of all probability spaces and measure-preserving maps modulo almost sure equality,²⁷ since we already know $\mathbf{Stoch}$ to be causal (Example 11.35).

13.10. Remark. If $\mathbf{C}$ has conditionals, we now sketch how Bayesian inversion equips the category $\mathbf{ProbStoch}(\mathbf{C})$ with a symmetric monoidal dagger functor in the sense of dagger categories.²⁸ Since this applies in particular to $\mathbf{BorelStoch}$ by Example 11.7, this

²⁷ Assuming that almost sure equality is defined appropriately, namely as in Example 13.3.

²⁸ This means that $-^\dagger : \mathbf{C}^{\text{op}} \rightarrow \mathbf{C}$ is identity-on-objects, satisfies $(gf)^\dagger = f^\dagger g^\dagger$ and $f^{\dagger\dagger} = f$ and $\text{id}^\dagger = \text{id}$ and $(g \otimes f)^\dagger = g^\dagger \otimes f^\dagger$, and is such that the dagger of any symmetric monoidal structure isomorphism is its inverse. The latter holds in our case because the dagger of any deterministic isomorphism is its inverse,

generalizes a recent result due to Clerc, Danos, Dahlqvist and Garnier, who proved this in explicit measure-theoretic terms for **BorelStoch** [23, Theorem 2.10]. Concretely, if $f : (X, \psi) \rightarrow (Y, \phi)$ is a morphism represented by $f : X \rightarrow Y$ with $f\psi = \phi$, then we get a morphism $f^\dagger : (Y, \phi) \rightarrow (X, \psi)$ represented by any $f^\dagger : Y \rightarrow X$ satisfying the equation

$$\begin{array}{c} \text{---} \\ | \\ \text{---} \end{array} \begin{array}{c} \boxed{f} \\ | \\ \bullet \\ \triangle \psi \end{array} = \begin{array}{c} \text{---} \\ | \\ \text{---} \end{array} \begin{array}{c} \boxed{f^\dagger} \\ | \\ \bullet \\ \triangle \phi \end{array} \quad (13.2)$$

This equation expresses the property that $f^\dagger$ should be a conditional on Y for the joint $I \rightarrow X \otimes Y$ depicted in (13.2). In terms of the characterization as the category of couplings stated in Proposition 13.9(b), this dagger functor is a mere bookkeeping device, swapping couplings $I \rightarrow X \otimes Y$ to couplings $I \rightarrow Y \otimes X$ by composition with $\text{swap}_{X,Y} : X \otimes Y \rightarrow Y \otimes X$. This characterization can be used to give a simple proof of the fact that the defining properties of a symmetric monoidal dagger functor are satisfied.

In conclusion, for any Markov category $\mathbf{C}$ which has conditionals, Bayesian inversion is implemented by a symmetric monoidal dagger functor on $\text{ProbStoch}(\mathbf{C})$.

Returning to the theory of almost surely, we believe that other concepts of probability and statistics can and should be relativized similarly with respect to almost surely in a similar way as Definition 13.1 does for equality, by starting with the original definition and composing it with p followed by $\text{copy}_X : X \rightarrow X \otimes X$. For concreteness and future use, we show what this amounts to in the case of Definition 10.1.

13.11. Definition. Given $p : \Theta \rightarrow X$, we say that $f : X \rightarrow Y$ is *p-a.s. deterministic* if

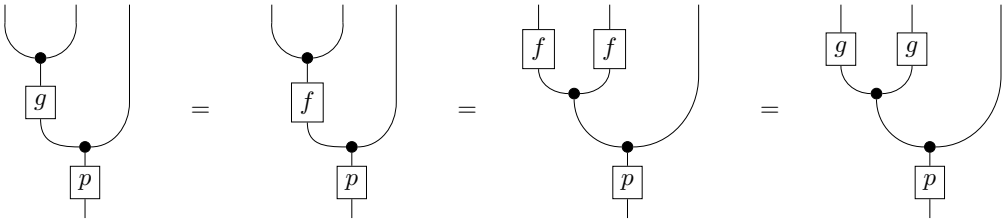
$$\begin{array}{c} \text{---} \\ | \\ \text{---} \end{array} \begin{array}{c} \bullet \\ | \\ \boxed{f} \\ | \\ \bullet \\ \boxed{p} \end{array} = \begin{array}{c} \text{---} \\ | \\ \text{---} \end{array} \begin{array}{c} \boxed{f} \quad \boxed{f} \\ | \quad | \\ \bullet \\ | \\ \bullet \\ \boxed{p} \end{array}$$

In other words, we now postulate the determinism equation only with respect to p -a.s. equality. We typically expect such relativizations to respect a.s. equality, which is indeed what happens in this case:

which applies to the structure isomorphisms by Lemma 10.12. All of these properties are straightforward to check.

13.12. Lemma. Let $p : \Theta \rightarrow X$. If $f, g : X \rightarrow Y$ are such that $g =_{p\text{-a.s.}} f$ and f is p -a.s. deterministic, then so is g .

Proof. We use the assumptions to compute



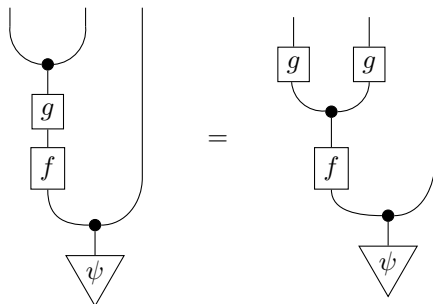
where the last step is by Lemma 13.5. $\square$

13.13. Example. Generalizing Example 10.4, in **Stoch** a morphism f is p -a.s. deterministic if for all $\theta \in \Theta$ we have

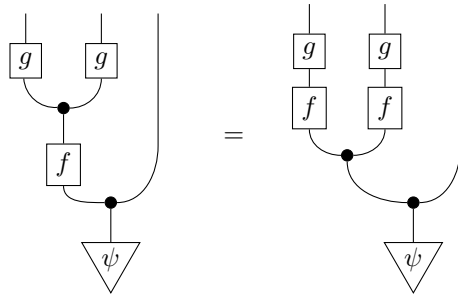
$$\int_{x \in R} f(S|x) f(T|x) p(dx|\theta) = \int_{x \in R} f(S \cap T) p(dx|\theta)$$

for all $R \in \Sigma_X$ and $S, T \in \Sigma_Y$. The universal quantification over R implies that the integrands $x \mapsto f(S|x)f(T|x)$ and $x \mapsto f(S \cap T|x)$ are $p(-|\theta)$ -almost surely equal for all θ . By reasoning analogous to Example 10.4, we thereby obtain that f is p -a.s. deterministic if and only if for every $S \in \Sigma_X$ we have $f(S|x) \in \{0, 1\}$ at least $p(-|\theta)$ -almost surely for every θ .

13.14. Remark. Assuming that **C** is causal, we can restrict the category $\mathbf{ProbStoch}(\mathbf{C})$ considered in Proposition 13.9 to the almost surely deterministic morphisms, resulting in a symmetric monoidal subcategory $\mathbf{Prob}(\mathbf{C}) \subseteq \mathbf{ProbStoch}(\mathbf{C})$ which we call the *category of probability spaces in C*. More precisely, a morphism $(X, \psi) \rightarrow (Y, \phi)$ represented by $f \in \mathbf{C}(X, Y)$ is in $\mathbf{ProbStoch}(\mathbf{C})$ if and only if f is ψ -a.s. deterministic. Lemma 13.12 shows that this property does not depend on the choice of representative. The fact that $\mathbf{Prob}(\mathbf{C})$ is closed under composition again follows from the causality property: the assumption that g is $f\psi$ -a.s. deterministic gives



thanks to causality, and the assumption that f is ψ -a.s. deterministic allows us to write



so that both steps together prove the claim.

Finally, it is easy to see that $\text{Prob}(\mathbf{C})$ is indeed closed under monoidal product and contains the structure isomorphisms by Lemma 10.12, so that it indeed forms a symmetric monoidal subcategory of $\text{ProbStoch}(\mathbf{C})$ if $\mathbf{C}$ is causal.

We now consider some relations between almost sure determinism and the positivity axiom discussed in Definition 11.22 and after. This following observation is in the spirit of Lemma 10.9.

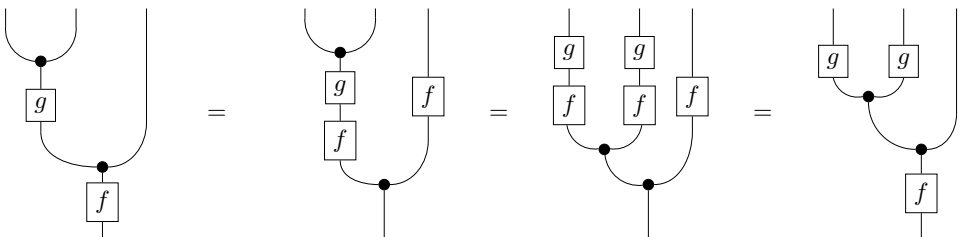
13.15. Proposition. *Suppose that $f : X \rightarrow Y$ and $g : Y \rightarrow Z$ in $\mathbf{C}$ are such that gf is deterministic. Suppose also that*

- (a) f is deterministic, or
- (b) $\mathbf{C}$ is positive (Definition 11.22).

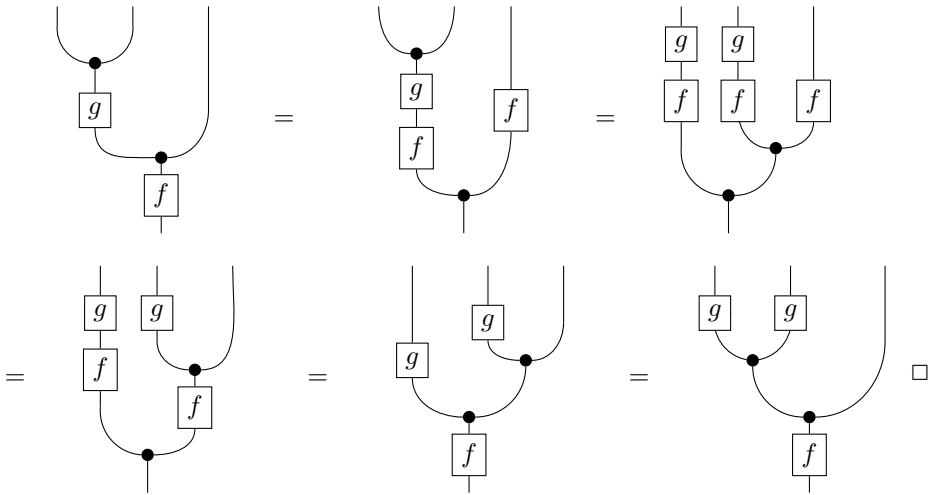
Then g is f -a.s. deterministic.

In particular, if $gf = \text{id}$ and $\mathbf{C}$ is positive, then g is f -a.s. deterministic. Example 11.27 shows that $gf = \text{id}$ alone is not enough to imply that g is f -a.s. deterministic.

Proof. (a) This is a straightforward string diagram computation,



- (b) We use the positivity equation of Lemma 11.24 repeatedly together with associativity,



In the proof of Theorem 14.5, we will need a stronger positivity condition than Definition 11.22 obtained by relativizing almost surely.

13.16. Definition. We say that $\mathcal{C}$ is *strictly positive* if whenever morphisms

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

are such that gf is p -a.s. deterministic, then also

(13.3)

13.17. Remark. Carrying an extra $p : \Theta \rightarrow X$ around in the proof of Proposition 13.15 shows the following more general statement. Suppose that gf is p -a.s. deterministic, and that also

- (a) f is p -a.s. deterministic, or
- (b) $\mathcal{C}$ is strictly positive.

Then g is fp -a.s. deterministic.

As a generalization of Lemma 11.24, we have:

13.18. Lemma. If $\mathcal{C}$ has conditionals, then $\mathcal{C}$ is strictly positive.

Proof. Use the same s as in (11.9), and then do the analogous computation as in the proof of Lemma 11.24 while carrying the additional p along. $\square$

In particular, **FinStoch** and **Gauss** are strictly positive. For **Stoch**, we need a separate argument generalizing Example 11.25.

13.19. Example. **Stoch** is strictly positive, as we show now. By the characterization of p -a.s. deterministic morphisms of Example 13.13, the assumption is that

$$\int_{y \in Y} g(S|y) f(dy|x) \in \{0, 1\} \quad (13.4)$$

for all $S \in \Sigma_X$ and $p(-|\theta)$ -almost all $x \in X$ and all $\theta \in \Theta$. What we want to prove is that the equation

$$\int_{y \in T} g(S|y) f(dy|x) = f(T|y) \int_{y \in Y} g(S|y) f(dy|x)$$

holds for all $S \in \Sigma_X$ and $T \in \Sigma_Y$ for $p(-|\theta)$ -almost all $x \in X$ and all θ . This is by the same argument as in Example 11.25.

As a final note, let us propose one more definition which can be nicely formulated in terms of almost sure equality, although we will not study it any further in this paper.

13.20. Definition. Let $\mathbf{C}$ be a Markov category. For any morphism $p : A \rightarrow X$, the *support* $\text{supp}(p)$ is any object which represents the functor

$$\mathbf{C}(X, -) / \sim_{p\text{-a.s.}} : \mathbf{C} \longrightarrow \mathbf{Set}$$

mapping every object Y to the set of equivalence classes of morphisms $X \rightarrow Y$ modulo p -a.s. equality.

More concretely, this means that $\text{supp}(p)$ comes equipped with a *universal morphism* $i : \text{supp}(p) \rightarrow X$ having the property that for every $Y \in \mathbf{C}$, composition with i induces a bijection between morphisms $\text{supp}(p) \rightarrow Y$ and p -a.s. equivalence classes of morphisms $X \rightarrow Y$.

13.21. Remark. While this is somewhat speculative at the present stage, our hope is that this notion of support will specialize to the usual notion of support of a Radon probability measure on a locally compact space, or more generally that of a τ -smooth probability measure on any topological space, when applied to morphisms $I \rightarrow X$ in a suitable Markov category with topological spaces as objects. So far we have not considered such a category yet.

As is usually the case for an object specified in terms of a universal property, the support $\text{supp}(p)$ may or may not exist for a given morphism p in a given category $\mathcal{C}$.

13.22. Question. Suppose that $\text{supp}(p)$ exists. Then does $p : A \rightarrow X$ necessarily factor through $i : \text{supp}(p) \rightarrow X$? Is i necessarily deterministic?

As easily follows from the following example, the answer to both questions is positive in the case of **FinStoch**.

13.23. Example. In **FinStoch**, the support of any morphism p exists, and is exactly the support in the usual sense,

$$\text{supp}(p) = \{x \in X \mid \exists a : p(x|a) > 0\},$$

where the universal morphism $i : \text{supp}(p) \rightarrow X$ is the inclusion map. The universality is easy to prove as a consequence of Example 13.2.

Finally, we remark that the existence of supports can be useful for simplifying arguments which involve relativization with respect to almost surely.

13.24. Proposition. *If every morphism in $\mathcal{C}$ has a support and $\mathcal{C}$ is positive (Definition 11.22), then $\mathcal{C}$ is also strictly positive (Definition 13.16).*

Proof. Straightforward. $\square$

14. Sufficient statistics and the Fisher–Neyman factorization theorem

We finally move on to the treatment of statistics within our formalism, comprising this section and the following two. Informally, a *statistic* is a function of a random sample intended to capture the essential information contained in this sample. Perhaps the most common example is the *sample mean*: tossing a coin $n \gg 1$ times results in a probability distribution over 2^n outcomes, or in our setting a morphism $I \rightarrow \{\text{heads}, \text{tails}\}^n$. The sample mean is the map $\{\text{heads}, \text{tails}\}^n \rightarrow [0, 1]$ which maps a sequence of outcomes to the relative frequency of (say) heads, which is the essential information contained in the sequence, assuming that the individual throws are independent. This independence assumption is a *statistical model*: a hypothesis about a distribution which comes in the form of a collection of distributions on the sample space X , in this case $\{\text{heads}, \text{tails}\}^n$, parametrized by a parameter space Θ , in this case the space of probability distributions describing a single toss, which we can identify with $[0, 1]$. In other words, a statistical model is a *Markov kernel* $\Theta \rightarrow X$ [90].

14.1. Definition. Given a sample space $X \in \mathcal{C}$, a *statistical model* with values in X consists of another object $\Theta \in \mathcal{C}$ and a morphism $p : \Theta \rightarrow X$.

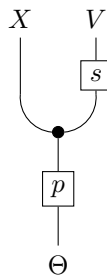
Put yet differently, a statistical model for X is an object in the slice category over X . This notion of statistical model is the simplest and most naive version, where one considers only a single fixed sample space X ; for a more refined notion and detailed discussion, see [90] as well as the *repetitive structures* of [76].²⁹

While our definition of statistical model keeps track of an explicit parametrization of all distributions via a parameter space Θ , one can alternatively omit this parametrization and consider categories of unparametrized statistical models, which are mere sets of probability distributions [69, Definition 2.20]. We view this difference as relatively minor, since a set of probability distributions can always be taken to be parametrized by itself.

14.2. Definition. A *statistic* for a statistical model $p : \Theta \rightarrow X$ is a deterministic morphism $s : X \rightarrow V$ to some other object V .

This is essentially the traditional definition (modulo subtleties like those of Example 10.4, depending on which category $\mathbf{C}$ we are working in).

By combining a statistic $t : X \rightarrow V$ with a statistical model $p : \Theta \rightarrow X$, we obtain a morphism $\Theta \rightarrow V \otimes X$ which describes the joint distribution over sample space and statistic as a function of the model parameter,



By construction, this morphism displays the conditional independence $\Theta \perp\!\!\!\perp V \mid X$ (Definition 12.19).

An important property which a statistic may or may not have is its *sufficiency*. Intuitively, a sufficient statistic for a statistical model is as informative than the full sample.

14.3. Definition. A statistic $s : X \rightarrow V$ for a statistical model $p : \Theta \rightarrow X$ is *sufficient* if there is $\alpha : V \rightarrow X$ such that

²⁹ It seems plausible to us that McCullagh's and Lauritzen's refined definitions fit perfectly well into our formalism as well by working within a suitable category of diagrams (Section 7), but we have not yet worked out the details of this.

$$(14.1)$$

Formulating sufficiency in terms of the existence of such a morphism is certainly not new and has been considered previously in more concrete ways, e.g. in Dawid’s work on statistical operations [25, Example III(a)]. Also Golubtsov [53, Section 7] already seems to have envisioned the possibility of giving such an abstract categorical definition of sufficiency.

We also say that α is a *sufficiency witness* for s if the above condition holds; another way to phrase it is to say that the morphism on the left in (14.1) displays the conditional independence $\Theta \perp\!\!\!\perp X \mid V$ in addition to $\Theta \perp\!\!\!\perp V \mid X$, which is a well-known way to formulate sufficiency. Also note the similarity of this condition with Bayesian inversion (13.2). Remarkably, the condition automatically implies that α *almost surely splits* s in the sense that $s\alpha =_{sp\text{-a.s.}} \text{id}_V$,

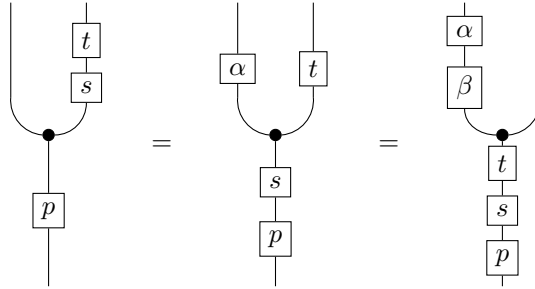
$$(14.2)$$

where the first equation is by sufficiency and the second one because s is deterministic. So intuitively, sufficiency means that given a value of the statistic s , applying α plays the role of making an *inference* over the corresponding outcomes in X which give rise to that value of s , and in such a way that this distribution coincides with the conditional distribution over outcomes given any parameter value. While the morphism α does play a role similar to that of a conditional, our definition may also be of interest in categories which do not have conditionals, such as **Stoch** (Example 11.3).

Sufficient statistics enjoy the following compositionality property, inspired by a similar compositionality result of Jost, L  , Luu and Tran in their more specific setting [69, Theorem 2.28(1)].

14.4. Lemma. *Let $s : X \rightarrow V$ and $t : V \rightarrow W$ be statistics. If s is sufficient for p and t is sufficient for sp , then also the composite ts is sufficient for p .*

Proof. Suppose that $\alpha : V \rightarrow X$ is a sufficiency witness for s with respect to p and $\beta : W \rightarrow V$ is a sufficiency witness for t with respect to the pushforward model sp . Then we show that $\alpha\beta$ is a sufficiency witness for ts with respect to p ,



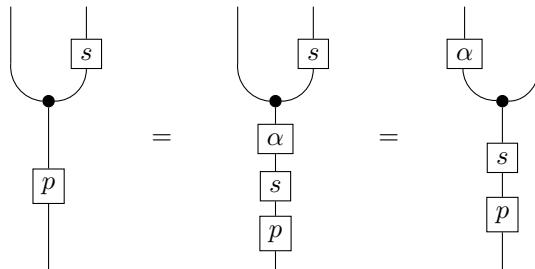
using the respective assumed instances of (14.1). $\square$

Here is an abstract characterization of sufficient statistics, which seems to be a close relative of or version of the Fisher–Neyman factorization theorem.

14.5. Theorem. Suppose that $\mathbb{C}$ is strictly positive (Definition 13.16). Then a statistic $s : X \rightarrow V$ is sufficient for a statistical model $p : \Theta \rightarrow X$ if and only if there is $\alpha : V \rightarrow X$ such that $\alpha sp = p$ and $s\alpha =_{sp\text{-a.s.}} \text{id}_V$.

More precisely, $\alpha : V \rightarrow X$ is a sufficiency witness if and only if it satisfies these alternative conditions. Since we have already assembled all the relevant ingredients, the proof is now essentially trivial.

Proof. The “only if” direction is trivial by marginalization of (14.1) and by (14.2). For the “if” direction, we show that $\alpha sp = p$ and $s\alpha =_{sp\text{-a.s.}} \text{id}$ together imply (14.1). Indeed,



where both steps are by assumption, and the second one also uses the strict positivity property of (13.3). $\square$

We now illustrate the relation to the Fisher–Neyman factorization theorem in the case of our basic running example.

14.6. Example. Consider FinStoch, where statistics are just functions between finite sets. Now consider the equation $\alpha sp = p$ from Theorem 14.5, which becomes

$$p(x|\theta) = \sum_{v, x'} \alpha(x|v) s(v|x') p(x'|\theta).$$

The condition $s\alpha =_{sp\text{-a.s.}} \text{id}_V$ implies that the sum over v has only one contributing term, namely $v = s(x)$, where s now also denotes the underlying function of finite sets by abuse of notation. Therefore

$$p(x|\theta) = \alpha(x|s(x)) \sum_{x' : s(x')=s(x)} p(x'|\theta).$$

This is the Fisher–Neyman factorization into one term which only depends on x , and one term which depends on θ and where the dependence on x is only through $s(x)$. Conversely, given a Fisher–Neyman factorization

$$p(x|\theta) = g_\theta(s(x)) h(x),$$

we construct α satisfying the required properties. We can assume without loss of generality that for every $v \in V$ there is $x \in s^{-1}(v)$ and $\theta \in \Theta$ with $p(x|\theta) > 0$, since on all other v we can define $\alpha(-|v)$ arbitrarily. Then we get that the denominator in the first case of the expression

$$\alpha(x|v) := \begin{cases} \left(\sum_{x' \in s^{-1}(v)} h(x') \right)^{-1} h(x) & \text{if } s(x) = v, \\ 0 & \text{otherwise} \end{cases}$$

is nonzero, making it well-defined. Now $s\alpha =_{sp\text{-a.s.}} \text{id}$ holds by construction, and plugging in the above expressions gives

$$(\alpha sp)(x|\theta) = \sum_{x' \in X} \alpha(x|s(x')) p(x'|\theta) = \sum_{x' \in s^{-1}(x)} \alpha(x|s(x)) p(x'|\theta) = p(x|\theta),$$

as was to be shown.

We leave the investigation of further instances of Theorem 14.5 to future work, and in particular the problem of determining whether our result also specializes to the standard Fisher–Neyman factorization theorem in its general measure-theoretic form [60, Corollary 1].

15. Complete morphisms, ancillary statistics, and Basu’s theorem

Basu’s theorem [5, Theorem 2] is a classical result on the independence of a *complete* sufficient statistic from any ancillary statistic. Let us start by discussing the more mysterious notion of completeness, introduced by Lehmann and Scheffé in 1950 [81, Section 3].

We first explain how it secretly is a concept that applies to general morphisms in every Markov category.

Recall from Remark 13.6 that if $g =_{f\text{-a.s.}} h$ for $f : X \rightarrow Y$ and $g, h : Y \rightarrow Z$, then also $gf = hf$. Now f is defined to be complete if the converse holds as well.

15.1. Definition. A morphism $f : X \rightarrow Y$ is *complete* if $gf = hf$ implies $g =_{f\text{-a.s.}} h$.

This definition is in the spirit of Dawid's [25, Definition 4.1]. Note that at our level of abstraction, we have no way to distinguish between *complete* and *bounded complete*. We will prove in Example 15.4 that in **Stoch**, our definition actually reproduces the latter.

15.2. Remark. The main difference between our definition and the standard one is terminological: while standardly, completeness is a property of a *statistic* s relative to a *statistical model* p , we consider it rather as a property of the composite morphism sp , since it only depends on the latter. In other words, completeness is best viewed a property of Markov kernels, or more generally morphisms in a Markov category.

15.3. Example. In **FinStoch**, a morphism $f : X \rightarrow Y$ is complete if and only if the affine hull of $\text{im}(f)$ is a face of the probability simplex over Y . For example, an $f : I \rightarrow Y$ is complete if and only if it is a Dirac measure, or equivalently if and only if it is deterministic.

To see that this characterization holds, it is most instructive to present a geometrical argument. We think of morphisms $g, h : Y \rightarrow Z$ as affine maps from the probability simplex over Y to the probability simplex over Z . Now recall from Example 13.23 that $g =_{f\text{-a.s.}} h$ if and only if the restrictions of g and h to $\text{supp}(f)$ coincide, where $\text{supp}(f) \subseteq Y$ is the set of all $y \in Y$ for which there is $x \in X$ with $f(y|x) > 0$. We can identify the support $\text{supp}(f)$ with the probability simplex over $\text{supp}(f)$, as included in the probability simplex over Y as a face. Therefore $g =_{f\text{-a.s.}} h$ if and only if g and h coincide on the smallest face containing $\text{im}(f)$.

Similarly, $gf = hf$ holds if and only if g and h coincide on $\text{im}(f)$, or equivalently on the affine hull of $\text{im}(f)$. We can now show that $gf = hf$ implies $g =_{f\text{-a.s.}} h$ if and only if the affine hull of $\text{im}(f)$ is a face. By the above, the “if” direction is clear; for the “only if” part, we prove the contrapositive: if the affine hull of $\text{im}(f)$ is not a face, then there are g and h with $gf = hf$ but not $g =_{f\text{-a.s.}} h$. We prove the equivalent statement that for any affine subset³⁰ of the probability simplex over Y , there is Z together with morphisms $g, h : Y \rightarrow Z$ which are equal on the affine subset, but not anywhere outside of it. Since the probability simplex over X is bounded, we can ignore the requirement that g and h must again land in a probability simplex, and instead find two such affine maps into some Euclidean space; composing with a suitable invertible contraction and translation

³⁰ An affine subset of the simplex is given by the intersection of the simplex with an affine subspace, or equivalently by suitably many hyperplanes.

will map them back into a finite-dimensional simplex. Now the solution is obvious: for example, take g to be the identity map and h the orthogonal projection onto the affine subspace (with respect to any choice of inner product).

15.4. Example. Now consider completeness in **Stoch**. Here, we claim that $f : X \rightarrow Y$ is complete if and only if for every bounded measurable $\beta : Y \rightarrow \mathbb{R}$,

$$\int_{y \in Y} \beta(y) f(dy|x) = 0 \quad \forall x \quad \implies \quad \beta(y) = 0 \quad f(-|x)\text{-a.s. } \forall x. \quad (15.1)$$

This is essentially the standard definition of *bounded complete*.

To prove the equivalence, suppose first that f is complete in our sense. Then for given β satisfying the vanishing integral assumption, suppose $\beta : Y \rightarrow [0, 1]$ without loss of generality by rescaling. Then with the power set σ -algebra on $\{0, 1\}$, we have a morphism in **Stoch** given by

$$g : Y \rightarrow \{0, 1\}, \quad g(\{1\}|y) := \beta(y),$$

where all other values of g are clearly uniquely specified and the measurability condition on g holds as β is measurable. We now have the composite morphism

$$gf : X \rightarrow \{0, 1\}, \quad (gf)(\{1\}|x) = \int_{y \in Y} \beta(y) f(dy|x) = 0,$$

which coincides with $hf : X \rightarrow \{0, 1\}$ for $h : Y \rightarrow \{0, 1\}$ the deterministic constant zero morphism, $h(\{1\}|y) = 0$ for all y . Therefore the completeness assumption on f implies $g =_{f\text{-a.s.}} h$. Using (13.1) with $T = \{1\}$ only, this is equivalent to

$$\int_{y \in S} \beta(y) f(dy|x) = \int_{y \in S} 0 f(dy|x) = 0$$

for all $S \in \Sigma_Y$. This implies that $\beta(y)$ indeed vanishes $f(-|x)$ -almost surely for every x , proving (15.1).

Conversely, suppose that f is bounded complete in the sense of (15.1). For given $g, h : Y \rightarrow Z$ with $gf = hf$, we will prove that

$$\int_{y \in S} g(T|y) f(dy|x) = \int_{y \in S} h(T|y) f(dy|x) \quad (15.2)$$

for all $x \in X$ and $S \in \Sigma_Y$ and $T \in \Sigma_Z$, which is what needs to be shown by (13.1). For fixed S and T , consider the function

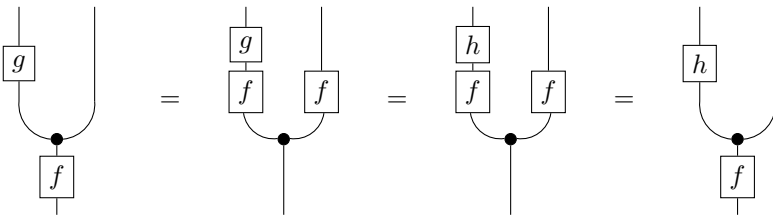
$$\beta : Y \longrightarrow [0, 1], \quad y \longmapsto g(T|y) - h(T|y),$$

which is bounded measurable by construction. Now the integral of β with respect to every $f(-|dx)$ vanishes by the assumption that $gf = hf$, which amounts exactly to equality of the integrals for $S = Y$. Therefore β itself vanishes $f(-|x)$ -almost surely for every x thanks to the assumption (15.1). This immediately implies (15.2).

We have therefore established that our abstract notion of completeness instantiates concretely to the one which is most relevant for measure-theoretic probability. Returning to the general theory, a simple first observation about completeness is as follows.

15.5. Lemma. *Every deterministic morphism is complete.*

Proof. Assuming $gf = hf$ for deterministic f , we compute



which means exactly $g =_{f\text{-a.s.}} h$. $\square$

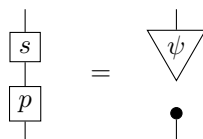
We have not yet studied other properties of complete morphisms, such as their compositionality. A simple use of completeness is to provide a simplified criterion for sufficiency.

15.6. Proposition. *Suppose that $\mathcal{C}$ is strictly positive (Definition 13.16). Let $p : \Theta \rightarrow X$ be a statistical model and $s : X \rightarrow V$ a statistic such that sp is complete. Then s is sufficient if and only if there is $\alpha : V \rightarrow X$ such that $\alpha sp = p$.*

Proof. Applying Theorem 14.5, the condition $s\alpha =_{sp\text{-a.s.}} \text{id}$ follows from $s\alpha sp = sp$ by the completeness assumption. $\square$

The other missing ingredient for Basu's theorem is the notion of *ancillary* statistic, which formalizes the idea that a statistic carries no information at all about the parameters of a statistical model.

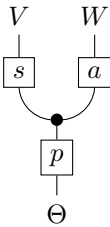
15.7. Definition. A statistic $s : X \rightarrow V$ for a statistical model $p : \Theta \rightarrow X$ is *ancillary* if sp displays $\Theta \perp\!\!\!\perp T \mid I$, which means that



for suitable $\psi : I \rightarrow V$.

Here is our version of Basu’s theorem:

15.8. Theorem. *Suppose that $p : \Theta \rightarrow X$ is a statistical model, and that $s : X \rightarrow V$ and $a : X \rightarrow W$ are statistics. If s is sufficient, the composite morphism sp is complete, and a is ancillary, then the morphism*

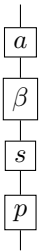


(15.3)

displays $V \perp W \parallel \Theta$.

The morphism (15.3) represents the joint distribution of the statistics s and a , given the model parameter.

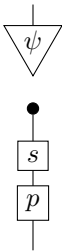
Proof. Let $\beta : V \rightarrow X$ be a splitting of s which witnesses its sufficiency. Then we have



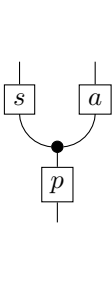
=



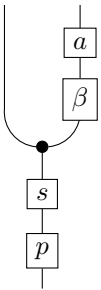
=



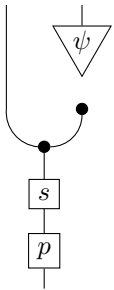
for suitable $\psi : I \rightarrow U$, where the first step is by the sufficiency of s and the second since a is ancillary. Using that sp is complete, we obtain



=



=



=





where the first step is by sufficiency of s and the second by completeness of sp and the previous equation. This factorization implies in particular the desired conditional independence $V \perp W \parallel \Theta$. $\square$

15.9. Remark. Thanks to the characterization of completeness as bounded completeness in *Stoch* (Example 15.4), Theorem 15.8 reproduces essentially the standard version of Basu's theorem in the case of *Stoch*.

16. Minimal sufficient statistics and Bahadur's theorem

While sufficient statistics are useful in that they contain maximal information about the parameters, in practice one also wants a statistic to *discard as much irrelevant information* as possible. This is formalized by the notion of *minimal* sufficient statistic, which this section develops within our framework. We will prove Bahadur's theorem, a classical result giving a sufficient criterion for the minimality of a sufficient statistic in terms of completeness.

Leading up to the definition of minimality, we first investigate how different statistics may determine each other:

16.1. Definition. Let $s : X \rightarrow V$ and $t : X \rightarrow W$ be statistics for a statistical model $p : \Theta \rightarrow X$. Then we write $t \leq s$ if there is $c : V \rightarrow W$ such that $t =_{p\text{-a.s.}} cs$.

Such a c must be sp -a.s. deterministic by Remark 13.17, since s is deterministic and cs is p -a.s. deterministic by Lemma 13.12.

It is easy to see that the $\leq$ -relation on statistics is reflexive and transitive. In this way, the collection of statistics for a given statistical model forms a preordered set,³¹ with the more informative statistics further up. This preorder is reminiscent of the standard theory of subobjects in category theory [87, Section 1.3], but with the direction of arrows reversed. The idea of defining an informativeness ordering of this type in categorical terms is certainly not new and appears e.g. also in the work of Golubtsov [54, Section 5], who has used it to develop aspects of decision theory in purely categorical terms [54, Sections 6 and 7].

We now restrict to the set of sufficient statistics with the induced preorder. Recall that a *least element* s in a preordered set is an element such that $s \leq r$ for all other elements r .

16.2. Definition. A statistic $s : X \rightarrow T$ is *minimal sufficient* if it is a least element in the preordered set of sufficient statistics for the given model $p : \Theta \rightarrow X$.

³¹ Or, depending on how large the category $\mathbf{C}$ is, a proper class equipped with a preorder. We will ignore these size issues here, for which Grothendieck universes provide the standard working solution.

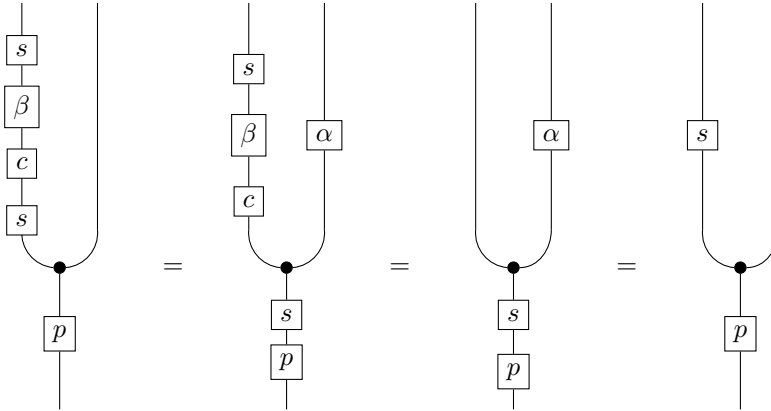
Since any two least elements are ordered in both directions, a minimal sufficient statistic is essentially unique: any two minimal sufficient statistics can be computed from each other in the sense of Definition 16.1.

Here is our version of Bahadur’s theorem, showing that under mild assumptions, a complete sufficient statistic is guaranteed to be minimal.

16.3. Theorem. *Suppose that $\mathcal{C}$ is strictly positive. Let $p : \Theta \rightarrow X$ be a statistical model which has a minimal sufficient statistic. If $s : X \rightarrow V$ is a complete sufficient statistic for p , then s is minimal sufficient.*

Proof. If there is a minimal sufficient statistic, then s must lie above it. By Definition 16.1, this means that there is some $c : V \rightarrow W$ such that cs is minimal. In order to show that s itself is minimal, we therefore only need to show that $s \leq cs$. Since cs is still sufficient, we have a sufficiency witness $\beta : W \rightarrow X$ for cs , satisfying $cs\beta =_{csp\text{-a.s.}} \text{id}$ and $p = \beta csp$ by Theorem 14.5.

Since s is sufficient as well, we have a witness $\alpha : V \rightarrow X$ with $s\alpha =_{sp\text{-a.s.}} \text{id}$ and $p = \alpha sp$. Hence $\alpha sp = \beta csp$, and completeness of sp now implies $\alpha =_{sp\text{-a.s.}} \beta c$, and therefore $s\beta c =_{sp\text{-a.s.}} \text{id}$ by postcomposing with s . We now claim that $s =_{p\text{-a.s.}} s\beta cs$. Indeed,



using sufficiency of s in the first and third step. This proves that $s \leq cs$, as was to be shown. $\square$

We end with two possible avenues for further research.

16.4. Problem. Under which conditions on $\mathcal{C}$ does every statistical model in $\mathcal{C}$ have a minimal sufficient statistic?

We suspect that it is possible to derive the existence of minimal sufficient statistics from the assumption that $\mathcal{C}$ has all colimits and that these colimits interact suitably

nicely with the Markov category structure, in a precise way which still needs to be determined. For then it may be possible to obtain a minimal sufficient statistic by taking the colimit of the diagram consisting of all sufficient statistics out of X for a given statistical model $p : \Theta \rightarrow X$. (A suitable co-well-poweredness assumption may be relevant for this as well.)

16.5. Remark. Following Lauritzen [76], another interesting avenue may be to consider the *inverse problem* of finding all statistical models $p : \Theta \rightarrow X$ for which a given deterministic morphism $s : X \rightarrow V$ is a sufficient statistic.

References

- [1] R.J. Aumann, Borel structures for function spaces, *Ill. J. Math.* 5 (1961) 614–630.
- [2] T. Avery, Codensity and the Giry monad, *J. Pure Appl. Algebra* 220 (3) (2016) 1229–1251.
- [3] G. Bacci, R. Mardare, P. Panangaden, G. Plotkin, An algebraic theory of Markov processes, in: *LICS '18—33rd Annual ACM/IEEE Symposium on Logic in Computer Science*, ACM, New York, 2018, p. 10, cs.mcgill.ca/~prakash/Pubs/Markov-LICS.pdf.
- [4] H. Barnum, J. Barrett, M. Leifer, A. Wilce, A generalized no-broadcasting theorem, *Phys. Rev. Lett.* 99 (2007) 240501, [arXiv:0707.0620](https://arxiv.org/abs/0707.0620).
- [5] D. Basu, On statistics independent of a complete sufficient statistic, *Sankhya* 15 (1955) 377–380. Reprinted at [6, p. 62].
- [6] D. Basu, *Selected Works of Debabrata Basu, Selected Works in Probability and Statistics*, Springer, New York, 2011, edited by Anirban DasGupta.
- [7] P. Billingsley, *Probability and Measure*, third edition, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, Inc., New York, 1995, A Wiley-Interscience Publication.
- [8] V.I. Bogachev, I.I. Malofeev, Kantorovich problems and conditional measures depending on a parameter, *J. Math. Anal. Appl.* 486 (1) (2020) 123883, [arXiv:1904.03642](https://arxiv.org/abs/1904.03642).
- [9] A. Carboni, R.F.C. Walters, Cartesian bicategories. I, *J. Pure Appl. Algebra* 49 (1–2) (1987) 11–32.
- [10] A. Carboni, Matrices, relations, and group representations, *J. Algebra* 136 (2) (1991) 497–529.
- [11] N.N. Čencov, The categories of mathematical statistics, *Dokl. Akad. Nauk SSSR* 164 (1965) 511–514.
- [12] N.N. Čencov, Algebraic foundation of mathematical statistics, *Math. Oper.forsch. Stat., Ser. Stat.* 9 (2) (1978) 267–276.
- [13] N.N. Čencov, *Statistical Decision Rules and Optimal Inference*, Translations of Mathematical Monographs, vol. 53, American Mathematical Society, Providence, RI, 1982, translation from the Russian edited by Lev J. Leifman.
- [14] G. Chiribella, M. D’Ariano, P. Perinotti, Probabilistic theories with purification, *Phys. Rev. A* 81 (2010) 062348, [arXiv:0908.1583](https://arxiv.org/abs/0908.1583).
- [15] K. Cho, B. Jacobs, Disintegration and Bayesian inversion via string diagrams, *Math. Struct. Comput. Sci.* 29 (2019) 938–971, [arXiv:1709.00322](https://arxiv.org/abs/1709.00322).
- [16] K. Cho, B. Jacobs, B. Westerbaan, A. Westerbaan, An introduction to effectus theory, [arXiv:1512.05813](https://arxiv.org/abs/1512.05813).
- [17] Y.S. Chow, H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales*, third edition, Springer Texts in Statistics, Springer-Verlag, New York, 1997.
- [18] B. Coecke, Terminality implies non-signalling, in: *Proceedings of the 11th Workshop on Quantum Physics and Logic*, in: *Electron. Proc. Theor. Comput. Sci. (EPTCS)*, vol. 172, 2014, pp. 27–35, [arXiv:1405.3681](https://arxiv.org/abs/1405.3681).
- [19] B. Coecke, R. Lal, Causal categories: relativistically interacting processes, *Found. Phys.* 43 (4) (2013) 458–501, [arXiv:1107.6019](https://arxiv.org/abs/1107.6019).
- [20] B. Coecke, É.O. Paquette, D. Pavlovic, Classical and quantum structuralism, in: *Semantic Techniques in Quantum Computation*, Cambridge Univ. Press, Cambridge, 2010, pp. 29–69.
- [21] B. Coecke, R.W. Spekkens, Picturing classical and quantum Bayesian inference, *Synthese* 186 (3) (2012) 651–696, [arXiv:1102.2368](https://arxiv.org/abs/1102.2368).

- [22] O. Cunningham, C. Heunen, Purity through factorisation, in: Proceedings 14th International Conference on Quantum Physics and Logic, in: *Electron. Proc. Theor. Comput. Sci. (EPTCS)*, vol. 266, EPTCS, 2018, pp. 315–328, arXiv:1705.07652.
- [23] F. Dahlqvist, A. Silva, V. Danos, I. Garnier, Borel kernels and their approximation, categorically, in: Proceedings of the Thirty-Fourth Conference on the Mathematical Foundations of Programming Semantics, MFPS XXXIV, in: *Electron. Notes Theor. Comput. Sci.*, vol. 341, 2018, pp. 91–119, arXiv:1803.02651.
- [24] A.P. Dawid, Conditional independence in statistical theory, *J. R. Stat. Soc. Ser. B* 41 (1) (1979) 1–31.
- [25] A.P. Dawid, Conditional independence for statistical operations, *Ann. Stat.* 8 (3) (1980) 598–617.
- [26] A.P. Dawid, Separoids: a mathematical framework for conditional independence and irrelevance, *Ann. Math. Artif. Intell.* 32 (1–4) (2001) 335–372, representations of uncertainty.
- [27] A.P. Dawid, M. Studený, Conditional products: an alternative approach to conditional independence, in: *Artificial Intelligence and Statistics '99*, 1999, pp. 32–40, <ftp://ftp.utia.cas.cz/pub/staff/studený/daw-stu-99-pdf.pdf>.
- [28] J.L. Doob, *Stochastic Processes*, Wiley Classics Library, John Wiley & Sons, Inc., New York, 1990, reprint of the 1953 original, a Wiley-Interscience Publication.
- [29] M.J. Druzdzel, H.A. Simon, Causality in Bayesian belief networks, in: *Uncertainty in Artificial Intelligence*, Washington, DC, 1993, Morgan Kaufmann Publishers, 1993, pp. 3–11, arXiv:1303.1454.
- [30] M.L. Eaton, *Multivariate Statistics: A Vector Space Approach*, Institute of Mathematical Statistics Lecture Notes—Monograph Series, vol. 53, Institute of Mathematical Statistics, Beachwood, OH, 2007, reprint of the 1983 original.
- [31] A.M. Faden, The existence of regular conditional probabilities: necessary and sufficient conditions, *Ann. Probab.* 13 (1) (1985) 288–298.
- [32] C. Flori, T. Fritz, Compositories and gleaves, *Theory Appl. Categ.* 31 (33) (2016) 928–988, arXiv:1308.6548.
- [33] B. Fong, Causal theories: a categorical perspective on Bayesian networks, Master's thesis, University of Oxford, 2012, arXiv:1301.6201.
- [34] B. Fong, *The Algebra of Open and Interconnected Systems*, PhD thesis, University of Oxford, 2016, arXiv:1609.05382.
- [35] B. Fong, D. Spivak, Supplying bells and whistles in symmetric monoidal categories, arXiv:1908.02633.
- [36] B. Fong, D. Spivak, Hypergraph categories, *J. Pure Appl. Algebra* 223 (2019) 4746–4777, arXiv:1806.08304.
- [37] U. Franz, What Is Stochastic Independence? In *Non-commutativity, Infinite-Dimensionality and Probability at the Crossroads*, QP–PQ: Quantum Probab. White Noise Anal, vol. 16, World Sci. Publ., River Edge, NJ, 2002, pp. 254–274, arXiv:math/0206017.
- [38] D.H. Fremlin, *Measure Algebras*. Vol. 3: Measure Algebras, Torres Fremlin, Colchester, 2004, corrected second printing of the 2002 original essex.ac.uk/math/people/fremlin/mt.htm.
- [39] D.H. Fremlin, *Measure Theory*. Vol. 4: Topological Measure Spaces. Part I, II, Torres Fremlin, Colchester, 2006, corrected second printing of the 2003 original essex.ac.uk/math/people/fremlin/mt.htm.
- [40] T. Fritz, Beyond Bell's theorem II: scenarios with arbitrary causal structure, *Commun. Math. Phys.* 341 (2) (2016) 391–434, arXiv:1404.4812.
- [41] T. Fritz, P. Perrone, Bimonoidal structure of probability monads, in: Proceedings of the Thirty-Fourth Conference on the Mathematical Foundations of Programming Semantics, MFPS XXXIV, in: *Electron. Notes Theor. Comput. Sci.*, vol. 341, Elsevier Sci. B. V., Amsterdam, 2018, pp. 121–149, arXiv:1804.03527.
- [42] T. Fritz, P. Perrone, S. Rezagholi, Probability, valuations, hyperspaces: three monads on top and the support as a morphism, arXiv:1910.03752.
- [43] T. Fritz, E.F. Rischel, The zero-one laws of Kolmogorov and Hewitt–Savage in categorical probability, arXiv:1912.02769.
- [44] C. Führmann, Varieties of effects, in: *Foundations of Software Science and Computation Structures*, Grenoble, 2002, in: *Lecture Notes in Comput. Sci.*, vol. 2303, Springer, Berlin, 2002, pp. 144–158.
- [45] R. Furber, *Categorical Duality in Probability and Quantum Foundations*, PhD thesis, Radboud University Nijmegen, 2017, people.cs.aau.dk/~furber/phdthesis.pdf.
- [46] R. Furber, B. Jacobs, From Kleisli categories to commutative C*-algebras: probabilistic Gelfand duality, *Log. Methods Comput. Sci.* 11 (2015) 1–28, arXiv:1303.1115.

- [47] M. Gerhold, S. Lachs, M. Schürmann, Categorical Lévy processes, arXiv:1612.05139.
- [48] M. Giry, A categorical approach to probability theory, in: *Categorical Aspects of Topology and Analysis*, Ottawa, Ont., 1980, in: *Lecture Notes in Math.*, vol. 915, Springer, 1982, pp. 68–85.
- [49] P.V. Golubtsov, Informativity in the category of linear measurement systems, *Probl. Pereda. Inf.* 28 (2) (1992) 30–46 (in Russian). English translation in *Probl. Inf. Transm.* 28 (2) (1992) 125–140.
- [50] P.V. Golubtsov, Theory of fuzzy sets as a theory of uncertainty and decision-making problems in fuzzy experiments, *Probl. Pereda. Inf.* 30 (3) (1994) 47–67 (in Russian). English translation in *Probl. Inf. Transm.* 30 (3) (1994) 232–250.
- [51] P.V. Golubtsov, Relative informativity and a priori information in the category of linear information transformers, *Probl. Pereda. Inf.* 31 (3) (1995) 3–23 (in Russian). English translation in *Probl. Inf. Transm.* 31 (3) (1995) 195–215.
- [52] P.V. Golubtsov, Informativity in the category of multivalued information transformers, *Probl. Pereda. Inf.* 34 (3) (1998) 60–80 (in Russian). English translation in *Probl. Inf. Transm.* 34 (3) (1998) 259–276.
- [53] P.V. Golubtsov, Axiomatic description of categories of information converters, *Probl. Pereda. Inf.* 35 (3) (1999) 80–98 (in Russian). English translation in *Probl. Inf. Transm.* 35 (3) (1999) 259–274.
- [54] P.V. Golubtsov, Monoidal Kleisli category as a background for information transformers theory, *Информационные процессы (Information Theory and Information Processing)* 2 (2002) 62–84, translated from Russian jip.ru/2002/GOLU1.pdf.
- [55] P.V. Golubtsov, Information transformers: category-theoretical structure, informativeness, decision-making problems, *Hadronic J. Suppl.* 19 (3) (2004) 375–424.
- [56] P.V. Golubtsov, S.S. Moskaliuk, Method of additional structures on the objects of a monoidal Kleisli category as a background for information transformers theory, *Hadron. J.* 25 (2) (2002) 179–237, arXiv:math-ph/0211067.
- [57] J. Goubault-Larrecq, X. Jia, Algebras of the extended probabilistic powerdomain monad, arXiv:1903.07472.
- [58] A. Grothendieck, Sur quelques points d'algèbre homologique, *Tohoku Math. J.* 2 (9) (1957) 119–221.
- [59] R. Guitart, Tenseurs et machines, *Cah. Topol. Géom. Différ.* 21 (1) (1980) 5–62.
- [60] P.R. Halmos, L.J. Savage, Application of the Radon-Nikodym theorem to the theory of sufficient statistics, *Ann. Math. Stat.* 20 (1949) 225–241.
- [61] R. Heckmann, Spaces of valuations, in: *Papers on General Topology and Applications*, Gorham, ME, 1995, in: *Ann. New York Acad. Sci.*, vol. 806, New York Acad. Sci., New York, 1996, pp. 174–200.
- [62] C. Heunen, O. Kammar, S. Staton, H. Yang, A convenient category for higher-order probability theory, in: 2017 32nd Annual ACM/IEEE Symposium on Logic in Computer Science, LICS, IEEE, [Piscataway], NJ, 2017, p. 12, arXiv:1701.02547.
- [63] C. Heunen, J. Vicary, *Categories for Quantum Theory: An Introduction*, Oxford University Press, 2020.
- [64] M. Hyland, J. Power, Symmetric monoidal sketches, in: *Second International Conference on Principles and Practice of Declarative Programming*, 2000, pp. 280–288, <https://webdpmms.maths.cam.ac.uk/~martin/Research/Pub91-00/hp00.pdf>.
- [65] B. Jacobs, Semantics of weakening and contraction, *Ann. Pure Appl. Log.* 69 (1) (1994) 73–106.
- [66] B. Jacobs, Structured probabilistic reasoning, 2019. Forthcoming book, available at cs.ru.nl/B.Jacobs/PAPERS/ProbabilisticReasoning.pdf.
- [67] B. Jacobs, C. Heunen, I. Hasuo, Categorical semantics for arrows, *J. Funct. Program.* 19 (3–4) (2009) 403–438.
- [68] B. Jacobs, F. Zanasi, A predicate/state transformer semantics for Bayesian learning, in: *The Thirty-Second Conference on the Mathematical Foundations of Programming Semantics, MFPS XXXII*, in: *Electron. Notes Theor. Comput. Sci.*, vol. 325, Elsevier Sci. B. V., Amsterdam, 2016, pp. 185–200.
- [69] J. Jost, H.V. Lê, D.H. Luu, T.D. Tran, Probabilistic mappings and Bayesian nonparametrics, arXiv:1905.11448.
- [70] A. Joyal, R. Street, The geometry of tensor calculus. I, *Adv. Math.* 88 (1) (1991) 55–112.
- [71] Y. Kifer, *Ergodic Theory of Random Transformations*, Progress in Probability and Statistics, vol. 10, Birkhäuser Boston, Inc., Boston, MA, 1986.
- [72] A. Kissinger, Finite matrices are complete for (dagger-)hypergraph categories, arXiv:1406.5942.
- [73] A. Kissinger, M. Hoban, B. Coecke, Equivalence of relativistic causal structure and process terminality, arXiv:1708.04118.

- [74] H. Kleisli, Every standard construction is induced by a pair of adjoint functors, *Proc. Am. Math. Soc.* 16 (1965) 544–546.
- [75] A. Klenke, *Probability Theory: A Comprehensive Course*, second edition, Universitext., Springer, London, 2014.
- [76] S.L. Lauritzen, *Extremal Families and Systems of Sufficient Statistics*, *Lecture Notes in Statistics*, vol. 49, Springer-Verlag, New York, 1988.
- [77] S.L. Lauritzen, *Graphical Models*, *Oxford Statistical Science Series*, vol. 17, The Clarendon Press, Oxford University Press, New York, 1996, Oxford Science Publications.
- [78] S.L. Lauritzen, F. Jensen, Stable local computation with conditional Gaussian distributions, *Stat. Comput.* 11 (2) (2001) 191–203.
- [79] B. Lawvere, The category of probabilistic mappings, 1962, ncatlab.org/nlab/files/lawvereprobability1962.pdf.
- [80] E.L. Lehmann, *Selected Works of E.L. Lehmann, Selected Works in Probability and Statistics*, Springer, New York, 2012, edited by Javier Rojo.
- [81] E.L. Lehmann, H. Scheffé, Completeness, similar regions, and unbiased estimation. I, *Sankhya* 10 (1950) 305–340. Reprinted at [80, p. 239].
- [82] T. Leinster, Monoidal categories with projections, 2016, *n-Category Café* blog post golem.ph.utexas.edu/category/2016/08/monoidal_categories_with_proje.html.
- [83] T. Leinster, Monoidal categories with projections, 2016, *n-Category Café* comment golem.ph.utexas.edu/category/2016/08/monoidal_categories_with_proje.html#c051133.
- [84] P.B. Levy, J. Power, H. Thielecke, Modelling environments in call-by-value programming languages, *Inf. Comput.* 185 (2) (2003) 182–210.
- [85] R.B.B. Lucyshyn-Wright, Functional distribution monads in functional-analytic contexts, *Adv. Math.* 322 (2017) 806–860, [arXiv:1701.08152](https://arxiv.org/abs/1701.08152).
- [86] S. Mac Lane, *Categories for the Working Mathematician*, second edition, *Graduate Texts in Mathematics*, vol. 5, Springer-Verlag, New York, 1998.
- [87] S. Mac Lane, I. Moerdijk, *Sheaves in Geometry and Logic: A First Introduction to Topos Theory*, Universitext., Springer-Verlag, New York, 1994, corrected reprint of the 1992 edition.
- [88] E.G. Manes, A class of fuzzy theories, *J. Math. Anal. Appl.* 85 (2) (1982) 409–451.
- [89] M.S. Massa, S.L. Lauritzen, Combining statistical models, in: *Algebraic Methods in Statistics and Probability II*, in: *Contemp. Math.*, vol. 516, Amer. Math. Soc., Providence, RI, 2010, pp. 239–259, web.math.ku.dk/~lauritzen/papers/conm10179.pdf.
- [90] P. McCullagh, What is a statistical model?, *Ann. Stat.* 30 (5) (2002) 1225–1310, with comments and a rejoinder by the author.
- [91] E.A. Morozova, N.N. Čencov, Markov invariant geometry on state manifolds, in: *Current Problems in Mathematics. Newest Results*, vol. 36 (Itogi Nauki i Tekhniki), Akad. Nauk SSSR, Vsesoyuz. Inst. Nauchn. i Tekhn. Inform., Moscow, 1989, pp. 69–102, 187 (in Russian). Translated in *J. Sov. Math.* 56 (5) (1991) 2648–2669.
- [92] N. Morse, R. Sacksteder, Statistical isomorphism, *Ann. Math. Stat.* 37 (1966) 203–214.
- [93] J. Neveu, *Bases mathématiques du calcul des probabilités*, Préface de R. Fortet. Deuxième édition, revue et corrigée, Masson et Cie, Éditeurs, Paris, 1970.
- [94] P. Panangaden, *Labelled Markov Processes*, Imperial College Press, London, 2009.
- [95] J. Pearl, *Causality: Models, Reasoning, and Inference*, second edition, Cambridge University Press, Cambridge, 2009.
- [96] J. Pearl, A. Paz, Graphoids: a graph-based logic for reasoning about relevance relations, Technical report, 1985, ftp.cs.ucla.edu/pub/stat_ser/r53-L.pdf.
- [97] P. Perrone, *Categorical Probability and Stochastic Dominance in Metric Spaces*, PhD thesis, Max Planck Institute for Mathematics, in the Sciences, 2018, paoloperrone.org/phdthesis.pdf.
- [98] N. Popesco, P. Gabriel, Caractérisation des catégories abéliennes avec générateurs et limites inductives exactes, *C. R. Acad. Sci. Paris* 258 (1964) 4188–4190.
- [99] K.P.S.B. Rao, B.V. Rao, Borel Spaces, *Dissertationes Math. (Rozprawy Mat.)*, vol. 190, 1981, p. 63, eudml.org/doc/268562.
- [100] N. Saavedra Rivano, *Catégories Tannakiennes*, *Lecture Notes in Mathematics*, vol. 265, Springer-Verlag, Berlin-New York, 1972.
- [101] R. Sacksteder, A note on statistical equivalence, *Ann. Math. Stat.* 38 (1967) 787–794.
- [102] A. Šcibior, O. Kammar, M. Vákár, S. Staton, H. Yang, Y. Cai, K. Ostermann, S.K. Moss, C. Heunen, Z. Ghahramani, Denotational validation of higher-order Bayesian inference, in: *Proc. ACM Program. Lang.*, vol. 2, ACM, 2017, 60, [arXiv:1711.03219](https://arxiv.org/abs/1711.03219).

- [103] G.J. Seal, Tensors, monads and actions, *Theory Appl. Categ.* 28 (15) (2013) 403–433, arXiv:1205.0101.
- [104] P. Selinger, A survey of graphical languages for monoidal categories, in: *New Structures for Physics*, in: *Lecture Notes in Phys.*, vol. 813, Springer, Heidelberg, 2011, pp. 289–355.
- [105] M. Shulman, Framed bicategories and monoidal fibrations, *Theory Appl. Categ.* 20 (18) (2008) 650–738, arXiv:0706.1286.
- [106] A. Simpson, Probability sheaves and the Giry monad, <http://drops.dagstuhl.de/opus/volltexte/2017/8051/pdf/LIPIcs-CALCO-2017-1.pdf>, 2017.
- [107] A. Simpson, Category-theoretic structure for independence and conditional independence, in: *The Thirty-Third Conference on the Mathematical Foundations of Programming Semantics, MFPS XXXIII*, in: *Electron. Notes Theor. Comput. Sci.*, vol. 336, Elsevier Sci. B. V., Amsterdam, 2018, pp. 281–297, coalg.org/mfps-calco2017/mfps-papers/6-simpson.pdf.
- [108] S. Staton, D. Stein, H. Yang, N.L. Ackerman, C.E. Freer, D.M. Roy, The beta-Bernoulli process and algebraic effects, in: I. Chatzigiannakis, C. Kaklamanis, D. Marx, D. Sannella (Eds.), *45th International Colloquium on Automata, Languages, and Programming, ICALP 2018*, in: *Leibniz International Proceedings in Informatics (LIPIcs)*, vol. 107, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2018, pp. 1–15, arXiv:1802.09598.
- [109] M. Studený, Multi-information and the problem of characterization of conditional independence relations, *Probl. Control Inf. Theory (Probl. Upr. Teor. Inform.)* 18 (1) (1989) 3–16, ftp.utia.cas.cz/pub/staff/studený/multiinformation-PCIT-89.pdf.
- [110] M. Studený, *Probabilistic Conditional Independence Structures*, Information Science and Statistics, Springer, London, 2005.
- [111] T. Świrszcz, Monadic functors and convexity, *Bull. Acad. Pol. Sci., Sér. Sci. Math. Astron. Phys.* 22 (1974) 39–42.
- [112] H. Thielecke, *Categorical Structure of Continuation Passing Style*, PhD thesis, University of Edinburgh, 1997, <https://www.cs.bham.ac.uk/~hxt/research/THESIS-FINAL.ps>.
- [113] R.M. Wald, *General Relativity*, University of Chicago Press, Chicago, IL, 1984.