Representable Markov Categories and Comparison of Statistical Experiments in Categorical Probability

Tobias Fritz*[1], Tomáš Gonda†[2,3], Paolo Perrone‡[4], and Eigil Fjeldgren Rischel§[5]

[1]Department of Mathematics, University of Innsbruck, Austria
[2]Perimeter Institute for Theoretical Physics, Waterloo ON, Canada
[3]School of Physics and Astronomy, University of Waterloo, Canada
[4]Massachusetts Institute of Technology, Cambridge MA, U.S.A.
[5]University of Strathclyde, Glasgow, Scotland

May 9, 2023

Abstract

Markov categories are a recent categorical approach to the mathematical foundations of probability and statistics. Here, this approach is advanced by stating and proving equivalent conditions for second-order stochastic dominance, a widely used way of comparing probability distributions by their spread. Furthermore, we lay the foundation for the theory of comparing statistical experiments within Markov categories by stating and proving the classical Blackwell–Sherman–Stein Theorem. Our version not only offers new insight into the proof, but its abstract nature also makes the result more general, automatically specializing to the standard Blackwell–Sherman–Stein Theorem in measure-theoretic probability as well as a Bayesian version that involves prior-dependent garbling. Along the way, we define and characterize *representable* Markov categories, within which one can talk about Markov kernels to or from spaces of distributions. We do so by exploring the relation between Markov categories and Kleisli categories of probability monads.

Keywords—Categorical probability; Markov category; Kleisli category; Blackwell–Sherman–Stein Theorem; Second-order stochastic dominance; Comparison of statistical experiments

*tobias.fritz@uibk.ac.at
†tomas.gonda@uibk.ac.at
‡paolo.perrone@cs.ox.ac.uk
§eigil.rischel@strath.ac.uk

This is a peer reviewed, accepted author manuscript of the following research article: Fritz, T., Gonda, T., Perrone, P., & Fjeldgren Rischel, E. (2023). Representable Markov categories and comparison of statistical experiments in categorical probability. *Theoretical Computer Science*, 961, [113896]. https://doi.org/10.1016/j.tcs.2023.113896

Contents

1 Introduction

Traditionally, the foundations of mathematical statistics are rooted in measure theory and measure-theoretic probability. More generally, mathematical statistics and probability theory are typically considered as mathematical subjects of a clearly analytical nature. While this has worked well in practice, it is also often the case in mathematics that higher abstraction leads ultimately to deeper understanding, greater generality and ultimately facilitates the development of results and methods of greater complexity.

This is what the growing field of categorical probability attempts to do by developing a category-theoretical foundation for probability theory and mathematical statistics. A promising approach is provided by *Markov categories* which, in line with categorical thinking, focuses on the morphisms involved in probabilistic reasoning, namely stochastic maps (or Markov kernels). There is growing evidence that Markov categories can serve both as a categorical foundation for, as well as a generalization of, ordinary measure-theoretic probability theory. Indeed, similar to how a computer can be programmed either in terms of low-level machine code or in a more accessible and hardware-independent abstract language, it seems to be the case that probability theory likewise can be practiced either in concrete analytical terms based on Kolmogorov's axioms, or in a more abstract *synthetic* form based on the structural axioms of Markov categories.

More specifically, Markov categories allow one to study and make use of:

- Bayes' theorem and Bayesian updating: This was first considered by Golubtsov in [20] and rediscovered recently by Cho and Jacobs [9], with further results on the dagger functor structure of Bayesian inversion in the first named author's [13].
- Conditional independence: This was also defined within this framework by Cho and Jacobs [9], and more generally in [13, Section 12].
- Almost sure equality: Again, first done by Cho and Jacobs [9] and then generalized and developed further in [13, Section 13].
- Sufficient statistics: Some of the basic theorems on sufficient statistics were proven abstractly in [13, Sections 14–16].
- Kolmogorov extension theorem and 0/1-laws: The *Kolmogorov products* developed by the first-named and last-named author, which arise as infinite products in Markov categories formalizing the Kolmogorov extension theorem, have facilitated synthetic proofs of the classical 0/1-laws of Kolmogorov and Hewitt–Savage [16].
- patterson2020models has developed an algebraic approach to statistical models, drawing and exploiting relations to categorical logic [29].

Of course, this only lists those aspects of probability theory and statistics which have been developed synthetically up to the present time and to our knowledge. The present paper has two goals: first, to continue the development of the general categorical theory; and second, to add two more items to the above list, namely second-order stochastic dominance and the classical Blackwell–Sherman–Stein Theorem on the comparison of statistical experiments. We now summarize our results on both of these goals, which are related through the latter applications drawing on the former categorical developments.

Outline and results. In practice, Markov categories often arise as Kleisli categories of affine symmetric monoidal monads. For example, this is the case for $\mathsf{BorelStoch}$, the category of standard Borel spaces and measurable Markov kernels, or equivalently the Kleisli category of the Giry monad on the category of standard

Borel spaces and measurable maps.[1] In Section 3, we clarify the relation between Markov categories and Kleisli categories of this type, namely Kleisli categories of affine symmetric monoidal monads on categories with finite products. We find that for a Markov category C, the question of whether C arises as a Kleisli category like this is closely linked to the existence of a right adjoint to the inclusion functor $\mathsf{C}_{\mathrm{det}} \hookrightarrow \mathsf{C}$, where $\mathsf{C}_{\mathrm{det}}$ is the cartesian monoidal subcategory of deterministic morphisms in C; for if C is supposed to be the Kleisli category of a monad on $\mathsf{C}_{\mathrm{det}}$, then this right adjoint must exist for purely formal reasons. The existence of such $P\colon \mathsf{C} \to \mathsf{C}_{\mathrm{det}}$ amounts to the natural bijection

$$\mathsf{C}_{\mathrm{det}}(A, PX) \cong \mathsf{C}(A, X), \tag{1.1}$$

which we interpret as the existence of a *distribution functor* that, for $A = I$, identifies the deterministic morphisms $I \to PX$ with the "distributions" over X (morphisms $I \to X$). More generally, (not necessarily deterministic) morphisms $A \to X$ can be thought of as being "classified" by deterministic morphisms $A \to PX$. Not every Markov category has such a distribution functor. For example, the category of finite sets and stochastic matrices $\mathsf{FinStoch}$ does not, since $\mathsf{FinStoch}(A, X)$ is generically infinite, while instead its putative counterpart $\mathsf{FinStoch}_{\mathrm{det}}(A, PX)$ would necessarily have to be finite.

Our results on the connection between Markov categories and Kleisli categories are then as follows:

- We prove that if P is an affine symmetric monoidal monad on a cartesian monoidal category D, and P satisfies a certain pullback condition, then the Kleisli category $\mathsf{Kl}(P)$ is a Markov category such that the subcategory of deterministic morphisms $\mathsf{Kl}(P)_{\mathrm{det}}$ is exactly the original category D.
- Conversely, if a Markov category C has a distribution functor P, meaning a right adjoint for the inclusion $\mathsf{C}_{\mathrm{det}} \hookrightarrow \mathsf{C}$, then the induced monad on $\mathsf{C}_{\mathrm{det}}$ satisfies the pullback condition, and C is isomorphic to the Kleisli category of P (Theorem 3.19).[2]

We end Section 3 by studying the interaction between the distribution functor P and the notion of almost sure equality in C. If these are compatible in a suitable sense, then we say that C is *a.s.-compatibly representable.* Distribution functors and a.s.-compatible representability will then play a central role in the subsequent two sections that focus on applications of the general theory.

[1] Or yet equivalently Polish spaces and measurable maps.

[2] A similar reconstruction of strong monads from their Kleisli categories seems to be known [5]. Nevertheless, our result is not an immediate consequence of this construction.

In Section 4, we provide a categorical description and generalization of *second-order stochastic dominance*, which is a way of comparing probability distributions with respect to how "spread out" they are. This notion also appears in Blackwell–Sherman–Stein (BSS) theorem, a classical and widely used fundamental result that connects it to the question of comparing statistical experiments in terms of their informativeness about the tested hypotheses.

In Section 5, we introduce the informativeness preorder in Markov categories, and prove Theorem 5.4 that characterizes it in terms of the notions of sufficient statistics and conditional independence. Most of this section, however, is devoted to a categorical version of the BSS theorem. In fact, we have a few variations thereof. The closest to the standard version that concerns a *discrete* parameter space is Corollary 5.15, but it more generally applies to any a.s.-compatibly representable Markov category other than $\mathsf{BorelStoch}$. In our presentation, this result arises as a corollary of Theorem 5.13 for more general parameter spaces, which considers a fixed prior distribution and compares experiments with respect to whether they are "almost surely more informative". Concretely, in the context of standard Borel spaces, Theorem 5.13 says the following.

Theorem. *Let X, Y and Θ be standard Borel spaces, and let $(f_\theta)_{\theta\in\Theta}$ and $(g_\theta)_{\theta\in\Theta}$ be families of probability measures on X and Y respectively, parametrized measurably in θ. Let m be a probability measure on Θ. Then the following are equivalent:*

1. *There is a Markov kernel $c\colon X \to Y$ such that $g_\theta = cf_\theta$ holds for m-almost all θ.*

2. *The standard measures[3] $\hat{f}_m$ and $\hat{g}_m$ (probability measures on $P\Theta$—the space of probability measures) are such that $\hat{g}_m$ second-order dominates $\hat{f}_m$.*

In this formulation of the BSS Theorem, we do not need to assume that the parameter space Θ be finite or even countable.

We then present a completely prior-independent version of the BSS theorem in Section 5.4. This result avoids the need for a prior by effectively considering all priors at once. In our categorical formulation, it turns out to be a special case of the earlier Theorem 5.13; but when instantiated in $\mathsf{BorelStoch}$, we obtain the following statement.

Theorem. *Let X, Y and Θ be standard Borel spaces, and let $(f_\theta)_{\theta\in\Theta}$ and $(g_\theta)_{\theta\in\Theta}$ be families of probability measures on X and Y respectively, parametrized measurably in θ. Then the following are equivalent:*

[3] Standard measures have been introduced in [2]. Here, we provide a synthetic definition in Section 5.3.

1. *There is a family of Markov kernels $(c_m\colon X \to Y)_{m\in P\Theta}$, depending measurably on the prior m, such that $g_\theta = cf_\theta$ holds for m-almost all θ and all $m \in P\Theta$.*
2. *The standard measures $\hat{f}_m$ and $\hat{g}_m$ are such that $\hat{g}_m$ second-order dominates $\hat{f}_m$ for every choice of prior $m \in P\Theta$, as witnessed by a family of dilations $(t_m)_{m\in P\Theta}$ that depend measurably on m.*

Moreover, as we show in Proposition 5.19, these conditions are not in general equivalent to f being more informative than g with respect to a *prior-independent* garbling map.

Outlook. Given the relevance of the theory of comparison of experiments in a wide array of situations, such as hypothesis testing or error correction, proving versions of celebrated results—such as the BSS Theorem—in the abstract context of Markov categories leads to a greater level of generality which has the potential for new domains of applications. The understanding of these results in a synthetic way also has the potential to overcome some of the limitations of the standard approaches, such as the discreteness of the parameter spaces involved.

With the recent development of quantum Markov categories [28], it is conceivable that one could obtain a synthetic version of the quantum BSS Theorem [7] and related results, with potential applications to quantum hypothesis testing or quantum error correction.

Finally, the categorical approach also lends itself to the considerations of variants of the theory in which additional restrictions are placed on the garbling maps. For example, such variations can be studied under the hood of resource theories of distinguishability as introduced in [21, Appendix C]. Many interesting restrictions arise from requiring equivariance of the garbling maps with respect to group actions. Others include adaptive garbling maps or garbling via independent action of multiple agents, both of which are considered in [11]. Although we have not done this yet, it should be straightforward to instantiate Theorem 5.13 and Corollary 5.20 in suitable categories, so as to obtain measure-theoretic BSS theorems which apply in such contexts.

Acknowledgments. We thank Robert Furber for helpful feedback on measure-theoretic aspects, Luciano Pomatto for helpful feedback on a draft, Jean-Simon Pacaud Lemay for pointers to the literature, and an anonymous referee for additional detailed feedback on an earlier version. Research for the first author is supported by FWF (Austrian Science Fund) P 35992-N. Research for the third author is funded by AFOSR grants FA9550-19-1-0113 and FA9550-17-1-0058. Research for the second author is supported by NSERC Discovery grant RGPIN 2017-04383, and by the Perimeter Institute for Theoretical Physics. Research at Perimeter Institute is supported in part by the Government of Canada through the Department of Innovation, Science and Economic Development Canada and by the Province of Ontario through the Ministry of Colleges and Universities.

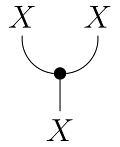
2 Markov Categories

2.1 Definition of Markov Categories and Basic Theory

We now recall the definition of Markov category. As far as we know, it was first proposed by Golubtsov as *category of information transformers* in slightly different form [20], used implicitly in Fong's work on Bayesian networks [12], and rediscovered recently by Cho and Jacobs as *affine CD-categories* [9]. The simpler term *Markov category* was subsequently coined in [13], based on the idea that Markov categories are abstract generalizations of the category of Markov kernels.

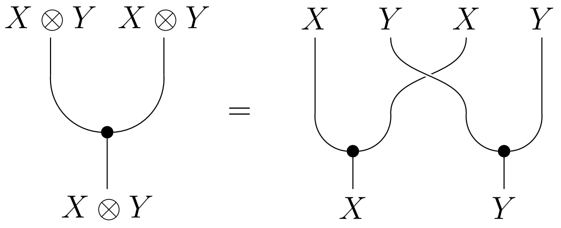
Definition 2.1. *A* Markov category C *is a semicartesian*[4] *symmetric monoidal category where every object* $X \in \mathsf{C}$ *is equipped with a distinguished morphism*

$\mathrm{copy}_X \quad =$



(2.1)

which, together with the unique morphism $\mathrm{del}_X \colon X \to I$*, makes* X *into a commutative comonoid, and such that*



(2.2)

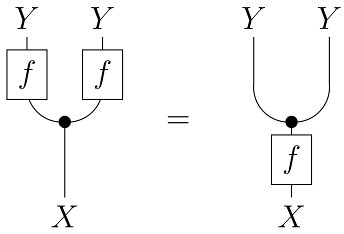
for all $X, Y \in \mathsf{C}$.

Throughout this manuscript, C denotes a Markov category.

[4] Recall that this means that the monoidal unit object I is terminal in C, among several equivalent characterizations; see [19, Theorem 3.5].

Among the prototypical examples of a Markov category is $\mathsf{BorelStoch}$, the category of standard Borel spaces and measurable Markov kernels. A more basic example is $\mathsf{FinStoch}$, the category of finite sets and stochastic matrices. In both cases, the comultiplications $\mathrm{copy}_X \colon X \to X \otimes X$ are given by the diagonals $x \mapsto \delta_{(x,x)}$, assigning to every element $x \in X$ the Dirac delta distribution $\delta_{(x,x)}$; this is the stochastic way to talk about copying. Other examples of Markov categories can be obtained from categories of relations, such as Rel, by restricting to relations $R \colon X \rightsquigarrow Y$ which have the property that for every $x \in X$ there is $y \in Y$ with xRy; this is the relational analogue of the normalization of probability. This results in a Markov category with respect to the usual cartesian product as monoidal structure, and the copy maps are again given by the obvious diagonals. Another interesting class of examples arises by noting that diagram categories of Markov categories are again Markov categories (when suitably defined [13, Section 7]), and we expect that this can be used in future work as a basis for a synthetic theory of stochastic processes.

Definition 2.2. *A morphism $f \colon X \to Y$ in C is* deterministic *if it respects the copy maps:*



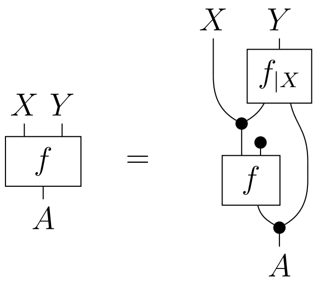
(2.3)

The subcategory of C that consists of its deterministic morphisms is denoted by $\mathsf{C}_{\mathrm{det}}$.

This type of condition goes back to the seminal paper of Carboni and Walters on cartesian bicategories [8]. Intuitively, it means that applying f to two independent copies of its input is guaranteed to result in the same pair of output values than applying f directly to the input and copying its output. C is a cartesian monoidal category with respect to the monoidal structure inherited from C, and all structure morphisms of C, including the copy maps, are in $\mathsf{C}_{\mathrm{det}}$ [13, Remark 10.13].

Other key notions within Markov categories that we use in Sections 4 and 5 include conditionals, Bayesian inverses, almost sure equality, and domination (in the sense of absolute continuity). All but the last of these notions have been introduced in earlier works [9, 13]. We now recall their definitions.

Definition 2.3. *Given $f \colon A \to X \otimes Y$ in C, a morphism $f_{|X} \colon X \otimes A \to Y$ in C is* *called a* conditional *of f with respect to X if the equation*



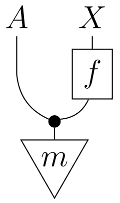
(2.4)

holds. We say that C has conditionals *provided that such a conditional exists for all objects $A, X, Y \in \mathsf{C}$ and all $f \colon A \to X \otimes Y$ in C.*

We can also consider conditionals of $f \colon A \to X \otimes Y$ with respect to Y, which are defined in the analogous way. Using the symmetry of C shows that these automatically exist if C has conditionals.

Example 2.4. $\mathsf{BorelStoch}$ has conditionals [13, Example 11.7]. As far as we know at the moment, the earliest reference for this measure-theoretic fact is in Kallenberg's textbook on random measures [23, Theorem 1.25].

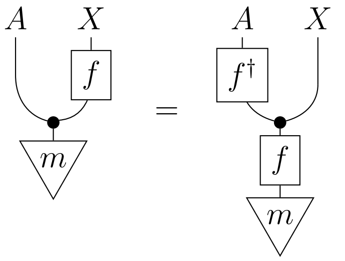
Definition 2.5. *Given two morphisms $m \colon I \to A$ and $f \colon A \to X$, a* Bayesian inverse *of f with respect to (the prior) m is a conditional of*



(2.5)

with respect to X.

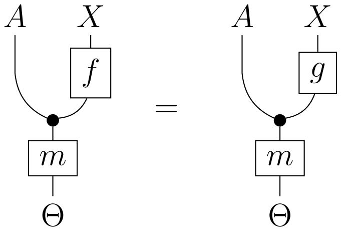
The choice of a prior is often clear from context, so we denote a Bayesian inverse of f simply by $f^\dagger \colon X \to A$ with the dependence on m left implicit. Thus a Bayesian inverse $f^\dagger$ is defined to be a morphism satisfying the equation:



(2.6)

Even though conditionals and Bayesian inverses are generally not unique when they exist, it is clear from the definition that they *are* unique up to almost sure equality [13, Proposition 13.6], which in general is defined as follows.

Definition 2.6. *Given any morphism $h\colon A \to X$, we say that any two parallel $f, g\colon X \to Y$ are h-almost surely equal, denoted by $f =_{h\text{-a.s.}} g$, if we have*



(2.7)

Example 2.7. In the context of $\mathsf{BorelStoch}$, Definition 2.6 recovers the expected notion of equality almost surely as has been shown in [9, Proposition 5.4]. In particular, given Markov kernels $f, g\colon X \to Y$ and $\nu\colon I \to X$, the relation $f =_{\nu\text{-a.s.}} g$ means exactly that for all $S \in \Sigma_X$ and $T \in \Sigma_Y$, we have

$$\int_S f(T|x)\,\nu(dx) = \int_S g(T|x)\,\nu(dx), \tag{2.8}$$

or equivalently that the integrands $f(T|\_)$ and $g(T|\_)$ are ν-almost everywhere equal for all T.

The following notion of measure domination is new in the context of Markov categories. We consider this definition tentative for the moment; we will be using it in this form in the present paper, but note that we may adopt a different variant of this definition in future work.

Definition 2.8. *Given two morphisms $\mu, \nu\colon I \to X$, we say that μ is* absolutely continuous *with respect to ν, denoted $\nu \gg \mu$ or $\mu \ll \nu$, if for all objects Y and all morphisms $f, g\colon X \to Y$ we have*

$$f =_{\nu\text{-a.s.}} g \qquad \Longrightarrow \qquad f =_{\mu\text{-a.s.}} g. \tag{2.9}$$

Example 2.9. In $\mathsf{BorelStoch}$, Definition 2.8 recovers the standard notion of domination of probability measures (also known as absolute continuity preorder), given by the condition that for all measurable sets $S \in \Sigma_X$, we have

$$\nu(S) = 0 \qquad \Longrightarrow \qquad \mu(S) = 0. \tag{2.10}$$

To prove that this is indeed the case, suppose first that condition (2.10) holds. One can then replace ν with μ in equation (2.8), so that $f =_{\nu\text{-a.s.}} g$ indeed implies $f =_{\mu\text{-a.s.}} g$ as necessary to conclude $\nu \gg \mu$ according to Definition 2.8.

In the converse direction, suppose that $\nu \gg \mu$ holds in the sense of Definition 2.8, and that $\nu(S) = 0$ for some $S \in \Sigma_X$. Consider f and g to be the Markov kernels associated to the measurable functions $1_S \colon X \to \{0,1\}$ and $X \to \{0,1\}$, $x \mapsto 0$ respectively. Then we have $f =_{\nu\text{-a.s.}} g$ by $\nu(S) = 0$. However, together with $\nu \gg \mu$ this gives $f =_{\mu\text{-a.s.}} g$, which is just a different way to write $\mu(S) = 0$ given our choice of f and g.

2.2 Parametric Markov Categories

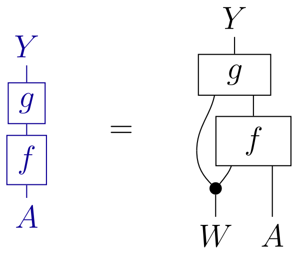
In order to demonstrate the power of the synthetic treatment of the notions of second-order stochastic dominance and comparison of statistical experiments later, we use the following new class of Markov categories throughout this paper.

Given any Markov category C and any object $W \in \mathsf{C}$, we now define a new Markov category C_W which we call the *Markov category parametrized by* W, or simply a *parametric Markov category* when referring to no particular choice of W. This is essentially a known construction for symmetric monoidal categories that has been called *comonoid indexing* [22].

The objects of C_W coincide with those of C, and its morphisms $A \to X$ are defined to be precisely the morphisms $W \otimes A \to X$ in C, that is

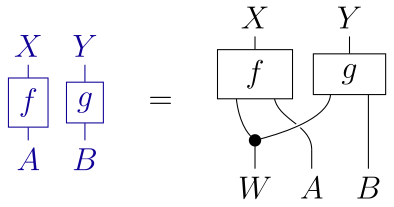
$$\mathsf{C}_W(A, X) \coloneqq \mathsf{C}(W \otimes A, X). \tag{2.11}$$

We think of the object W as playing the role of a "parameter space" which indexes a family of morphisms $A \to X$. In order to distinguish notationally between morphisms $A \to X$ in C_W and their representatives $W \otimes A \to X$ in C, we use blue colored text and diagrams whenever the former representation is used, but otherwise use the same symbol to denote the two. The composition of morphisms in C_W is defined by distributing the parameter in W via the copy map copy_W:



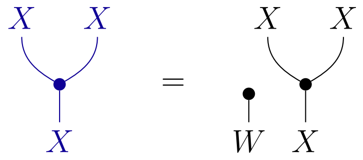
(2.12)

The tensor product of morphisms in C_W is likewise defined by supplying copies of W to the respective morphisms,



(2.13)

and with the monoidal structure morphisms being precisely those of C itself. The discarding operation del_X in C_W just consists of discarding both W and X. Finally, the copying in C_W also discards the parameter,



(2.14)

It is then straightforward to verify that C_W is indeed also a Markov category.

We can alternatively think of C_W as the co-Kleisli category of the reader comonad[5] $W \otimes \_$ on C (see for example [30, Section 5.3]). Note that, while the reader comonad is usually defined on cartesian monoidal categories, the only property of cartesian monoidal categories that is actually used in the definition is that the object W has a comonoid structure, and thus this co-Kleisli category still makes sense in our context.

Lemma 2.10. *If C has conditionals, then so does every parametric Markov category C_W.*

Proof. If $f \colon A \to X \otimes Y$ is a morphism in C_W represented by $f \colon W \otimes A \to X \otimes Y$ in C, then every conditional $f_{|X} \colon X \otimes W \otimes A \to Y$ of f with respect to X represents a conditional $f_{|X}$ of f in C_W upon permuting its input factors to $W \otimes (X \otimes A)$. □

3 Representable Markov Categories

3.1 Kleisli Categories as Markov Categories

It was argued by Kock [25] that affine commutative monads provide a convenient categorical framework for theories of distributions. The following result, which is a

[5] Depending on the literature, this is also known as "writer comonad", since its underlying functor is the same as the writer monad in case W is a monoid object, as well as "product comonad".

special case of [13, Proposition 3.1] gives one direction of the connection between this monadic approach and Markov categories.

Recall first that a monad P on a category with a terminal object I is called *affine* if $P(I) \cong I$ holds. Since commutative monads and symmetric monoidal monads are equivalent concepts [6, Proposition 6.3.5], the following result can be viewed as taking a variant of Kock's framework as its starting point.

Note that term "commutative monad" is more commonly used than the equivalent notion of a "symmetric monoidal monad", especially in the computer science literature. However, we prefer working with the latter because its monoidal structure maps given in (3.1) have a clear probabilistic interpretation. Intuitively, if $\mu \in PX$ and $\nu \in PY$ are probability distributions, then $\nabla(\mu, \nu) \in P(X \times Y)$ can be thought of as the corresponding product distribution (see equation (3.7)).

Proposition 3.1. *Let D be a cartesian monoidal category, and let (P, E, δ) be an affine symmetric monoidal monad on D with unit δ, multiplication E, and monoidal structure maps*

$$\nabla \colon P(\_) \times P(\_) \to P(\_ \times \_). \tag{3.1}$$

Then the Kleisli category $\mathsf{Kl}(P)$ is a Markov category with respect to the following pieces of structure:

- *The monoidal structure on objects is given by products in D, and the monoidal product of Kleisli morphisms $f \colon A \to PX$ and $g \colon B \to PY$ represented by the composite*

$$A \times B \xrightarrow{f \times g} PX \times PY \xrightarrow{\nabla} P(X \times Y),$$

- *The copy maps copy_X are represented by the overall composite of the diagram*

$$\begin{tikzcd}
X \arrow[r, "\delta"] \arrow[d, "{(\mathrm{id},\mathrm{id})}"'] & PX \arrow[d, "{(\mathrm{id},\mathrm{id})}"] \\
X \times X \arrow[r, "\delta \times \delta"] \arrow[dr, "\delta"'] & PX \times PX \arrow[d, "\nabla"] \\
& P(X \times X)
\end{tikzcd} \tag{3.2}$$

Note that the upper square in equation (3.2) commutes trivially, while the lower triangle commutes as one of the defining properties of monoidal monads.

Example 3.2. This construction reproduces $\mathsf{BorelStoch}$ as the Kleisli category of the Giry monad on the category of standard Borel spaces and measurable maps; the definition of the copy maps reproduces exactly the maps $x \mapsto \delta_{(x,x)}$ described above.

Example 3.3. Let $(R, +, \cdot, 0, 1)$ be a commutative semiring, i.e. a set R equipped with algebraic structure like that of a commutative ring except for the assumption of additive inverses. Then R induces an affine symmetric monoidal monad D_R on Set, given by the R-linear combinations monad together with a normalization constraint. This is spelled out, for example, in [10, Section 5.1], which we recall here.

For each set X, denote by $D_R X$ the set of functions $p\colon X \to R$ which are nonzero on finitely many elements, and such that the normalization constraint

$$\sum_{x \in X} p(x) = 1 \tag{3.3}$$

holds. This sum is well-defined thanks to the fact that it has at most a finite number of nonzero summands, which is also the case for all other sums appearing in this example.

For every set function $f\colon X \to Y$, we can construct the corresponding function $D_R f\colon D_R X \to D_R Y$ as follows. Given $p \in D_R X$, we define $(D_R f)(p)$ to be the map

$$y \mapsto \sum_{x \in f^{-1}(y)} p(x). \tag{3.4}$$

This makes D_R into a functor. The unit of the monad has components $\eta\colon X \to D_R X$ that map each $x \in X$ to $\eta(x)\colon X \to R$ defined by

$$x' \mapsto \begin{cases} 1 & \text{if } x = x', \\ 0 & \text{if } x \neq x', \end{cases} \tag{3.5}$$

generalizing the Dirac delta distribution to the commutative semiring setting. The monad multiplication map $\mu\colon D_R D_R X \to D_R X$ is given by

$$\mu(\phi)(x) \coloneqq \sum_{p \in D_R X} \phi(p) \cdot p(x) \tag{3.6}$$

for all $\phi \in D_R D_R X$ and $x \in X$, where the product is taken in R. The monoidal unit map is uniquely determined because $D_R I \cong I$ is the terminal object. Finally, the monoidal multiplication map $\nabla\colon D_R X \times D_R Y \to D_R(X \times Y)$ is given by

$$\nabla(p, q)(x, y) \coloneqq p(x) \cdot p(y) \tag{3.7}$$

for all $p \in D_R(X)$, $q \in D_R(Y)$, $x \in X$ and $y \in Y$. The commutativity of R is relevant for showing that this lax monoidal structure is symmetric. We leave the detailed verifications to the reader.

Hence we have specified D_R as an affine symmetric monoidal monad on Set, which we call the *(generalized) distribution monad valued in R*. By Proposition 3.1, its Kleisli category is canonically a Markov category.

Returning to the general theory, we consider the relation between D and the subcategory of deterministic morphisms $\mathsf{Kl}(P)_{\mathrm{det}}$ in the Kleisli category. Clearly, the canonical identity-on-objects functor $\mathsf{D} \to \mathsf{Kl}(P)$ lands in $\mathsf{Kl}(P)_{\mathrm{det}}$. For particular monads P it often happens that this functor is fully faithful, and hence an isomorphism of categories: The original category D is precisely the category of deterministic morphisms.

For example, this happens with the Giry monad on standard Borel spaces, for which the Kleisli category is $\mathsf{BorelStoch}$ [13, Example 10.5]. On the other hand, it does *not* happen for Stoch as the Kleisli category of the Giry monad on all measurable spaces: There are $\{0,1\}$-valued probability measures on suitable measurable spaces (X, Σ_X) which are not delta measures [13, Example 10.4]. Some unfolding of the definitions shows that such a measure defines a deterministic morphism $I \to (X, \Sigma_X)$ in Stoch which does not correspond a measurable map $I \to (X, \Sigma_X)$, since the latter correspond exactly to the delta measures on X.

We now present a general criterion which guarantees that there are no such "accidental" deterministic morphisms. Intuitively, it states that the delta distributions should be precisely those distributions which are independent of themselves, or equivalently, that they should be the only product measures supported on the diagonal.

Proposition 3.4. *Let* D *be a cartesian monoidal category. Let* (P, E, δ) *be an affine symmetric monoidal monad on* D*. Then the canonical functor* $\mathsf{D} \to \mathsf{Kl}(P)_{\mathrm{det}}$ *is an isomorphism of categories if and only if the diagram*

$$\begin{tikzcd} X \arrow[r, "\delta"] \arrow[d, "{(\delta,\delta)}"'] \arrow[dr, phantom, "\lrcorner", very near start] & PX \arrow[d, "{P(\mathrm{id},\mathrm{id})}"] \\ PX \times PX \arrow[r, "\nabla"] & P(X \times X) \end{tikzcd} \tag{3.8}$$

is a pullback for every $X \in \mathsf{D}$*.*

Proof. The monoidal structure map ∇ has a left inverse given by the canonical map

$$\Delta \colon P(X \times X) \to PX \times PX$$

induced from the cartesian monoidal structure of D (note that this map corresponds to marginalization in the probability context [14]). Therefore, ∇ is a monomorphism. Since monomorphisms are stable under pullback, it follows that $\delta \colon X \to PX$ is a monomorphism as well. This implies that the canonical functor $\mathsf{D} \to \mathsf{Kl}(P)_{\mathrm{det}}$ is faithful.

To prove fullness, let $f \colon A \to PX$ be the representative of a deterministic Kleisli morphism $A \to X$ in the Markov category $\mathsf{Kl}(P)$. Some unfolding of the definitions shows that the determinism assumption amounts exactly to commutativity of the diagram

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[d, "{(f,f)}"'] & PX \arrow[d, "{P(\mathrm{id},\mathrm{id})}"] \\
PX \times PX \arrow[r, "\nabla"] & P(X \times X)
\end{tikzcd} \tag{3.9}$$

But now the assumption that diagram (3.8) is a pullback lets us obtain the dashed arrow $\tilde{f}$ in

$$\begin{tikzcd}
A \arrow[rr, bend left, "f"] \arrow[dr, dashed, "\tilde{f}"] \arrow[dr, phantom] \arrow[ddr, bend right, "{(f,f)}"'] & & \\
& X \arrow[r, "\delta"] \arrow[d, "{(\delta,\delta)}"] & PX \arrow[d, "{P(\mathrm{id},\mathrm{id})}"] \\
& PX \times PX \arrow[r, "\nabla"] & P(X \times X)
\end{tikzcd} \tag{3.10}$$

which is exactly the factorization of f needed to show that it is in the image of $\mathsf{D} \to \mathsf{Kl}(P)_{\mathrm{det}}$.

Conversely, suppose that $\mathsf{D} \to \mathsf{Kl}(P)_{\mathrm{det}}$ is an isomorphism. Our goal is now to show the unique existence of the dashed arrow in diagram (3.10). We observe that the arrow $f\colon A \to PX$ represents an arrow $A \to X$ in $\mathsf{Kl}(P)$. The commutativity of the outer square entails that this arrow is deterministic, so that there is a unique preimage $\bar{f}\colon A \to X$ in D. The condition that $\bar{f}$ is sent to f is precisely the condition that the upper triangle commutes—the lower left triangle then commutes automatically by construction of the arrows. □

Example 3.5. Consider the distribution monad D_R valued in a commutative semiring R as in Example 3.3. Then depending on what R is, the diagram (3.8) for $P = D_R$ may or may not be a pullback for all sets X. For example when $R = \mathbb{R}_+$, we recover the usual distribution monad involving finitely supported probability measures, and (3.8) is a pullback since every $\{0,1\}$-valued and finitely supported probability measure is a Dirac delta.

The most trivial examples when (3.8) is not a pullback occur when δ does not have monomorphism components. For instance, if R is the zero semiring, then the associated distribution monad D_R on Set is the terminal monad, since every $D_R X$ is a singleton set containing the unique map $X \to R$. In this case, it is clear that (3.8) is a pullback only when X itself is a singleton set.

For a less trivial example, namely one in which $\delta\colon X \to PX$ is in fact injective but (3.8) is still not a pullback, let P be the distribution monad $D_{R \oplus R}$ for any nonzero commutative semiring R, where the addition and multiplication in $R \oplus R$ are component-wise. Consider the set $X \coloneqq \{a, b\}$ and the distribution

$$s \coloneqq (0,1)\,\delta_a + (1,0)\,\delta_b \;\in\; PX \tag{3.11}$$

for $0, 1 \in R$. Clearly s is not a delta distribution, since the only two delta distributions in PX are $(1,1)\delta_a$ and $(1,1)\delta_b$. Nevertheless, both the product distribution $s \otimes s$ and $P(\mathrm{id}, \mathrm{id})(s)$ are equal to

$$(0,1)\,\delta_{(a,a)} + (1,0)\,\delta_{(b,b)} \;\in\; P(X \times X). \tag{3.12}$$

Therefore, thinking of s as a morphism $I \to PX$ in Set and using it in place of f in diagram (3.10) proves that diagram (3.8) is not a pullback in this case. Although s is not a delta measure, $s\colon I \to X$ is a deterministic morphism in $\mathsf{Kl}(D_{R\oplus R})$, correctly capturing the intuition that s does not produce any randomness.

A semiring R is *entire* if $R \not\cong 0$ and R has no zero divisors. In contrast to example 3.5, we now establish entirety as a sufficient condition for the deterministic morphisms in the Kleisli category of D_R to be precisely the ones in the image of the functor $\mathsf{Set} \to \mathsf{Kl}(D_R)$.

Proposition 3.6. *For an entire commutative semiring R, the diagram (3.8) with $P = D_R$ is a pullback for all X.*

Proof. Since δ has monomorphism components by $1 \neq 0$ in R, it is enough to prove that for every $r_1, r_2, s \in PX$ such that

$$r_1 \otimes r_2 = P(\mathrm{id}, \mathrm{id})(s) \tag{3.13}$$

holds, we necessarily have $r_1 = r_2 = s = \delta_x$ for some $x \in X$. Equation (3.13) unfolds to

$$r_1(x_1)\, r_2(x_2) = \begin{cases} s(x_1) & \text{if } x_1 = x_2 \\ 0 & \text{otherwise} \end{cases} \tag{3.14}$$

for all $x_1, x_2 \in X$. Since $\sum_{x_1} r_1(x_1)$ is equal to 1 by normalization, there must be an $\tilde{x} \in X$ such that $r_1(\tilde{x}) \neq 0$. We then necessarily have $r_2(x_2) = 0$ for all $x_2 \neq \tilde{x}$, because R is entire. Therefore, $r_2(\tilde{x}) = 1$ by normalization; and applying the same argument the other way around yields the analogous statement for r_1, so that $r_1 = r_2 = \delta_{\tilde{x}}$, from which $s = \delta_{\tilde{x}}$ follows as well. □

3.2 Markov Categories as Kleisli Categories

Many common Markov categories are indeed Kleisli categories of affine symmetric monoidal monads, as per Proposition 3.1. In this subsection, we prove a partial converse to this result. As we will see, the resulting *representable* Markov categories carry additional structure which we put to use in the rest of the paper: For every object X, there is a *distribution object* PX, to be interpreted as the space of probability measures on the given space X.

But let us start by asking under what conditions a given Markov category C arises from the construction of Proposition 3.1. If the monad P on D satisfies the assumption that (3.8) is a pullback, then Proposition 3.4 provides us with the natural bijection

$$\mathsf{Kl}(P)_{\mathrm{det}}\big(A, PX\big) \;\cong\; \mathsf{Kl}(P)\big(A, X\big), \tag{3.15}$$

intuitively stating that Markov kernels $A \to X$ are in bijection with ordinary maps $A \to PX$, where PX is the "object of distributions" on the object X.

In particular, P uniquely extends to a right adjoint to the inclusion functor $\mathsf{Kl}(P)_{\mathrm{det}} \hookrightarrow \mathsf{Kl}(P)$, resulting in a functor $\mathsf{Kl}(P) \to \mathsf{Kl}(P)_{\mathrm{det}}$ which we also denote P by abuse of notation. On a Kleisli morphism represented by $f \colon X \to PY$ in the original category D, the naturality of equation (3.15) in X shows that this functor acts by assigning to it the corresponding morphism of free P-algebras, namely the composite

$$PX \xrightarrow{Pf} PPY \xrightarrow{E} PY,$$

where E is the monad multiplication. In the probability context, the units and counits of the Kleisli adjunction (3.15) instantiate to maps intimately familiar from probability theory. The unit component $A \to PA$ is of course the maps that assigns delta distributions. The counit $PX \to X$ in $\mathsf{Kl}(P)$, which is the Kleisli morphism represented by $\mathrm{id}_{PX} \colon PX \to PX$, has been less commonly considered explicitly. It can be thought of as the Markov kernel $PX \to X$ which assigns to every probability distribution $\mu \in PX$ a random element (a "sample") of X distributed according to μ. We thus call it the *sampling map* and denote it by $\mathsf{samp} \colon PX \to X$.

In summary, if P is an affine symmetric monoidal monad satisfying the relevant pullback condition, then we obtain the natural bijection of (3.15). From right to left, a Markov kernel $A \to X$ can be reinterpreted as a deterministic map $A \to PX$; from left to right, composing a deterministic map $A \to PX$ by sampling from its output distribution produces a Markov kernel $A \to X$. By construction, we have $\mathsf{samp} \circ \delta = \mathrm{id}$, which can be interpreted to mean that sampling from a delta distribution δ_x for $x \in X$ returns x.

Based on these considerations, it is natural to consider bijections of the same type for arbitrary Markov categories now.

Definition 3.7. *Let C be a Markov category and $X \in \mathsf{C}$ an object. A* distribution object *for X is an object PX equipped with a morphism $\mathsf{samp}_X \colon PX \to X$ so that the induced map*

$$\mathsf{samp}_X \circ \_ : \ \mathsf{C}_{\mathrm{det}}(A, PX) \to \mathsf{C}(A, X) \tag{3.16}$$

is a bijection for all $A \in \mathsf{C}$.

Notation 3.8. *As before, we call samp_X the* sampling map *and often drop the subscript if no confusion is likely to arise. We write*

$$(\_)^\sharp \colon \mathsf{C}(A, X) \to \mathsf{C}_{\mathrm{det}}(A, PX) \tag{3.17}$$

for the inverse of $\mathsf{samp} \circ \_$. Using this notation, the abstract version of the delta distribution map can be identified as

$$\delta_X \coloneqq (\mathrm{id}_X)^\sharp, \tag{3.18}$$

i.e. it is the unique deterministic morphism $\delta \colon X \to PX$ satisfying

$$\mathsf{samp} \circ \delta = \mathrm{id}. \tag{3.19}$$

In other words, PX is a distribution object if it represents the hom-functor

$$\mathsf{C}(\_, X) : \ \mathsf{C}_{\mathrm{det}}^{\mathrm{op}} \to \mathsf{Set}$$

in $\mathsf{C}_{\mathrm{det}}$. The distinguished sampling morphism then arises as one represented by $\mathrm{id}_{PX} \colon PX \to PX$.

Note that the term "distribution object" is motivated by the fact that the global elements $I \to X$ in C, which are the abstract versions of probability distributions on X, correspond to the global elements $I \to PX$ in $\mathsf{C}_{\mathrm{det}}$.

Lemma 3.9. *If every $X \in \mathsf{C}$ has a distribution object PX, then the assignment $X \mapsto PX$ is the object part of a functor $P \colon \mathsf{C} \to \mathsf{C}_{\mathrm{det}}$ which is right adjoint to the inclusion $\mathsf{C}_{\mathrm{det}} \hookrightarrow \mathsf{C}$, and with the counit of the adjunction being the transformation whose components are the sampling maps.*

Proof. This is part of the standard theory of adjunctions. □

Definition 3.10. *A Markov category is termed* representable *if every object has a distribution object. We call the corresponding right adjoint functor $P \colon \mathsf{C} \to \mathsf{C}_{\mathrm{det}}$ the* distribution functor *for C.*

Let's now see some properties of representable Markov categories. First of all, for any $f\colon A \to X$ in a representable Markov category, its deterministic counterpart $f^\sharp$ from Notation 3.8 is the adjunct of f given by the composite

$$A \xrightarrow{\delta} PA \xrightarrow{Pf} PX. \tag{3.20}$$

Also, the faithfulness of the left adjoint $\mathsf{C}_{\mathrm{det}} \hookrightarrow \mathsf{C}$ also implies that the unit components $\delta\colon X \to PX$ are all monomorphisms [31, Lemma 4.5.13].

Remark 3.11. An important caveat is that δ is a natural transformation between $\mathrm{id}\colon \mathsf{C}_{\mathrm{det}} \to \mathsf{C}_{\mathrm{det}}$ and $P\colon \mathsf{C}_{\mathrm{det}} \to \mathsf{C}_{\mathrm{det}}$, and in particular natural with respect to deterministic morphisms. But δ is generally *not* natural with respect to non-deterministic morphisms. This is one way in which denoting the two functors $P\colon \mathsf{C} \to \mathsf{C}_{\mathrm{det}}$ and $P\colon \mathsf{C}_{\mathrm{det}} \to \mathsf{C}_{\mathrm{det}}$ by the same letter may be initially confusing.

On the other hand, the sampling transformation samp from $P\colon \mathsf{C} \to \mathsf{C}$ to $\mathrm{id}\colon \mathsf{C} \to \mathsf{C}$ is natural with respect to all morphisms in C. In particular, the diagram

$$\begin{tikzcd} PPX \arrow[r, "P\mathsf{samp}"] \arrow[d, "\mathsf{samp}"'] & PX \arrow[d, "\mathsf{samp}"] \\ PX \arrow[r, "\mathsf{samp}"] & X \end{tikzcd} \tag{3.21}$$

commutes for all $X \in \mathsf{C}$, which amounts to the usual associativity of the monad multiplication.

This situation, where samp is natural but δ is not, can be captured by the notion of a *thunk–force category*, which can be interpreted as "a category that looks like the Kleisli category of a monad" [17, 18].

Definition 3.12 ([17]). *A* thunk–force structure *on a category* C *amounts to*

- *an endofunctor* $L\colon \mathsf{C} \to \mathsf{C}$*;*
- *a family of maps* $\mathsf{thunk}_X\colon X \to LX$ *for each object* X*; and*
- *a family of maps* $\mathsf{force}_X\colon LX \to X$*,*

such that

- *the maps* $\mathsf{force}_X\colon LX \to X$ *assemble to a natural transformation* $L \Rightarrow \mathrm{id}$*;*
- *the maps* $\mathsf{thunk}_X\colon X \to LX$ *may not in general assemble to a natural transformation* $\mathrm{id} \Rightarrow L$*, but the maps* $\mathsf{thunk}_{LX}\colon LX \to LLX$ *do assemble to a natural transformation* $L \Rightarrow LL$*; and*

- *the following diagrams commute.*

$$\begin{tikzcd}
A \arrow[r, "\mathsf{thunk}_A"] \arrow[d, "\mathsf{thunk}_A"'] & LA \arrow[d, "L(\mathsf{thunk}_A)"] \\
LA \arrow[r, "\mathsf{thunk}_{LA}"'] & LLA
\end{tikzcd}
\qquad
\begin{tikzcd}
A \arrow[r, "\mathsf{thunk}_A"] \arrow[dr, "\mathrm{id}"'] & LA \arrow[d, "\mathsf{force}_A"] \\
& A
\end{tikzcd}
\qquad
\begin{tikzcd}
LA \arrow[r, "\mathsf{thunk}_{LA}"] \arrow[dr, "\mathrm{id}"'] & LLA \arrow[d, "L(\mathsf{force}_A)"] \\
& LA
\end{tikzcd}$$

A category equipped with a thunk–force structure is called a thunk–force category *or* abstract Kleisli category*.*

A representable Markov category is a thunk–force category, where the endofunctor L is the distribution functor $P\colon \mathsf{C} \to \mathsf{C}$, and the maps thunk and force are given by δ and samp respectively. See also [27], but keep in mind that in that paper, the name samp is used for the map thunk composed with copying. Now, as we saw in Remark 3.11, δ may not be natural against non-deterministic morphisms. In the context of thunk–force categories, this idea is captured by the notion of thunkable morphisms.

Definition 3.13 ([17])**.** *A morphism $f\colon X \to Y$ in a thunk–force category $(\mathsf{C}, L, \mathsf{thunk}, \mathsf{force})$ is called* thunkable *if and only if the following diagram commutes.*

$$\begin{tikzcd}
X \arrow[r, "f"] \arrow[d, "\mathsf{thunk}_X"'] & Y \arrow[d, "\mathsf{thunk}_Y"] \\
LX \arrow[r, "Lf"'] & LY
\end{tikzcd} \tag{3.22}$$

It turns out that, for representable Markov categories, this class of morphisms coincides with that of deterministic morphisms.

Proposition 3.14. *A morphism $f\colon X \to Y$ in a representable Markov category is deterministic if and only if it is thunkable, i.e. if and only if we have*

$$\delta_Y \circ f = Pf \circ \delta_X. \tag{3.23}$$

See also [27, Theorem 3.14] for a more general context.

Proof. The “only if” direction was already noted in Remark 3.11. For the “if” part, we now prove that the top face of the following cube commutes.

$$\begin{tikzcd}
X \arrow[rr, "f"] \arrow[dr, "copy"'] \arrow[dd, "\delta"'] & & Y \arrow[dr, "copy"] \arrow[dd, "\delta", crossing over] & \\
& X \otimes X \arrow[rr, "f \otimes f"] \arrow[dd, "\delta \otimes \delta"] & & Y \otimes Y \arrow[dd, "\delta \otimes \delta"] \\
PX \arrow[rr, "Pf"', crossing over] \arrow[dr, "copy"'] & & PY \arrow[dr, "copy"] & \\
& PX \otimes PX \arrow[rr, "Pf \otimes Pf"'] & & PY \otimes PY
\end{tikzcd} \tag{3.24}$$

Now,

- The front and back faces commute by the assumed naturality equation (3.23);
- The two side faces commute since δ is deterministic;
- The bottom face commutes since Pf is deterministic.

Therefore, the top face commutes after postcomposing with the front right leg $\delta \otimes \delta$. By equation (3.19), i.e. $\mathsf{samp} \circ \delta = \mathrm{id}$, we conclude that the top face of the cube also commutes as such. □

Now, if $\mathsf{C} = \mathsf{Kl}(P)$ is a Markov category arising from the construction of Proposition 3.1 and the monad P satisfies the pullback condition of Equation (3.8), then C is representable.

Somewhat conversely, if C is a representable Markov category, then the defining adjunction induces a monad on $\mathsf{C}_{\mathrm{det}}$. We denote its underlying functor also by $P \colon \mathsf{C}_{\mathrm{det}} \to \mathsf{C}_{\mathrm{det}}$, since it differs from $P \colon \mathsf{C} \to \mathsf{C}_{\mathrm{det}}$ from Lemma 3.9 merely by restriction to the subcategory $\mathsf{C}_{\mathrm{det}}$. This monad has unit δ and multiplication $P\mathsf{samp}$. Indeed, in the probability context, sampling from the "inner" distribution of a distribution of distributions returns the expected distribution, which is consistent with the idea that $P\mathsf{samp}$ is the multiplication of a probability monad. In fact, we can also compose $P \colon \mathsf{C} \to \mathsf{C}_{\mathrm{det}}$ with the inclusion functor on the other side, considering P as a functor $\mathsf{C} \to \mathsf{C}$ instead. Hence, P comes in three versions which we do not distinguish notationally; we leave it understood that P can act on any morphism of C and always returns a deterministic morphism.

For every representable Markov category C with distribution functor P, there is a canonical isomorphism $\mathsf{C} \cong \mathsf{Kl}(P)$. This is an instance of the elementary fact that if any identity-on-objects functor $\mathsf{D}_1 \to \mathsf{D}_2$ has a right adjoint, then this makes D_2 canonically isomorphic to the Kleisli category of the induced monad on D_1.[6]

However, the Markov category structure on C equips this monad with additional structure and properties. Next, we show that P is an affine symmetric monoidal monad in a canonical way and that it automatically satisfies the pullback condition of Proposition 3.4. As a consequence, if the right adjoint of $\mathsf{C}_{\mathrm{det}} \hookrightarrow \mathsf{C}$ exists, then the canonical isomorphism of categories $\mathsf{C} \cong \mathsf{Kl}(P)$ is in fact an isomorphism of Markov categories.

Proposition 3.15. *Let C be a representable Markov category. Then the right adjoint $P\colon \mathsf{C} \to \mathsf{C}_{\mathrm{det}}$ has a canonical symmetric lax monoidal structure which makes the adjunction between P and the inclusion functor $\iota\colon \mathsf{C}_{\mathrm{det}} \hookrightarrow \mathsf{C}$ into a symmetric monoidal adjunction.*

The proof is best understood as an instance of the general theory of *doctrinal adjunctions* [24].

Proof. Since both composites and monoidal products of deterministic morphisms are again deterministic, and also all monoidal structure isomorphisms are deterministic, we can equip $\mathsf{C}_{\mathrm{det}}$ with the monoidal structure induced from C, and this makes the inclusion functor $\iota\colon \mathsf{C}_{\mathrm{det}} \hookrightarrow \mathsf{C}$ into a strict symmetric monoidal functor by definition.

By the general theory of doctrinal adjunctions,[7] a right adjoint to a strong monoidal functor is canonically lax monoidal, and the structure maps are given as follows:

- For all objects X and Y of C, the multiplication map ∇ of the functor $P\colon \mathsf{C} \to \mathsf{C}_{\mathrm{det}}$ is given by

$$\nabla \,:\; PX \otimes PY \xrightarrow{\delta} P(PX \otimes PY) \xrightarrow{P(\mathsf{samp} \otimes \mathsf{samp})} P(X \otimes Y),$$

  which is deterministic due to being a composite of deterministic morphisms. Naturality of ∇ means that the following diagram ought to commute for all (not necessarily deterministic) morphisms $f\colon X \to Y$ and $g\colon A \to B$:

$$\begin{tikzcd}
PX \otimes PA \arrow[r, "\delta"] \arrow[d, "Pf \otimes Pg"] & P(PX \otimes PA) \arrow[r, "P(\mathsf{samp} \otimes \mathsf{samp})"] \arrow[d, "P(Pf \otimes Pg)"] & P(X \otimes A) \arrow[d, "P(f \otimes g)"] \\
PY \otimes PB \arrow[r, "\delta"] & P(PY \otimes PB) \arrow[r, "P(\mathsf{samp} \otimes \mathsf{samp})"] & P(Y \otimes B)
\end{tikzcd}$$

[6] We thank Sam Staton for pointing this fact out to us.

[7] While the paper [24] is not open access, the result we are using appears as Proposition 2.1 on the nLab page ncatlab.org/nlab/show/monoidal+adjunction.

The left square commutes by naturality of δ with respect to the deterministic morphism $Pf \otimes Pg$ and the right one by naturality of samp, so that ∇ is indeed natural in both arguments. This can be interpreted as the fact that processing two independent random variables independently preserves their independence.

A straightforward but tedious diagrammatic argument, involving the given properties of δ and samp including $\mathsf{samp} \circ \delta = \mathrm{id}$, then shows that the relevant associativity condition for ∇ to be a lax monoidal structure holds as well. Compatibility with the braiding $X \otimes Y \to Y \otimes X$ is obvious.

- The natural isomorphism

$$\mathsf{C}_{\mathrm{det}}\big(X, PI\big) \cong \mathsf{C}(X, I)$$

shows that $PI \cong I$ by the assumed terminality of I. The unit $I \to PI$ is thus the unique morphism of this type, and it automatically satisfies the relevant compatibility conditions with the multiplication.

Hence, the right adjoint $P\colon \mathsf{C} \to \mathsf{C}_{\mathrm{det}}$ is a symmetric lax monoidal functor. It remains to be shown that δ and samp, as unit and counit of the adjunction, are monoidal transformations.

The fact that δ is a monoidal natural transformation means that the following diagram

$$\begin{tikzcd}
X \otimes Y \arrow[r, "\delta\otimes\delta"] \arrow[dr, "\delta"'] & PX \otimes PY \arrow[d, "\nabla"] \\
& P(X \otimes Y)
\end{tikzcd}$$

commutes. This can be interpreted as the fact that products of Dirac deltas are again Dirac deltas. A formal proof follows via a standard naturality argument together with $\mathsf{samp} \circ \delta = \mathrm{id}$.

Dually, the fact that samp is a monoidal natural transformation means that the diagram

$$\begin{tikzcd}
PX \otimes PY \arrow[r, "\nabla"] \arrow[dr, "\mathsf{samp}\otimes\mathsf{samp}"'] & P(X \otimes Y) \arrow[d, "\mathsf{samp}"] \\
& X \otimes Y
\end{tikzcd}$$

commutes. This can be interpreted as the fact that sampling from a product distribution is the same as sampling from the two marginals independently, and again follows formally by similar arguments. □

Remark 3.16. The *strength* of the monoidal monad P is given by the deterministic maps $\sigma_{X,Y} : X \otimes PY \to P(X \otimes Y)$, natural in C_{det}, given by the composition[8]

$$X \otimes PY \xrightarrow{\delta \otimes \mathrm{id}} PX \otimes PY \xrightarrow{\nabla} P(X \otimes Y).$$

The strength satisfies the following commutative diagram,

$$\begin{tikzcd}
X \otimes PY \arrow[r, "\sigma"] \arrow[dr, "\mathrm{id} \otimes \mathsf{samp}"'] & P(X \otimes Y) \arrow[d, "\mathsf{samp}"] \\
& X \otimes Y
\end{tikzcd} \tag{3.25}$$

which has a similar, but "one-sided", interpretation to the analogous condition for ∇. Note that, since the unit δ of the adjunction is not natural on the whole of C (Remark 3.11), the strength $\sigma \colon X \otimes PY \to P(X \otimes Y)$ is natural with respect to general morphisms only in the second argument, and natural with respect to deterministic morphisms in the first argument.

Corollary 3.17. *Let C be a representable Markov category. Then the monad $(P, P\mathsf{samp}, \delta)$ on C_{det} arising from the underlying adjunction is symmetric monoidal and affine, thus inducing an isomorphism of Markov categories $\mathsf{C} \cong \mathsf{Kl}(P)$.*

Proof. We have already noted that there is a canonical isomorphism of categories $\mathsf{C} \cong \mathsf{Kl}(P)$. It is also an isomorphism of *monoidal* categories because the defining adjunction $\mathsf{C}_{\text{det}}(A, PX) \cong \mathsf{C}(A, X)$ is monoidal. Finally, to see that the copy maps are preserved, it is enough to note that on both sides, they are given by the diagonals $Y \to Y \times Y$ in the cartesian monoidal category C_{det}. □

Lemma 3.18. *Let C be a representable Markov category with distribution functor P. Then P satisfies the pullback condition of Proposition 3.4 on C_{det}.*

Proof. We need to show that for every $A, X \in \mathsf{C}$ and any diagram in C_{det} of the form

$$\begin{tikzcd}
A \arrow[drr, bend left, "f_1"] \arrow[dr, dashed, "g"] \arrow[ddr, bend right, "f_2"'] & & \\
& X \arrow[r, "\delta"] \arrow[d, "{(\delta,\delta)}"'] \arrow[dr, phantom, "\lrcorner", very near start] & PX \arrow[d, "{P(\mathrm{id},\mathrm{id})}"] \\
& PX \otimes PX \arrow[r, "\nabla"] & P(X \otimes X)
\end{tikzcd} \tag{3.26}$$

[8] One can equivalently start from a commutative strength and construct the monoidal structure in terms of it, see [6, Section 6.3].

without the dashed arrow, there is a unique dashed arrow such that the diagram commutes. Note that the diagonal in $\mathsf{C}_{\mathrm{det}}$ is given by the copy map in C, so that the two vertical morphisms in the diagram are

$$(\delta, \delta) = (\delta \otimes \delta) \circ \mathrm{copy}_X, \qquad\qquad P(\mathrm{id}, \mathrm{id}) = P(\mathrm{copy}_X). \tag{3.27}$$

Since δ is a monomorphism by $\mathsf{samp} \circ \delta = \mathrm{id}$, the uniqueness is automatic and it is enough to find some g which makes the diagram commute.

To this end, note first that composing the whole diagram with the two marginalization maps $P(X \otimes X) \to PX$ shows that $f_2 = (f_1, f_1)$, again as a pairing with respect to the universal property of $PX \otimes PX$ as a product in $\mathsf{C}_{\mathrm{det}}$.

We now show that $g \coloneqq \mathsf{samp} \circ f_1$ does the job. To this end, it is enough to prove that g is deterministic, because then we have

$$\delta \circ g = Pg \circ \delta \tag{3.28}$$

resulting in commutativity of the upper triangle by

$$Pg \circ \delta = P\mathsf{samp} \circ Pf_1 \circ \delta = f_1 \circ \mathsf{samp} \circ \delta = f_1 \tag{3.29}$$

Commutativity of the lower left triangle then also follows, thanks to $f_2 = (f_1, f_1)$.

The claim that g is deterministic amounts to the commutativity of the outermost rectangle in the diagram

$$\begin{tikzcd}
A \arrow[r, "f_1"] \arrow[dd, "\mathrm{copy}"'] \arrow[ddr, "f_2"] & PX \arrow[rr, "\mathsf{samp}"] \arrow[dr, "P(\mathrm{copy})"'] & & X \arrow[dd, "\mathrm{copy}"] \\
& & P(X \otimes X) \arrow[dr, "\mathsf{samp}"] & \\
A \otimes A \arrow[r, "f_1 \otimes f_1"'] & PX \otimes PX \arrow[ur, "\nabla"] \arrow[rr, "\mathsf{samp} \otimes \mathsf{samp}"'] & & X \otimes X
\end{tikzcd}$$

Here, the lower left triangle commutes by $f_2 = (f_1, f_1)$, the oddly shaped square by assumption, the upper right square by naturality of samp and the lower right triangle by Proposition 3.15. Therefore, choosing g to be $\mathsf{samp} \circ f_1$ makes the diagram (3.26) commute. □

We can summarize the previous results as follows.

Theorem 3.19. *For a Markov category* C*, the following are equivalent:*

1. C *is representable.*
2. *There is an affine symmetric monoidal monad* P *on* $\mathsf{C}_{\mathrm{det}}$ *such that:*
   - *The diagram* (3.8) *is a pullback for every* X*.*
   - *The identity functor on* $\mathsf{C}_{\mathrm{det}}$ *extends to an isomorphism of Markov categories* $\mathsf{C} \cong \mathsf{Kl}(P)$*.*

In particular, since representability is a property rather than extra structure, the monoidal monad P in the second condition is unique (up to unique isomorphism).

This result is similar, but unrelated, to [5, Theorem 4.7], where a correspondence is drawn between Freyd categories and Kleisli categories of *strong* (not necessarily commutative) monads.

Proof. If C is representable, Lemma 3.9 gives us the desired monad P, and the isomorphism of Markov categories is the one of Corollary 3.17. Finally, Lemma 3.18 states exactly that this monad satisfies the pullback condition.

For the converse, we only need to show that the inclusion functor $\mathsf{C}_{\mathrm{det}} \hookrightarrow \mathsf{C}$ has a right adjoint. This holds by the assumed isomorphism $\mathsf{C} \cong \mathsf{Kl}(P)$ together with the Kleisli adjunction, which gives us natural bijections

$$\mathsf{C}_{\mathrm{det}}(A, PX) \cong \mathsf{Kl}(P)(A, X) \cong \mathsf{C}(A, X).$$

Note that the pullback condition is not needed in this argument. □

Example 3.20. If C is a representable Markov category, then every parametric Markov category C_W as introduced in Section 2.2 is representable too. One can use the same distribution objects and take the sampling map samp_X in C_W to be represented by $\mathrm{del}_W \otimes \mathsf{samp}_X$, resulting in the desired bijection

$$\mathsf{C}_{W,\mathrm{det}}(A, P_W X) = \mathsf{C}_{\mathrm{det}}(W \otimes A, PX) \cong \mathsf{C}(W \otimes A, X) = \mathsf{C}_W(A, X).$$

Thus, the distribution functor $P_W \colon \mathsf{C}_W \to \mathsf{C}_{W,\mathrm{det}}$ acts the same on objects as the original $P \colon \mathsf{C} \to \mathsf{C}_{\mathrm{det}}$ does. The action on morphisms is then uniquely determined subject to making the bijection $\mathsf{C}_{W,\mathrm{det}}(A, P_W X) \cong \mathsf{C}_W(A, X)$ natural. Concretely, a morphism $f \in \mathsf{C}_W(A, X)$ represented by $f \colon W \otimes A \to X$ gets mapped to the morphism $P_W f \in \mathsf{C}_W(P_W A, P_W X)$ represented by the composite

$$W \otimes PA \xrightarrow{\sigma} P(W \otimes A) \xrightarrow{Pf} PX, \tag{3.30}$$

where $\sigma\colon W \otimes PA \to P(W \otimes A)$ denotes the strength of the monad P as introduced in Remark 3.16.

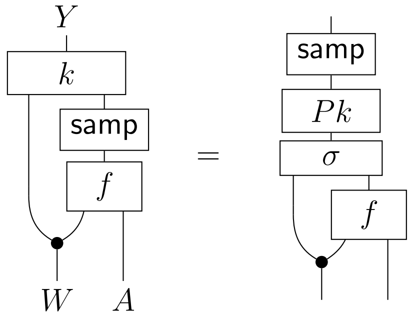
In order to see that this is how P_W must act on morphism in C_W, it is enough to show that this prescription indeed makes the bijection

$$\mathsf{C}_{W,\mathrm{det}}(A, P_W X) \cong \mathsf{C}_W(A, X)$$

natural in X. That is, we need to check that for each morphism $k\colon X \to Y$ represented by $k\colon W \otimes X \to Y$, the following diagram

$$\begin{tikzcd}
\mathsf{C}_{W,\mathrm{det}}(A, P_W X) \arrow[r, "\mathsf{samp}\circ\_", "\cong"'] \arrow[d, "P_W k\circ\_"] & \mathsf{C}_W(A, X) \arrow[d, "k\circ\_"] \\
\mathsf{C}_{W,\mathrm{det}}(A, P_W Y) \arrow[r, "\mathsf{samp}\circ\_", "\cong"'] & \mathsf{C}_W(A, Y)
\end{tikzcd}$$

commutes. Starting with a deterministic $f\colon W \otimes A \to PX$ in the top-left corner, commutativity of the diagram amounts to showing that the equation



(3.31)

holds in C, where we use (3.30) in order to express $P_W k$ in terms of Pk and σ. This equation follows straightforwardly if we apply the naturality of samp and property (3.25).

3.3 Almost-Surely-Compatible Representability

In a representable Markov category, it is not a priori clear whether the defining equation $\mathsf{C}_{\mathrm{det}}(A, PX) \cong \mathsf{C}(A, X)$ respects almost sure equality, in the following sense. An almost sure equality of two deterministic morphisms $A \to PX$ (with respect to some morphism $p\colon \Theta \to A$) implies the corresponding almost sure equality of the resulting morphisms $A \to X$, since the latter are obtained simply by composition with the sampling morphisms $PX \to X$. However, the other direction is not clear:

If two morphisms $A \to X$ are almost surely equal, does this means that also their deterministic counterparts $A \to PX$ must be almost surely equal?

Indeed, in Example 3.26 we provide a representable Markov category in which this converse implication fails to hold. But since such a converse is relevant to our upcoming applications of representable Markov categories, we now investigate representable Markov categories in which it does hold.

Definition 3.21. *A Markov category is* a.s.-compatibly representable *if it is representable and for any morphism $p\colon \Theta \to A$, the defining natural bijection*

$$\mathsf{C}_{\mathrm{det}}(A, PX) \cong \mathsf{C}(A, X)$$

respects almost sure equality. That is, for all $f, g\colon A \to X$, we have

$$f^\sharp =_{p\text{-}a.s.} g^\sharp \qquad \Longleftrightarrow \qquad f =_{p\text{-}a.s.} g. \tag{3.32}$$

As we already noted, the implication from left to right is automatic because of $f = \mathsf{samp}\, f^\sharp$.

Many representable Markov categories are actually a.s.-compatibly representable, including $\mathsf{BorelStoch}$ as the following example shows.

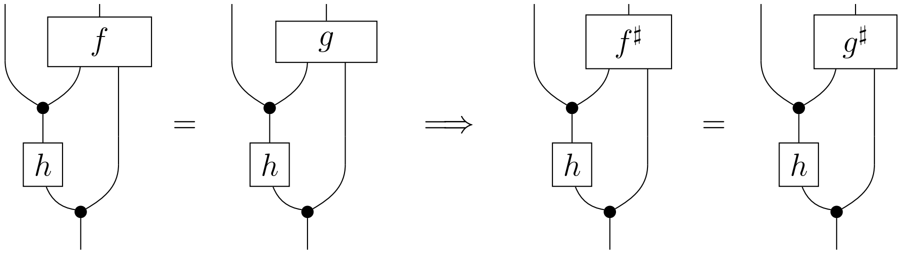
Example 3.22. For any two $f, g\colon A \to X$ in $\mathsf{BorelStoch}$, we have $f =_{p\text{-a.s.}} g$ if and only if for all $S \subseteq \Sigma_A$ and $T \subseteq \Sigma_X$,

$$\int_S f(T|a)\, p(da|\theta) = \int_S g(T|a)\, p(da|\theta), \tag{3.33}$$

or equivalently if and only if the two functions $f(T|\_), g(T|\_)\colon A \to \mathbb{R}$ are $p(\_|\theta)$-a.s. equal for every $\theta \in \Theta$ and every $T \in \Sigma_X$ [13, Example 13.3]. What we need to prove is that this holds uniformly in T, i.e. that the *measures* $f(\_|a)$ and $g(\_|a)$ are likewise $p(\_|\theta)$-almost surely equal with respect to $a \in A$. Since Σ_X is countably generated, say by a sequence of measurable sets $(T_n)_{n\in\mathbb{N}}$, it is enough to show that $f(T_n|a) = g(T_n|a)$ holds for all n with unit probability in a. But this is indeed the case by assumption, since a countable intersection of sets of full measure again has full measure. Therefore, $\mathsf{BorelStoch}$ is a.s.-compatibly representable.

A property equivalent to a.s.-compatible representability, which is useful in manipulations of string diagrams, turns out to be the following.

Definition 3.23. *A representable Markov category is said to satisfy the* sampling cancellation property *if, for any three morphisms $f, g \colon X \otimes A \to Y$ and $h \colon A \to X$, the following implication holds:*

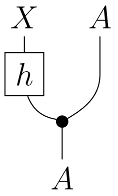


The name of this condition is explained by the equation $f = \mathsf{samp}\, f^\sharp$, so that the implication amounts to the possibility to cancel the sampling map in diagram equations of the above form.

Proposition 3.24. *A representable Markov category C satisfies the sampling cancellation property if and only if it is a.s.-compatibly representable.*

Proof. If in Definition 3.23 we use $f, g \colon X \otimes A \to Y$ of the form $f \otimes \mathrm{del}_A$ and $g \otimes \mathrm{del}_A$ for arbitrary $f, g \colon X \to Y$ and $h = p$, then we recover the non-trivial direction of (3.32).

Conversely, if in Definition 3.21 we use $p \colon A \to X \otimes A$ given by



(3.34)

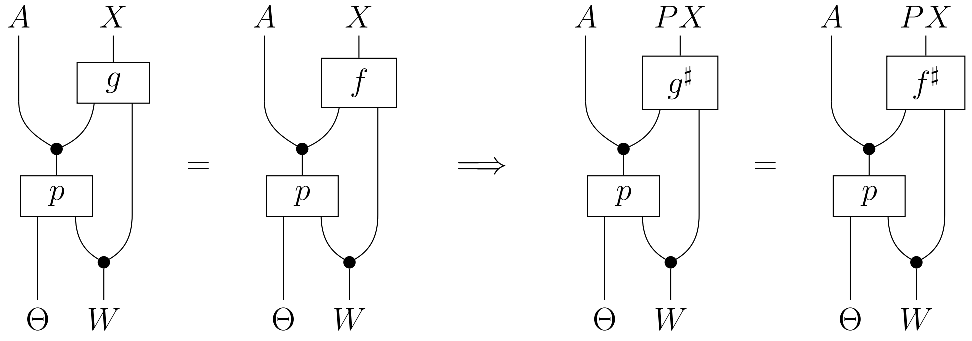
for a given $h \colon A \to X$, then the right-to-left implication of (3.32) implies the sampling cancellation property. □

In Example 3.20, we saw that if a Markov category C is representable, then so is every parametric Markov category C_W. As the following lemma shows, the same can be said about a.s.-compatible representability.

Lemma 3.25. *Let C be an a.s.-compatibly representable Markov category. For every $W \in \mathsf{C}$, the parametric Markov category C_W as introduced in Section 2.2 is likewise a.s.-compatibly representable.*

Proof. First of all, notice that if $f \in \mathsf{C}_W(A, X)$ is a morphism in C_W represented by $f \colon W \otimes A \to X$ in C, then its adjunct $f^\sharp \in \mathsf{C}_W(A, P_W X)$ is represented by $f^\sharp \colon W \otimes A \to PX$, the adjunct of f in C. This follows because $P_W X$ is a distribution object of C_W with respect to the same sampling map as PX is in C.

Therefore, the non-trivial part of checking a.s.-compatible representability of C_W, i.e. the right-to-left implication of (3.32), boils down to the following implication



(3.35)

in C. This holds because C satisfies the sampling cancellation property by Proposition 3.24. Consequently, C_W is a.s.-compatibly representable. □

Example 3.26. We now give an example of a Markov category which is representable but not a.s.-compatibly representable. Continuing on from Proposition 3.6, the Kleisli category $\mathsf{Kl}(D_R)$ of the distribution monad D_R for an entire commutative semiring R is representable with $\mathsf{Kl}(D_R)_{\mathrm{det}} = \mathsf{Set}$. As we elaborate below, there is an R such that $\mathsf{Kl}(D_R)$ is not a.s.-compatibly representable.

Concretely, let R be the semiring

$$R \coloneqq \{0, \varepsilon, 1\} \tag{3.36}$$

with addition and multiplication given by the following nontrivial cases:

$$\begin{gathered} 1 + 1 = 1, \quad 1 + \varepsilon = 1, \quad \varepsilon + \varepsilon = \varepsilon, \\ \varepsilon^2 = \varepsilon. \end{gathered} \tag{3.37}$$

Intuitively, we can think of assigning value $1 \in R$ to outcomes that are possible and happen with nonzero probability; assigning ε to outcomes that may be considered, but whose probability is 0 or negligibly small; and assigning $0 \in R$ to outcomes that could never happen. Then the above arithmetical rules acquire straightforward interpretations.

Since this commutative semiring R has idempotent addition and multiplication, it can also be understood as a distributive lattice with max as addition and min as multiplication. In this picture, R is simply the three-element totally ordered set with $0 < \varepsilon < 1$. It is also clear that R is entire.

In the representable Markov category $\mathsf{Kl}(D_R)$, we then consider $A = X = \{a, b\}$ together with $f, g \colon A \to A$ represented by the morphisms $f^\sharp, g^\sharp \colon A \to PA$ given by

$$\begin{aligned} f^\sharp(a) &\coloneqq \delta_a, & f^\sharp(b) &\coloneqq \varepsilon\delta_a + \ \delta_b, \\ g^\sharp(a) &\coloneqq \delta_a, & g^\sharp(b) &\coloneqq \ \delta_a + \varepsilon\delta_b. \end{aligned}$$

We also consider $p \colon I \to A$ represented by $p^\sharp \coloneqq \delta_a + \varepsilon\delta_b$. Then we have $f^\sharp \neq_{p\text{-a.s.}} g^\sharp$, since

$$\delta_{(a,\delta_a)} + \varepsilon\delta_{(b,\varepsilon\delta_a+\delta_b)} \neq \delta_{(a,\delta_a)} + \varepsilon\delta_{(b,\delta_a+\varepsilon\delta_b)}. \tag{3.38}$$

However, applying sampling to the second output reduces this to

$$\delta_{(a,a)} + \varepsilon^2\delta_{(b,a)} + \varepsilon\delta_{(b,b)} = \delta_{(a,a)} + \varepsilon\delta_{(b,a)} + \varepsilon^2\delta_{(b,b)}, \tag{3.39}$$

where the equation holds by $\varepsilon^2 = \varepsilon$. Therefore, we have $f =_{p\text{-a.s.}} g$ and $\mathsf{Kl}(D_R)$ is not a.s.-compatibly representable.

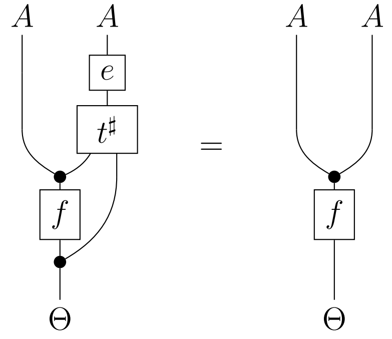
4 Second-Order Stochastic Dominance

In this and the following sections, we state and prove generalizations of existing concepts in probability theory and statistics to representable Markov categories. The first one for which we do this is *second-order stochastic dominance.* Traditionally, this is a partial order of probability distributions on $\mathbb{R}$ or $\mathbb{R}^n$ that expresses whether a given distribution is more "spread out" than another.

In the abstract setting of representable Markov categories, second-order dominance makes sense for all algebras A of the monad P on C_{det}. It is concerned with comparing two distributions on A, represented now by morphisms $p, q \colon I \to A$, and induces a preorder relation on $\mathsf{C}(I, A)$. More generally, it is a preorder relation on *every* hom-set $\mathsf{C}(\Theta, A)$, defined in terms of the P-algebra structure on A.

One way to define second-order dominance in terms of the existence of a so-called *dilation* $A \to A$ that takes q to p. Intuitively, a dilation is a specific kind of map which increases uncertainty without affecting the expected value of distributions on A.

Definition 4.1. *Let C be a representable Markov category and let $e \colon PA \to A$ be an algebra of the monad $(P, \delta, P\mathsf{samp})$ on C_{det}. Then, given a morphism $f \colon \Theta \to A$ in C, a morphism $t \colon \Theta \otimes A \to A$ is an* f-dilation *(with respect to the P-algebra* *structure e) if it satisfies*



(4.1)

Thus, in case Θ is the unit object, an f-dilation t has to satisfy

$$e \circ t^\sharp =_{f\text{-a.s.}} \mathrm{id}. \tag{4.2}$$

In the probability theory context, such dilations are also called *mean-preserving maps*, as the following example elucidates.

Example 4.2. Given the choice of $\mathsf{C} = \mathsf{BorelStoch}$ and a compact set $A \subseteq \mathbb{R}$, the canonical algebra map $e\colon PA \to A$ is the assignment of expectation values. Relative to it, an f-dilation is any Markov kernel $t\colon A \to PA$ that takes f-almost every point $a \in A$ to a distribution $t^\sharp(a)$ whose expectation value is a itself. Similarly, for a compact $A \subseteq \mathbb{R}^n$, or for a closed and bounded subset of a separable Banach space, the natural choice of $e\colon PA \to A$ is one that maps each distribution to its barycenter.

For general Θ, equation (4.1) in $\mathsf{BorelStoch}$ becomes

$$e\big(t(\_|\theta, a)\big) = a \tag{4.3}$$

for $f(\_|\theta)$-almost all $a \in A$ and all $\theta \in \Theta$. This again matches the intuition that the distribution $t(\_|\theta, a)$ dilates the point a, suitably almost surely, but now for every $\theta \in \Theta$ separately.

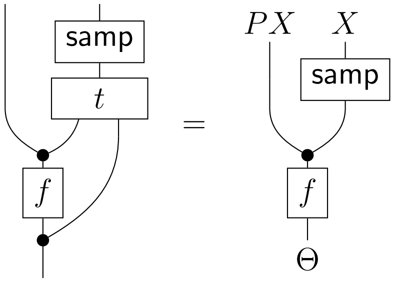
Remark 4.3. Throughout this section, we could restrict to the case $\Theta = I$ only, and then instantiate our definitions and results on the parametric Markov category C_Θ in order to recover the additional parameter dependence on any $\Theta \in \mathsf{C}$. We have decided against doing so for the benefit of making the greater generality of our formalism more explicitly apparent.

When working with second-order dominance, it is often helpful to assume a.s.-compatible representability rather than mere representability, which we do from now on.

Lemma 4.4. *Let C be an a.s.-compatibly representable Markov category. Consider the free algebra PX of the monad $(P, P\mathsf{samp}, \delta)$ on $\mathsf{C}_{\mathrm{det}}$, with algebra map*

$$P\mathsf{samp} \colon PPX \to PX,$$

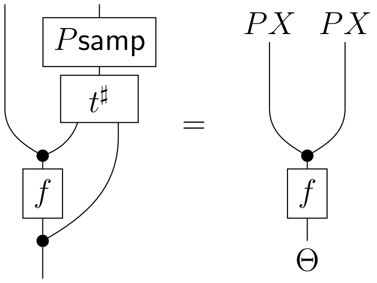
and a morphism $f \colon \Theta \to PX$ in C. Then $t \colon \Theta \otimes PX \to PX$ is an f-dilation if and only if it satisfies



(4.4)

Proof. Equation (4.4) follows from (4.1) for $e = P\mathsf{samp}$, namely by applying $\mathrm{id}_{PX} \otimes \mathsf{samp}_X$ to it and using the commutativity of diagram (3.21) together with $\mathsf{samp} \circ t^\sharp = t$.

Conversely, since the sampling cancellation property follows from a.s.-compatible representability by Proposition 3.24, we can follow the same reasoning backwards[9] and conclude the equation



(4.5)

as was to be shown. □

As the following proposition shows, the existence of dilations that convert between morphisms can be related to the existence of partial evaluations [15].

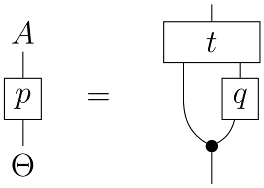
Proposition 4.5. *Let C be an a.s.-compatibly representable Markov category with distribution functor P and for which conditionals exist. Then for any P-algebra $e \colon PA \to A$ in $\mathsf{C}_{\mathrm{det}}$ and any morphisms $p, q \colon \Theta \to A$ in C, the following two conditions are equivalent:*

[9] Note that we cannot merely remove the sampling maps since t is generally not deterministic.

(i) There exists a morphism $r\colon \Theta \to PA$ such that the following diagram commutes:

$$\begin{tikzcd}
\Theta \arrow[dr, dashed, "r"] \arrow[drr, bend left, "p"] \arrow[ddr, bend right, "q"'] & & \\
& PA \arrow[r, "\mathsf{samp}"'] \arrow[d, "e"] & A \\
& A &
\end{tikzcd} \tag{4.6}$$

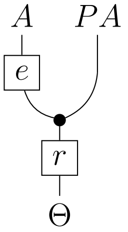
(ii) There exists a q-dilation $t\colon \Theta \otimes A \to A$ which converts q to p in the following sense:



(4.7)

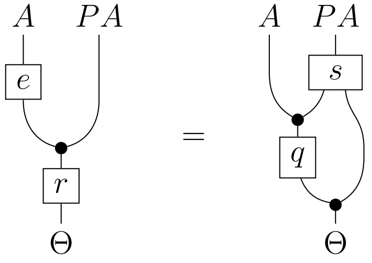
Proof. We split the proof into two parts each consisting of one of the implications.

(i) ⇒ (ii): Given an $r\colon \Theta \to PA$ that makes the diagram (4.6) commute, we construct $s\colon A \otimes \Theta \to A$ as a conditional of the morphism



(4.8)

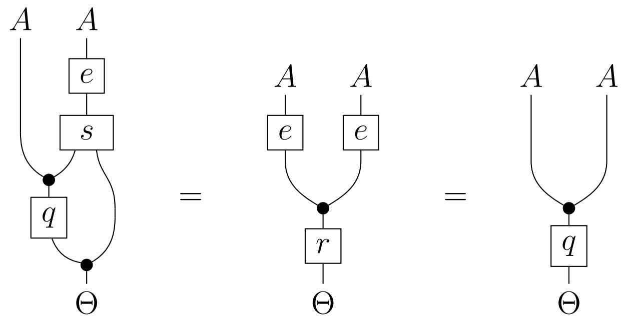
with respect to the left output, so that the defining equation of conditionals



(4.9)

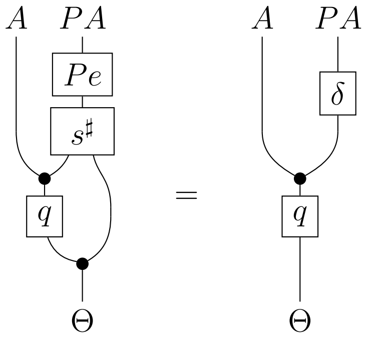
holds. Upon defining $t \coloneqq \mathsf{samp}_A \circ s$, the property (4.7) then follows immediately by applying $\mathrm{del}_A \otimes \mathrm{id}_{PA}$ to equation (4.9).

It remains to be shown that t is a q-dilation. Applying $\mathrm{id}_A \otimes e$ to equation (4.9), we get



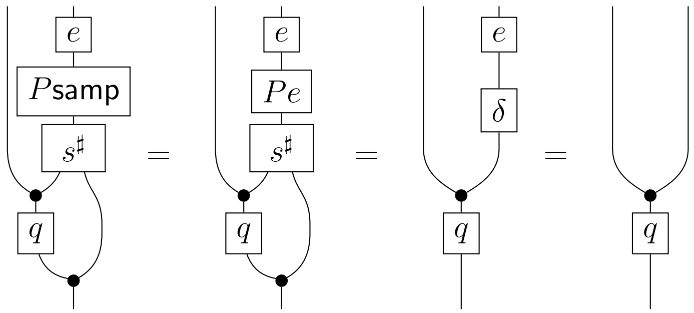
(4.10)

where the second step uses the assumption that e is deterministic. Applying the sampling cancellation property, we obtain



(4.11)

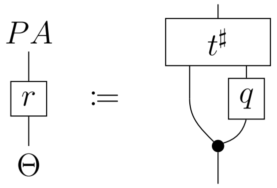
Now the fact that $t^\sharp = P\mathsf{samp} \circ s^\sharp$ satisfies equation (4.1) follows by applying $\mathrm{id}_A \otimes e$ to equation (4.11) and using the fact that e is a P-algebra:



(4.12)

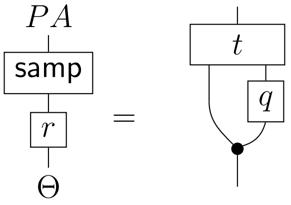
Thus, we have shown the implication (i) ⇒ (ii).

(ii) ⇒ (i): Given a morphism $t\colon \Theta \otimes A \to A$ satisfying (4.7) and (4.1), we define r via



(4.13)

Applying samp_A to this equation yields



(4.14)

whence $\mathsf{samp}_A r = p$ follows by virtue of the fact that t converts q to p.

On the other hand, applying e to equation (4.13) gives $e \circ r = q$ by the assumption that t is a q-dilation with respect to e. □

Definition 4.6. *If the equivalent conditions of Proposition 4.5 hold, then we say that* q second-order dominates p *with respect to* e*, and denote this relation as*

$$p \sqsubseteq q. \tag{4.15}$$

By Example 4.2, this recovers the usual notion of second-order stochastic dominance for probability measures on any compact $A \subseteq \mathbb{R}^n$ if we take C to be $\mathsf{BorelStoch}$ and $\Theta = I$.

5 Comparison of Statistical Experiments

5.1 Informativeness of Statistical Experiments

Dilations also appear in the context of comparison of statistical experiments, as introduced by Blackwell [2]. In this case, statistical experiments are modeled by families of measures indexed by a parameter set Θ that labels the distinct hypotheses one aims to discern by performing the experiment. Here, we represent statistical experiments abstractly as morphisms $\Theta \to X$ in a Markov category.

Commonly, the question of comparing two experiments f, g in terms of whether f is *more informative* about Θ than g is, amounts to the existence of a map c satisfying

$c\,f = g$. We can interpret c as a post-processing of the data generated by experiment f in a way that produces the data of experiment g. If such a c exists, we refer to it as *garbling map* and say that f is more informative than g, denoted by $f \succeq g$. A slightly more general version of this notion was introduced by Golubtsov in [20].

Example 5.1. For a concrete example, consider a company that manufactures metal rods. After production, a selection of rods is tested for structural integrity. The hypothesis space $\Theta = \{\mathsf{s}, \mathsf{f}\}$ thus consists of two possibilities: Either the selected rod is safe (s) or it is faulty (f).

There are two tests available. The threshold test g checks whether the rod withstands hanging a weight from its midpoint, while the oscillation test f checks whether the rod withstands transverse oscillations of a certain type. Each test has two possible outcomes: Either the rod passes ($\checkmark$) or fails ($\times$) the test, so that we have $X = Y = \{\checkmark, \times\}$. According to our model of metal rods and their defects (that we assume to be correct, up to the experimental precision) the outcomes of the two tests are generally distributed as follows:

$$\begin{aligned} g(\checkmark|\mathsf{s}) &= 0.72, & g(\checkmark|\mathsf{f}) &= 0.45, & f(\checkmark|\mathsf{s}) &= 0.96, & f(\checkmark|\mathsf{f}) &= 0.6, \\ g(\times|\mathsf{s}) &= 0.28, & g(\times|\mathsf{f}) &= 0.55, & f(\times|\mathsf{s}) &= 0.04, & f(\times|\mathsf{f}) &= 0.4. \end{aligned}$$

However, suppose that the test results cannot be trusted when both are executed on the same rod, for example because either test may afflict structural damage to the rod even upon passing the test. Then only one of g or f can be performed, and we are interested to know whether doing one of the experiments provides us with strictly more information about whether a given rod is safe. To that end, we can use the ordering $\succeq$. In particular, we have $f \succeq g$, because the stochastic map

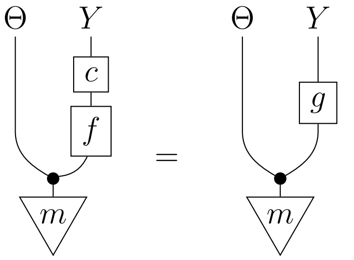
$$\begin{aligned} c(\checkmark|\checkmark) &= 0.75 & c(\checkmark|\times) &= 0 \\ c(\times|\checkmark) &= 0.25 & c(\times|\times) &= 1 \end{aligned}$$

achieves the conversion of the oscillation test to the threshold test by garbling.

The informativeness ordering defined via garbling maps can be equivalently expressed in terms of loss functions or Bayesian utilities in decision theory. For an account of the origins of this equivalence, see [26, Section 3] for example. Recently, it has also been expressed in categorical terms by de Oliveira in [11].

In what follows, we make use of the following relaxed informativeness ordering relative to a particular prior $m \colon I \to \Theta$.

Definition 5.2. *Given morphisms $m\colon I \to \Theta$, $f\colon \Theta \to X$ and $g\colon \Theta \to Y$ in any Markov category C, we say that f is m-*a.s. more informative *than g, denoted $f \succeq_{m\text{-a.s.}} g$, if there exists a morphism $c\colon X \to Y$ such that we have $c\,f =_{m\text{-a.s.}} g$,*

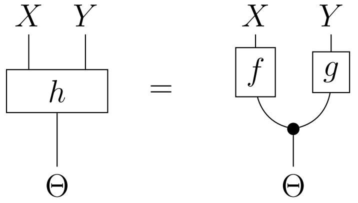


(5.1)

Note that the informativeness orderings are actually *preorders* on the collection of morphisms out of Θ, but we use the term "ordering" as synonymous with "preordering".

Remark 5.3. The informativeness ordering and its approximate versions that we leave out of our analysis here are particularly interesting in situations when the two experiments f and g *cannot* be implemented jointly or when it is costly to do so. They tell one which of the two mutually exclusive choices for an experiment is a better choice as far as learning about the hypothesis Θ is concerned.

On the other hand, if one *could* execute both f and g simultaneously, a more relevant question may be that of whether f (or g) is more informative than their joint implementation, which would be an experiment $h\colon \Theta \to X \otimes Y$ with marginals f and g. In practice, X and Y need not be conditionally independent given Θ. However, it is customary to assume so, in accordance with the assumption that Θ constitutes a complete set of relevant parameters. The joint experiment can then be expressed as



(5.2)

It is not hard to see that with this kind of joint implementation, the relation between f and g induced by $f \succeq_{m\text{-a.s.}} h$ is *different* from the basic informativeness ordering $f \succeq_{m\text{-a.s.}} g$. Nevertheless, as Theorem 5.4 below shows, the latter *is* closely connected to the question of whether X (or Y) constitutes a *sufficient statistic* for $X \times Y$, but the joint implementation of f and g is not in general the one given by equation (5.2) (in particular, see the left equation of (5.4)).

The considerations in the paragraph above do not necessarily take into account that there might be a cost associated with performing the experiments, which would make the comparison of the information contained in two samples (one from X and one from Y) with that of one sample (from X, say) less meaningful.[10] Instead, if one compares the same number of samples, either one from X or one from Y, then the relevant relation is the basic informativeness ordering given by $f \succeq_{m\text{-a.s.}} g$.

A slightly different context in which the informativeness ordering is relevant is the theory of communication. There, we commonly interpret f and g as distinct encodings of Θ in X and Y respectively. In general, f and g lose some of the information contained in Θ. The preorder $\succeq$ then tells us which of the two encodings unambiguously retains more of this information, and whether such a comparison can be made at all.

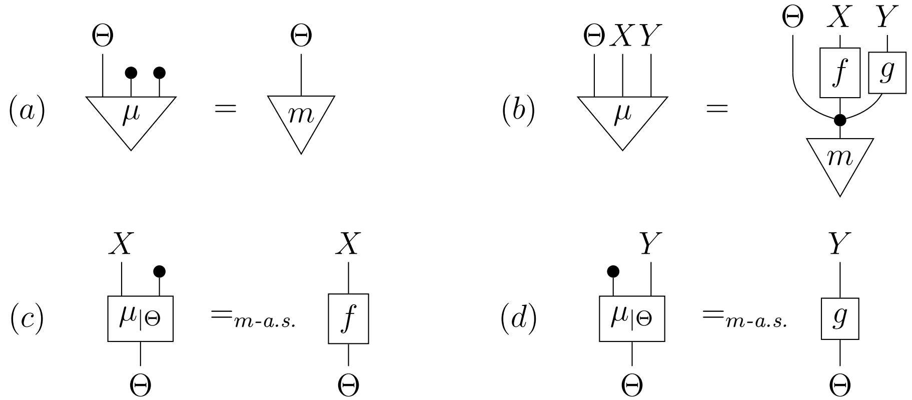
A measure-theoretic version the following theorem appears in [32, Theorem 7.2.16], but its roots can be traced back to [1].

Theorem 5.4. *Let $m \colon I \to \Theta$, $f \colon \Theta \to X$ and $g \colon \Theta \to Y$ be morphisms in a Markov category C. If C has conditionals, then the following are equivalent:*

(i) f is m-a.s. more informative than g.

(ii) There exists a morphism $h \colon \Theta \to X \otimes Y$ with marginals m-almost surely given by f and g respectively, such that the deterministic morphism $\mathrm{id}_X \otimes \mathrm{del}_Y$ is a sufficient statistic[11] for h.

(iii) There exists a morphism $\mu \colon I \to \Theta \otimes X \otimes Y$ satisfying



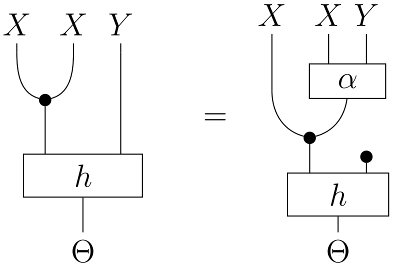
[10] For example, consider having to make a choice between one of two large-scale medical trials, in a situation where conducting both would not be feasible.

[11] See [13, Definition 14.3] for a synthetic definition of sufficiency in Markov categories.

The reading of condition (b) is that there are some morphisms that can take place of the empty boxes so that the equation holds. In other words, it states that μ displays the conditional independence [13, Definition 12.1] of Θ and Y given X. Also note that conditions (c) and (d) are independent of the choice of conditionals by the a.s.-uniqueness of conditionals.

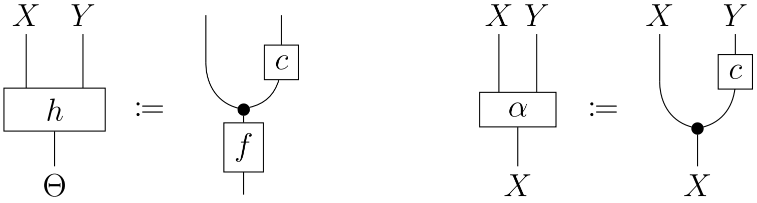
Proof. We split the proof into three implications.

(i) ⇒ (ii): The condition that $\mathrm{id}_X \otimes \mathrm{del}_Y$ is a sufficient statistic for h [13, Definition 14.3] translates to the existence of a morphism $\alpha \colon X \to X \otimes Y$ such that



(5.3)

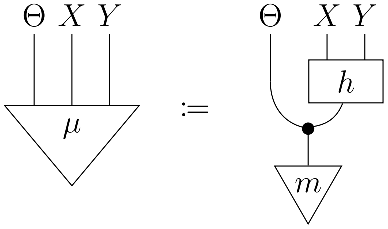
holds. Assuming condition (i) is true, we can use the garbling map c satisfying equation (5.1) to define the requisite h and α via



(5.4)

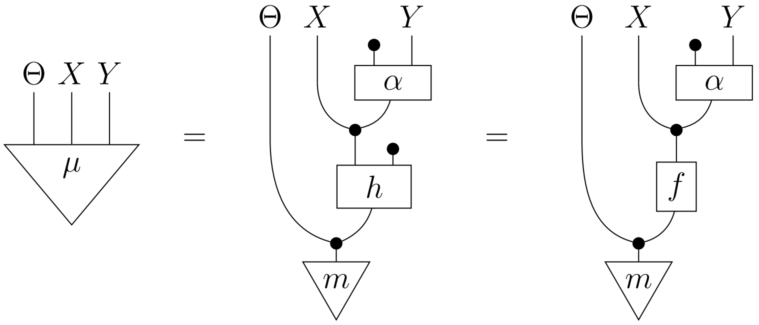
from which equation (5.3) follows. Since the marginals of h so defined are f and g respectively (the latter m-almost surely), we obtain the desired implication (i) $\implies$ (ii).

(ii) ⇒ (iii): We can define a morphism μ in terms of h and m as follows:



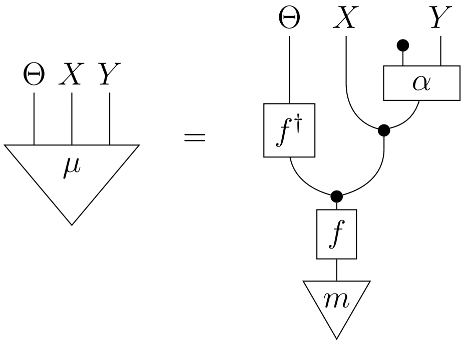
(5.5)

Conditions (a), (c), and (d) are then immediate. In order to show that condition (b) holds as well, we can use equation (5.3) with the middle output marginalized to get



(5.6)

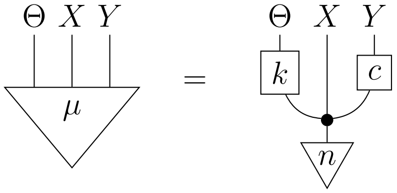
where the second equation follows from assuming condition (ii), in particular from the fact that the $\Theta \to X$ marginal of h is m-a.s. equal to f. In order to obtain condition (b), we can then apply the definition of a Bayesian inverse of f with respect to m to get



(5.7)

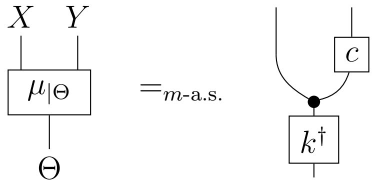
as required. Consequently, condition (ii) indeed implies condition (iii).

(iii) ⇒ (i): Let the decomposition of μ as given by condition (b) be



(5.8)

It remains to show that the morphism $c\colon X \to Y$ from such a decomposition can act as the garbling that achieves the conversion of f to g, m-almost surely. Taking the conditional of equation (5.8) and using $k\,n = m$, which follows from condition (a), yields



(5.9)

where $k^\dagger$ denotes the Bayesian inverse of k with respect to n. Marginalizing over Y in equation (5.9) and using condition (c) then gives $k^\dagger =_{m\text{-a.s.}} f$. Finally, by marginalizing equation (5.9) over X and using condition (d), we arrive at $g =_{m\text{-a.s.}} c\, f$ and the proof is thus complete. □

5.2 The Classical Blackwell–Sherman–Stein Theorem

Insofar as there is a host of equivalent ways to characterize the informativeness ordering [26, Theorem 1], we are interested in the extent to which one can generalize these results to the abstract setting of Markov categories and proved with synthetic methods. We already saw an instance of such a result in the form of Theorem 5.4, which relates informativeness to sufficient statistics and conditional independence. For the remainder of Section 5, we focus on the *Blackwell–Sherman–Stein Theorem* [2, Theorem 6], also known as the *dilation criterion*. It states that the informativeness order $\succeq$ for statistical experiments coincides with the second-order stochastic dominance order of the so-called *standard measures* on $P\Theta$, the description of which we turn to now.

An intuitive way to think of standard measures is to interpret a statistical experiment $f\colon \Theta \to X$ as a way of learning about the underlying hypothesis represented by the parameter set Θ. From this perspective, given a prior $m \in P\Theta$, one can use Bayesian updating to find the corresponding posterior on $P\Theta$, which of course depends on the value of X observed. Thus, it can be represented by the measurable map $(f^\dagger)^\sharp\colon X \to P\Theta$, which, recalling Notation 3.8, contains the same information as the Markov kernel $f^\dagger\colon X \to \Theta$—the Bayesian inverse of f. The standard measure is then an element of $PP\Theta$ obtained as a mixture of these posteriors with respect to the chosen prior m, or more precisely, with respect to the distribution of X one would expect, were the "true" value of the hypothesis Θ sampled from m.

The morphism given by the composite

$$\Theta \xrightarrow{f} X \xrightarrow{(f^\dagger)^\sharp} P\Theta \tag{5.10}$$

is commonly referred to as the *standard experiment* $\hat{f}$ of f. Intuitively, for a given "true" hypothesis as input, the standard experiment outputs the distribution over posteriors that results from conducting the experiment and applying Bayesian updating. Since the outcome of the experiment itself is random, we obtain a distribution over posteriors rather than a mere posterior. Therefore, the standard experiment is typically not deterministic and it clearly depends on the prior m. The standard measure is then nothing but $\hat{f}$ applied to m.

Example 5.5. Traditionally, one uses a uniform prior m to define standard measures for finite parameter sets, but any strictly positive probability measure would do. Continuing our Example 5.1 and choosing a uniform prior

$$m(\mathsf{s}) = 0.5 \qquad\qquad m(\mathsf{f}) = 0.5$$

gives the following posteriors (expressed as functions rather than Markov kernels):

$$\begin{aligned}
\left(g^{\dagger}\right)^{\sharp}(\checkmark) &= \frac{0.72\,\delta_{\mathsf{s}} + 0.45\,\delta_{\mathsf{f}}}{0.72 + 0.45} \approx 0.62\,\delta_{\mathsf{s}} + 0.38\,\delta_{\mathsf{f}} \\
\left(g^{\dagger}\right)^{\sharp}(\times) &= \frac{0.28\,\delta_{\mathsf{s}} + 0.55\,\delta_{\mathsf{f}}}{0.28 + 0.55} \approx 0.34\,\delta_{\mathsf{s}} + 0.66\,\delta_{\mathsf{f}} \\
\left(f^{\dagger}\right)^{\sharp}(\checkmark) &= \frac{0.96\,\delta_{\mathsf{s}} + 0.6\,\delta_{\mathsf{f}}}{0.96 + 0.6} \approx 0.62\,\delta_{\mathsf{s}} + 0.38\,\delta_{\mathsf{f}} \\
\left(f^{\dagger}\right)^{\sharp}(\times) &= \frac{0.04\,\delta_{\mathsf{s}} + 0.4\,\delta_{\mathsf{f}}}{0.04 + 0.4} \approx 0.09\,\delta_{\mathsf{s}} + 0.91\,\delta_{\mathsf{f}}.
\end{aligned}$$

The expected distributions of test outcomes if exactly half of the rods were faulty are

$$\begin{aligned}
g \circ m(\checkmark) &= 0.5\,g(\checkmark|\mathsf{s}) + 0.5\,g(\checkmark|\mathsf{f}) = 0.585 \\
g \circ m(\times) &= 0.5\,g(\times|\mathsf{s}) + 0.5\,g(\times|\mathsf{f}) = 0.415 \\
f \circ m(\checkmark) &= 0.5\,f(\checkmark|\mathsf{s}) + 0.5\,f(\checkmark|\mathsf{f}) = 0.78 \\
f \circ m(\times) &= 0.5\,f(\times|\mathsf{s}) + 0.5\,f(\times|\mathsf{f}) = 0.22
\end{aligned}$$

Therefore the standard measures of g and f, denoted by $\hat{g}_m$ and $\hat{f}_m$ respectively, are given by

$$\begin{aligned}
\hat{g}_m &\approx 0.585\,\delta_{0.62\,\delta_{\mathsf{s}}+0.38\,\delta_{\mathsf{f}}} + 0.415\,\delta_{0.34\,\delta_{\mathsf{s}}+0.66\,\delta_{\mathsf{f}}}, \\
\hat{f}_m &\approx 0.78\,\delta_{0.62\,\delta_{\mathsf{s}}+0.38\,\delta_{\mathsf{f}}} + 0.22\,\delta_{0.09\,\delta_{\mathsf{s}}+0.91\,\delta_{\mathsf{f}}}.
\end{aligned}$$

One may wonder why would one consider such a seemingly convoluted way to view a statistical experiment in terms of its standard measure. As mentioned at the beginning of Section 5.2, one reason is that we can express comparison of statistical experiments in terms of the second-order dominance ordering among their standard measures. Explicitly, this is the classical version of the BSS Theorem. One of its strong points is that it reduces the comparison of experiments $\Theta \to X$ for *arbitrary* X to the comparison of measures on a *fixed* sample space, namely $P\Theta$.

Theorem 5.6 (Blackwell–Sherman–Stein [26])**.** *Let Θ, X, and Y be standard Borel spaces with Θ finite. For any two Markov kernels $f\colon \Theta \to X$ and $g\colon \Theta \to Y$, the following are equivalent:*

1. *$f \succeq g$, i.e. f is more informative about Θ than g is.*
2. *$\hat{f}_m \sqsubseteq \hat{g}_m$, i.e. the standard measure of g second-order dominates the standard measure of f.*

Example 5.7. In terms of our running example of metal rods, we saw that f is more informative than g in Example 5.1 and constructed the standard measures in Example 5.5. Theorem 5.6 says that there should be a $\hat{g}_m$-dilation $t\colon P\Theta \to P\Theta$ that maps $\hat{g}_m$ to $\hat{f}_m$, meaning that the posteriors according to the oscillation test f are "more spread out" than those corresponding to the threshold test g. Concretely, a dilation that does the job can be chosen to be any measurable map $t^\sharp$ satisfying

$$\begin{aligned} &0.62\,\delta_{\mathsf{s}} + 0.38\,\delta_{\mathsf{f}} \mapsto \delta_{0.62\,\delta_{\mathsf{s}}+0.38\,\delta_{\mathsf{f}}} \\ &0.34\,\delta_{\mathsf{s}} + 0.66\,\delta_{\mathsf{f}} \mapsto 0.47\,\delta_{0.62\,\delta_{\mathsf{s}}+0.38\,\delta_{\mathsf{f}}} + 0.53\,\delta_{0.09\,\delta_{\mathsf{s}}+0.91\,\delta_{\mathsf{f}}} \end{aligned}$$

up to our convention of rounding to two decimal places. One can use Lemma 4.4 to convince oneself that these relations indeed give a $\hat{g}_m$-dilation t, since we have

$$0.34 \approx 0.47 * 0.62 + 0.53 * 0.09, \qquad 0.66 \approx 0.47 * 0.38 + 0.53 * 0.91.$$

5.3 The Blackwell–Sherman–Stein Theorem in Markov Categories

Our goal is now to state and prove a version of Theorem 5.6 for Markov categories. In order to arrive at such a synthetic generalization, we need a bit more than just a representable Markov category. Indeed, the requirements are the same as in Proposition 4.5.

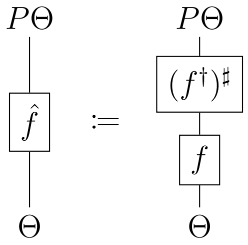
Assumption 5.8. *Throughout the rest of Section 5, let* C *be an a.s.-compatibly representable Markov category with conditionals, as well as* $f\colon \Theta \to X$ *and* $g\colon \Theta \to Y$ *arbitrary morphisms in* C *with the same domain.*

We mention this assumption again in the statements of our results, but otherwise leave it implicit.

The existence of conditionals is necessary in order to have an abstract notion of Bayesian inference, which is used to construct the basic elements of the BSS Theorem: standard experiments and standard measures. The representability is relevant again for the definition of the standard experiment and standard measure, which make reference to the distribution functor P. The a.s.-compatibility, in the sense of Definition 3.21, is relevant for proving a.s. uniqueness of the standard experiment, and for the interpretation of the second condition of the upcoming Theorem 5.13 as a second-order dominance relation via Lemma 4.4.

Next, we present the definitions of standard experiment and standard measure in the language of Markov categories.

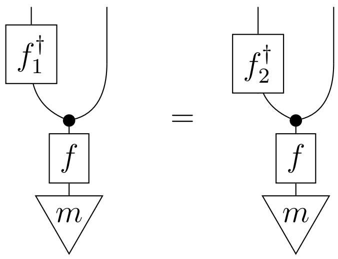
Definition 5.9. *Given morphisms* $m\colon I \to \Theta$ *and* $f\colon \Theta \to X$ *in a Markov category* C *with conditionals, the* standard experiment $\hat{f}\colon \Theta \to P\Theta$ *of* f *is given by*



(5.11)

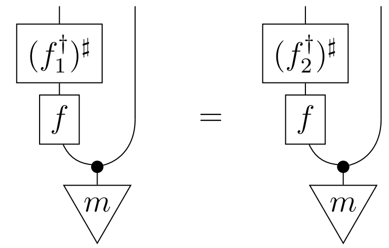
where $f^\dagger\colon X \to \Theta$ *is a Bayesian inverse of* f *with respect to* m.

Standard experiments are unique, m-almost surely, as long as the Markov category is a.s.-compatibly representable and has conditionals. To see this, consider two Bayesian inverses of f with respect to m, $f_1^\dagger$ and $f_2^\dagger$, which thus have to satisfy



(5.12)

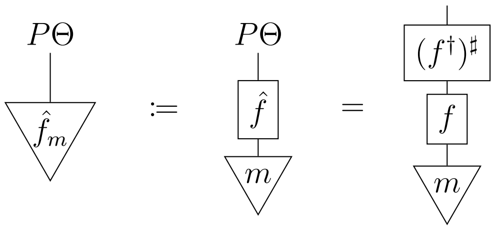
by the definition of Bayesian inverses. Applying the sampling cancellation property and the causality property [13, Definition 11.30] which follows from the existence of conditionals [13, Proposition 11.33] gives



(5.13)

which is exactly the equation needed to conclude that $\hat{f}$ is well-defined up to m-a.s. equality.

Definition 5.10. *The* standard measure $\hat{f}_m \colon I \to P\Theta$ *of* f *is then defined by*

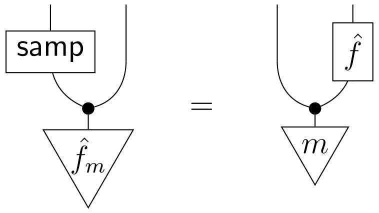


(5.14)

Per the m-almost sure uniqueness of $\hat{f}$, the standard measure $\hat{f}_m$ is unique as soon as C is a.s.-compatibly representable.

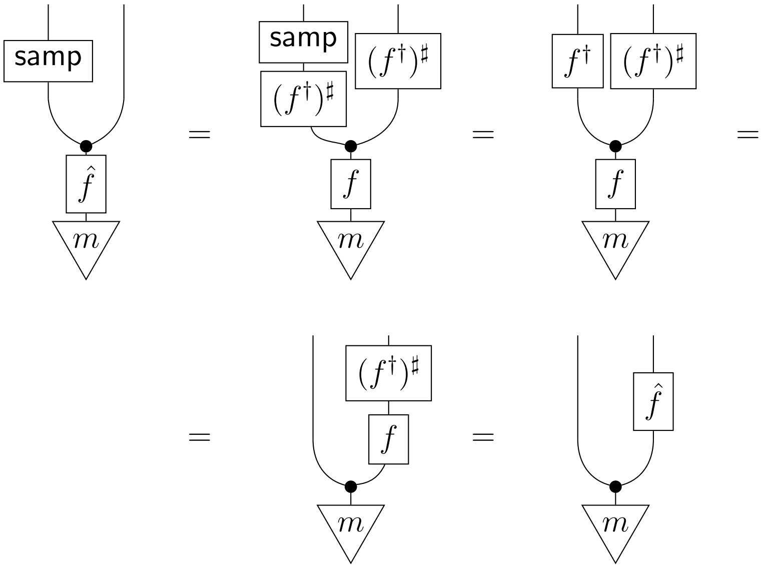
Having introduced the basic necessary ingredients of the theory of comparison of statistical experiments and the BSS Theorem in the language of Markov categories, we now present the results that build up to the BSS Theorem itself, assuming throughout that we are in an a.s.-compatibly representable Markov category with conditionals.

Lemma 5.11. *Let* $f \colon \Theta \to X$ *be a morphism with standard experiment* $\hat{f} \colon \Theta \to P\Theta$ *as defined above. Then the sampling map* $\mathsf{samp}_\Theta \colon P\Theta \to \Theta$ *is a Bayesian inverse of* $\hat{f}$ *with respect to* m*, i.e. we have*



(5.15)

Proof. Since, by definition, $(f^\dagger)^\sharp$ is deterministic and satisfies $\mathsf{samp} \circ (f^\dagger)^\sharp = f^\dagger$, we can prove the lemma as follows:



(5.16)

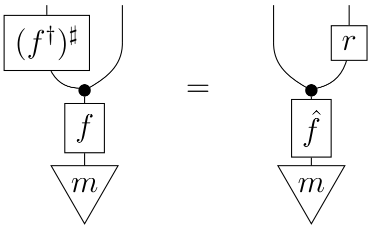
□

The terminology "standard experiment" of $\hat{f}$ is then justified by the following result, which states that $\hat{f}$ is exactly as informative about Θ as the original experiment f is, at least up to m-a.s. equality.

Proposition 5.12. *For any f, we have $f \succeq \hat{f}$ and $\hat{f} \succeq_{m\text{-a.s.}} f$.*

In standard measure-theoretic probability, this is [32, Proposition 7.2.2].

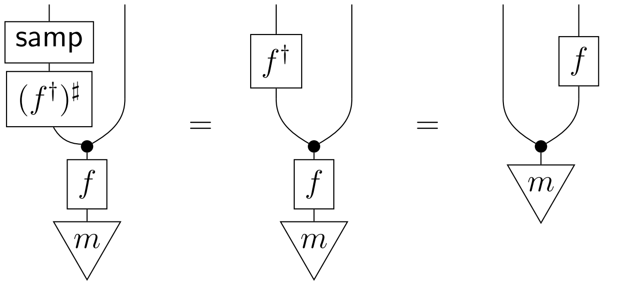
Proof. Since $f \succeq \hat{f}$ is clear from the definition of the standard experiment, the crux of the proof lies in showing that there exists a morphism $r \colon P\Theta \to X$ such that $r\,\hat{f} =_{m\text{-a.s.}} f$. That is, r is a garbling map which recovers f from its standard version. We now show that choosing r to be a Bayesian inverse of $(f^\dagger)^\sharp$ with respect to $f\,m$ does the job. With this choice, we thus have



(5.17)

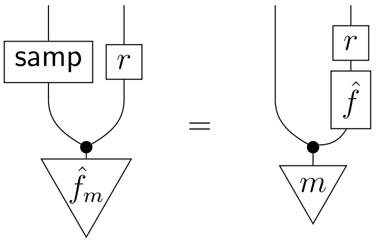
by the definition of Bayesian inverses.

Applying $\mathsf{samp}_\Theta \otimes \mathrm{id}_X$ to this equation yields



(5.18)

for its left-hand side and



(5.19)

for its right hand side, where we use Lemma 5.11 to obtain the latter. Consequently, f is m-almost surely equal to $r\,\hat{f}$, as we wanted to show. □

Now we have all the necessary ingredients to present a proof of our version of the Blackwell—Sherman—Stein Theorem in representable Markov categories.

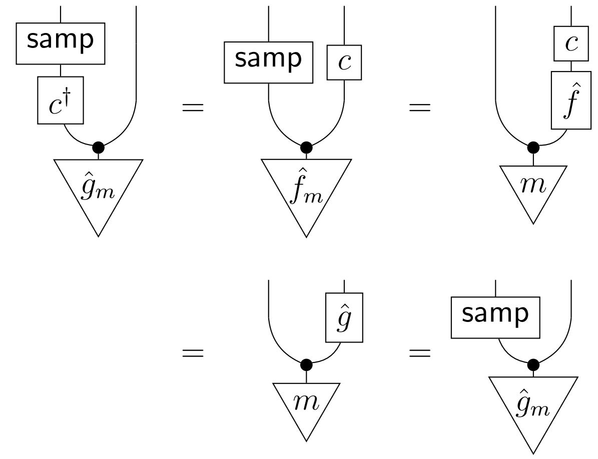
Theorem 5.13 (Blackwell—Sherman—Stein)**.** *Let C be an a.s.-compatibly representable Markov category with conditionals. Consider two morphisms $f\colon \Theta \to X$ and $g\colon \Theta \to Y$ in C, whose standard experiments are denoted by $\hat{f}$ and $\hat{g}$ respectively.*

Then the following are equivalent:

1. *$f \succeq_{m\text{-a.s.}} g$, i.e. there exists a morphism $c\colon X \to Y$ such that*
$$c\,f =_{m\text{-a.s.}} g. \tag{5.20}$$
2. *$\hat{f}_m \sqsubseteq \hat{g}_m$, i.e. there exists a $\hat{g}_m$-dilation $t\colon P\Theta \to P\Theta$ such that*
$$\hat{f}_m = t\,\hat{g}_m. \tag{5.21}$$

Proof. For the forward implication, let $c\colon P\Theta \to P\Theta$ be a morphism satisfying $c\,\hat{f} =_{m\text{-a.s.}} \hat{g}$, the existence of which is equivalent to $f \succeq_{m\text{-a.s.}} g$ by Proposition 5.12.

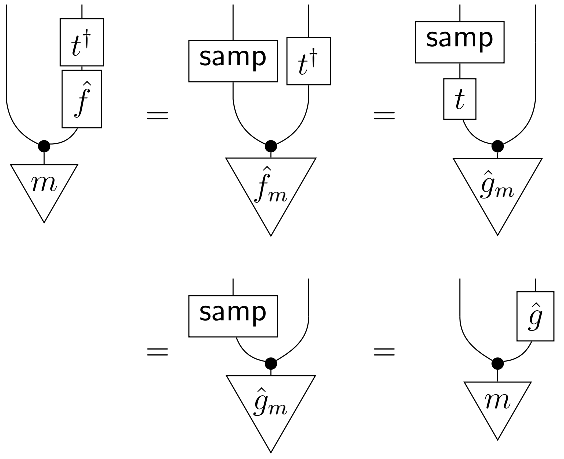
We then prove that $\hat{f}_m \sqsubseteq \hat{g}_m$ holds as well by constructing a $\hat{g}_m$-dilation that witnesses this second-order dominance relation. Let $c^\dagger$ denote the Bayesian inverse of c with respect to $\hat{f}_m$. Then $c^\dagger$ satisfies equation (5.21) by definition. In order to show that $c^\dagger$ is a $\hat{g}_m$-dilation, we can use Lemma 5.11 twice:



(5.22)

Overall, $c^\dagger$ is thus a $\hat{g}_m$-dilation that maps $\hat{g}_m$ to $\hat{f}_m$, as was to be shown.

For the converse implication, suppose that a $\hat{g}_m$-dilation t with $\hat{f}_m = t\,\hat{g}_m$ exists, and denote its Bayesian inverse with respect to $\hat{g}_m$ by $t^\dagger$. We can now use the same steps as in the computation above in order to show that $t^\dagger$ achieves the conversion of $\hat{f}$ into $\hat{g}$, at least m-almost surely:



(5.23)

We have thus shown that $t^\dagger$ achieves the conversion $\hat{f} \succeq_{m\text{-a.s.}} \hat{g}$ and consequently we get $f \succeq_{m\text{-a.s.}} g$ by Proposition 5.12. □

Traditionally, the BSS Theorem is stated with exact equalities rather then almost surely with respect some measure $m \in \mathsf{C}_{\mathrm{det}}(I, P\Theta)$. This is because the theorem is usually applied to finite (or countably infinite) parameter sets Θ, for which there exists a distribution $m\colon I \to \Theta$ having full support, so that m-almost sure equality coincides with plain equality,

$$f =_{m\text{-a.s.}} g \quad \iff \quad f = g \tag{5.24}$$

for all $f, g\colon \Theta \to A$ to any other object A. Thanks to this property, one can then remove all m-a.s. qualifications. This generalizes as follows.

Definition 5.14. *An object Θ of a Markov category is termed* discrete *if there exists a morphism $m\colon I \to \Theta$ such that (5.24) holds for all $f, g\colon \Theta \to A$ to any A.*

Clearly such an m satisfies $m \gg \mu$ for every other $\mu\colon I \to \Theta$ in the sense of Definition 2.8. In $\mathsf{BorelStoch}$, the only standard Borel spaces Θ for which such an m exists are the discrete measurable spaces, meaning that Θ must be finite or countably infinite. Then one can choose m to be any distribution of full support on Θ, where in the case of finite Θ the uniform distribution is the commonly used choice.

For discrete objects, we get the following version of the BSS Theorem, which is closer to the traditional account than Theorem 5.13.

Corollary 5.15. *Let C be an a.s.-compatibly representable Markov category with conditionals, and let Θ be a discrete object of C with respect to $m\colon I \to \Theta$. Consider two morphisms $f\colon \Theta \to X$ and $g\colon \Theta \to Y$ in C, whose standard experiments are denoted by $\hat{f}$ and $\hat{g}$ respectively.*

Then the following are equivalent:

1. *$f \succeq g$, i.e. there exists a morphism $c\colon X \to Y$ such that*

$$c\, f = g. \tag{5.25}$$

2. *$\hat{f}_m \sqsubseteq \hat{g}_m$, i.e. there exists a $\hat{g}_m$-dilation $t\colon P\Theta \to P\Theta$ such that*

$$\hat{f}_m = t\, \hat{g}_m. \tag{5.26}$$

As far as we know, there is no established genuine generalization of the BSS Theorem beyond discrete parameter sets Θ. In Theorem 5.13, we have given a version of the theorem that goes beyond discrete parameters by virtue of using a.s. equality with respect to a measure m. In the specific case of $\mathsf{BorelStoch}$, the possibility of doing this has been known, see for example [26]. However, we believe that our synthetic treatment exposes the key reasons for this to be the case and can therefore catalyze further development.

Arguably, the more interesting aspect of our Theorem 5.13 is the fact that it applies in a much wider context than the traditional measure-theoretic one. In particular, we have shown that the result can be used in any Markov category that

- allows one to perform Bayesian inference—i.e. one that has conditionals, and
- can describe spaces of measures internally—i.e. one that is a.s.-compatibly representable.

So far, we have not even begun to explore the full scope of this type of result. We explore one particular application in the next subsection.

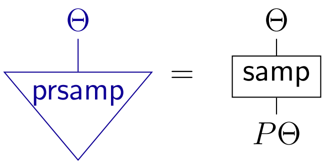
5.4 The Blackwell–Sherman–Stein Theorem Parametrized by Priors

In this subsection, we show that there is a (synthetic) version of the BSS Theorem which

- holds for an arbitrary hypothesis object Θ, and
- does not refer to any particular prior m,

in contrast to Theorem 5.13 and Corollary 5.15, which only satisfy one of these requirements each. The catch is that, in general, it characterizes a slightly different informativeness ordering, and can be thought of as a version of Theorem 5.13 which is uniform in the prior. Interestingly, this result actually arises as a *special case* of Theorem 5.13, namely when the latter is instantiated in a parametric Markov category as introduced in Section 2.2.

As before, we think of f and g as statistical experiments with hypothesis space or parameter space Θ. Given the corresponding distribution object $P\Theta$ associated to the hypothesis space Θ, we then consider the parametric Markov category $\mathsf{C}_{P\Theta}$. Intuitively, we think of morphisms in $\mathsf{C}_{P\Theta}$ as Markov kernels parametrized by a prior over the hypothesis space Θ. Throughout this section, we use a.s. equality in $\mathsf{C}_{P\Theta}$ with respect to the global element $\mathsf{prsamp} \in \mathsf{C}_{P\Theta}(I, \Theta)$ represented by the sampling map:

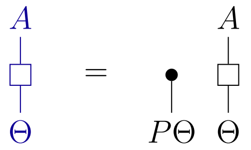


(5.27)

We think of this morphism as sampling a hypothesis in Θ distributed in accordance with the prior (which is not fixed here, but rather an extra parameter).

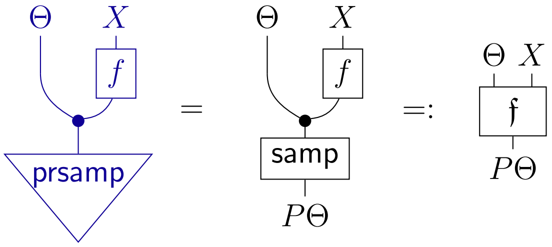
Before we discuss the BSS Theorem instantiated in $\mathsf{C}_{P\Theta}$, we consider how the basic notions of comparison of statistical experiments look in $\mathsf{C}_{P\Theta}$ when we reexpress them in terms of the corresponding morphisms in C.

First of all, the statistical models $\Theta \to X$ are thought of as models of the behavior of a system as a function of a model parameter Θ, typically not under the experimenter's control. The point of the theory of statistical experiments is to formalize and quantify the procedure of making inferences about an *unknown* parameter Θ by virtue of learning the value of X. We thus still assume that the statistical models in $\mathsf{C}_{P\Theta}$ are morphisms independent of the prior $P\Theta$ and are therefore represented by



(5.28)

For better readability in future expressions in the C-representation, we introduce a new notation for the *prior behavior* $\mathfrak{f} \colon P\Theta \to \Theta \otimes X$ of a statistical model f as follows



(5.29)

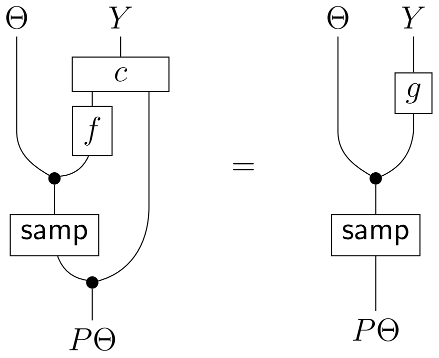
The morphism $\mathfrak{f}$ describes what a Bayesian experimenter expects to observe: For each prior distribution over hypotheses, it returns as outputs an experiment outcome, for a hypothesis randomly sampled from the prior, together with that hypothesis.

The process of Bayesian updating itself is then described by

$$\left(\mathfrak{f}_{|X}\right)^{\sharp} \colon X \otimes P\Theta \longrightarrow P\Theta,$$

where $\mathfrak{f}_{|X}$ denotes a conditional of $\mathfrak{f}$ with respect to X. The deterministic morphism $(\mathfrak{f}_{|X})^{\sharp}$ takes an outcome and a prior as input and returns the associated Bayesian posterior. Since conditionals are not unique in general, this Bayesian updating map is not uniquely determined by f itself. As before, Bayesian updating features in the upcoming definition of the standard experiment.

But let us consider the comparison of statistical experiments in $\mathsf{C}_{P\Theta}$ first. Interpreting Definition 5.2 in $\mathsf{C}_{P\Theta}$, the prsamp-a.s. informativeness ordering of two prior-independent morphisms as in equation (5.28) says that $f \succeq_{\mathsf{prsamp\text{-}a.s.}} g$ if and only if there exists $c \colon X \otimes P\Theta \to Y$ such that we have



(5.30)

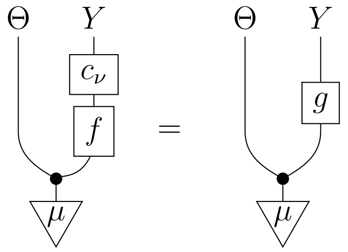
This amounts to allowing the garbling map c in the definition of the informativeness ordering to depend on the prior in addition to its dependence on the data variable of the experiment f used to simulate g. Since the condition (5.30) gives rise to an ordering on morphisms of C that differs from those considered in Sections 5.1 and 5.3, we give it a new name.

Notation 5.16. *Given two morphisms f and g in a Markov category C with a common domain Θ, we say that f is* more informative *than g* in the Bayesian sense*, denoted $f \succeq_{\mathrm{Bayes}} g$, if we have $f \succeq_{\mathsf{prsamp}\text{-}a.s.} g$ in $\mathsf{C}_{P\Theta}$ as expressed by the existence of a prior-dependent garbling c that achieves the conversion shown in equation* (5.30).

Before the exposition of our parametric BSS Theorem itself, we present some results on how the Bayesian informativeness ordering relates to the usual prior-independent one. While it is obvious that the existence of a prior-independent c with $c\,f = g$ implies the existence of a prior-dependent c as above—namely by choosing that dependence to be the trivial one which discards the prior—the question of whether the converse also holds is far more involved. We start addressing it with the following simple observation, which shows that the garbling map can be chosen in a uniform way for all distributions which are absolutely continuous with respect to a given one.

Proposition 5.17. *Let $\nu \colon I \to \Theta$ be arbitrary. If f is more informative than g in the Bayesian sense, then there exists a morphism $c_\nu \colon X \to Y$ in C such that we have*

$c_\nu f =_{\mu\text{-a.s.}} g$, i.e.

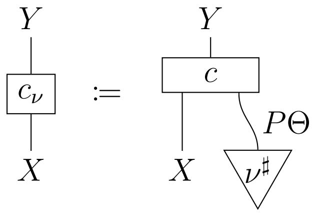


(5.31)

for every $\mu\colon I \to \Theta$ with $\nu \gg \mu$.

Recall that $\nu \gg \mu$ denotes the absolute continuity ordering from Definition 2.8.

Proof. In particular, c_ν can be constructed from c in equation (5.30) as



(5.32)

which makes (5.31) hold with ν in place of μ. The claim then follows from the definition of $\nu \gg \mu$. □

Corollary 5.18. *If the hypothesis object Θ is discrete, then*

$$f \succeq g \qquad \Longleftrightarrow \qquad f \succeq_{\text{Bayes}} g.$$

Proof. Discreteness implies that there is a distribution $\nu\colon I \to \Theta$ such that taking $\mu := \nu$ makes $f \succeq g$ follow from equation (5.31). □

However, the existence of a prior-dependent garbling map does not imply the existence of a completely prior-independent one in general. We now present an explicit counterexample in $\mathsf{BorelStoch}$ based on an example due to Blackwell and Ramamoorthi [4].

Proposition 5.19. *In $\mathsf{BorelStoch}$, there are $f\colon \Theta \to X$ and $g\colon \Theta \to Y$ such that a garbling c satisfying equation (5.30) exists, but there is no Markov kernel $c\colon X \to Y$ with $g = c\,f$.*

Proof. The following arguments amount to showing that the example of Blackwell and Ramamoorthi [4], which proved that classical sufficiency and Bayesian sufficiency are not equivalent, also similarly in our context of comparison of experiments.

The main difference is that our version of their example requires a more refined measurability analysis in the second part.

As the spaces of outcomes of the experiments, consider the two standard Borel spaces

$$X = \{a,b\}^{\mathbb{N}}, \qquad\qquad Y = \{a,b\}.$$

Writing $\pi_n \colon \{a,b\}^{\mathbb{N}} \to \{a,b\}$ for the n-th product projection, we define the sets

$$\begin{aligned} \Theta_a &\coloneqq \Big\{ \nu \in PX \;\Big|\; \lim_{n\to\infty} \nu\big(\pi_n^{-1}(a)\big) = 1 \Big\}, \\ \Theta_b &\coloneqq \Big\{ \nu \in PX \;\Big|\; \lim_{n\to\infty} \nu\big(\pi_n^{-1}(b)\big) = 1 \Big\}, \end{aligned}$$

to be thought of as containing those distributions for which the countably many $\{a,b\}$-valued random variables represented by a distribution on X converge in probability to a or b, respectively. Since we have

$$\Theta_a = \bigcap_{k\in\mathbb{N}} \bigcup_{n\in\mathbb{N}} \Big\{ \nu \in PX \;\Big|\; \nu\big(\pi_n^{-1}(b)\big) \leq 2^{-k} \Big\},$$

the set Θ_a is a measurable subset of PX, and we take it to be equipped with the induced σ-algebra. Moreover, it constitutes a standard Borel space since PX does too. A similar argument shows that Θ_b is likewise a standard Borel space.

We now consider the (disjoint) union $\Theta \coloneqq \Theta_a \cup \Theta_b$, and let a measurable map $f \colon \Theta \to X$ be defined by the restriction of $\mathsf{samp} \colon PX \to X$ to $\Theta \subseteq PX$, while the measurable map $g \colon \Theta \to Y$ is given by

$$g(\nu) \coloneqq \begin{cases} a & \text{if } \nu \in \Theta_a, \\ b & \text{if } \nu \in \Theta_b. \end{cases}$$

That is, $g(\nu) = a$ indicates that the sequence of random variables distributed according to ν converges in probability to a, and similarly for $g(\nu) = b$.

We first argue that there is no garbling from f to g, meaning that there is no Markov kernel $c \colon X \to Y$ with $g = c\,f$. If such a c did exist, then one could define a measurable map $s \colon X \to \{a,b\}$ by

$$s(x) \coloneqq \begin{cases} a & \text{if } c\big(\{a\}\big|x\big) > 0, \\ b & \text{if } c\big(\{a\}\big|x\big) = 0. \end{cases}$$

Since for $\nu \in \Theta_a$ we have $c(\{a\}|\_) =_{\nu\text{-a.s.}} 1$, it follows that also $s =_{\nu\text{-a.s.}} a$ holds. Similarly, for every $\nu \in \Theta_b$ we have $s =_{\nu\text{-a.s.}} b$. However, Blackwell has shown in [3] that such an s does not exist, and therefore neither does c.

Second, we construct a prior-dependent $c\colon X \otimes P\Theta \to Y$ such that equation (5.30) holds. In particular, note that even though Θ is defined so that the sequence of random variables converges in probability to either a or b, we cannot extract the limit from a sample sequence alone per the argument of the previous paragraph. We thus show that, given a prior $\mu \in P\Theta$, one can always find a subsequence of the random variables that converges ν-*almost surely* for μ-almost all measures $\nu \in \Theta$.

This requires a bit of preparation. For fixed $\mu \in P\Theta$, consider the average measures defined for all $S \in \Sigma_X$ as

$$\overline{\nu}_a(S) := \int_{\Theta_a} \nu(S)\, \mu(d\nu), \qquad\qquad \overline{\nu}_b(S) := \int_{\Theta_b} \nu(S)\, \mu(d\nu). \tag{5.33}$$

Note that these are both subnormalized, and that $\overline{\nu}_a + \overline{\nu}_b$ is exactly the expected probability measure $(Pf)(\mu)$ on X. By the monotone convergence theorem, the assumed convergence in probability of the product projections still holds on average, i.e. we have

$$\lim_{n\to\infty} \overline{\nu}_a\big(\pi_n^{-1}(b)\big) = 0, \qquad\qquad \lim_{n\to\infty} \overline{\nu}_b\big(\pi_n^{-1}(a)\big) = 0. \tag{5.34}$$

For $k \in \mathbb{N}$, let then $n_k(\mu)$ be the smallest natural number such that both the inequalities

$$\overline{\nu}_a\big(\pi_{n_k(\mu)}^{-1}(b)\big) \le 2^{-k} \qquad\qquad \overline{\nu}_b\big(\pi_{n_k(\mu)}^{-1}(a)\big) \le 2^{-k} \tag{5.35}$$

are satisfied. The existence of $n_k(\mu)$ is guaranteed by (5.34).

We now show that the map $\mu \mapsto n_k(\mu)$ is measurable. To see this, notice that for every $m \in \mathbb{N}$ we have $n_k(\mu) \le m$ if and only if both

$$\int 1_{\Theta_a}\, \nu\big(\pi_m^{-1}(b)\big)\, \mu(d\nu) \le 2^{-k} \quad \text{ and } \quad \int 1_{\Theta_b}\, \nu\big(\pi_m^{-1}(a)\big)\, \mu(d\nu) \le 2^{-k}$$

are satisfied. Preimages of the upper sets $\{m \mid k \le m\} \subseteq \mathbb{N}$ are thus measurable by the inequalities above, because integration of a fixed measurable function on Θ is measurable in $\mu \in P\Theta$. Since the discrete σ-algebra on $\mathbb{N}$ is generated by the upper sets, the assignment $\mu \mapsto n_k(\mu)$ is measurable.

Thanks to inequalities (5.35) and the Borel-Cantelli lemma, we have that the measurable sets

$$\begin{aligned} S_a(\mu) &:= \big\{ x \in X \;\big|\; \pi_{n_k(\mu)}(x) = a \text{ for infinitely many } k \big\}, \\ S_b(\mu) &:= \big\{ x \in X \;\big|\; \pi_{n_k(\mu)}(x) = b \text{ for infinitely many } k \big\}, \end{aligned}$$

satisfy

$$\overline{\nu}_a\big(S_b(\mu)\big) = \overline{\nu}_b\big(S_a(\mu)\big) = 0.$$

But since this must then also hold μ-almost surely for all the measures which form the averages $\overline{\nu}_a$ and $\overline{\nu}_b$, it follows that $\nu(S_b) = 0$ also holds for μ-almost all $\nu \in \Theta_a$, and similarly $\nu(S_a) = 0$ for μ-almost all $\nu \in \Theta_b$. But then we can decide whether $\nu \in \Theta_a$ or $\nu \in \Theta_b$ simply by testing membership of its sample in S_a. In other words, the function

$$c \colon X \otimes P\Theta \to \{a, b\}, \qquad (x, \mu) \mapsto \begin{cases} a & \text{if } x \in S_a(\mu), \\ b & \text{if } x \in \overline{S_a(\mu)}, \end{cases}$$

makes the desired equation (5.39) hold, since $c(x, \mu) = a$ for ν-almost every $x \in X$ and μ-almost every $\nu \in \Theta_a$, and similarly $c(x, \mu) = b$ for ν-almost every x and μ-almost every $\nu \in \Theta_b$.

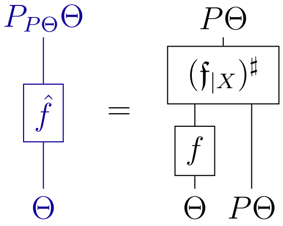
It remains to be shown that c is actually measurable. This follows because the complement $\overline{S_a(\mu)}$ is given by the countable disjoint union

$$\overline{S_a(\mu)} = \bigcup_F \Big\{ x \in \{a, b\}^{\mathbb{N}} \;\Big|\; \pi_{n_k(\mu)}(x) = b \iff k \in F \Big\}$$

over finite $F \subseteq \mathbb{N}$. Since for every fixed F the set of all pairs (x, μ) for which the internal condition holds is measurable by the measurability of $\mu \mapsto n_k(\mu)$, the set of all pairs (x, μ) with $c(x, \mu) = b$ is likewise measurable. Therefore, c is measurable, thereby making f indeed more informative than g in the Bayesian sense. □

We continue with the general theory, aiming at a BSS Theorem for characterizing informativeness in the Bayesian sense.

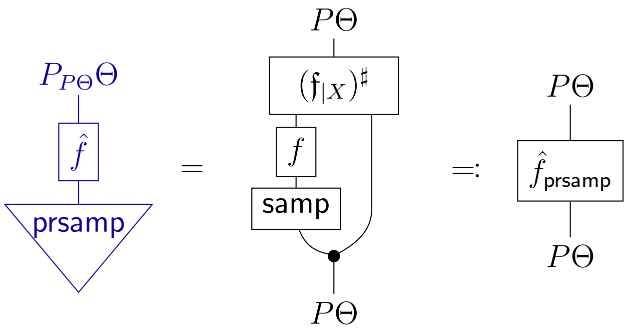
The standard experiment of a statistical model $f \colon \Theta \to X$ in $\mathsf{C}_{P\Theta}$, as introduced in Definition 5.9, now takes the form



(5.36)

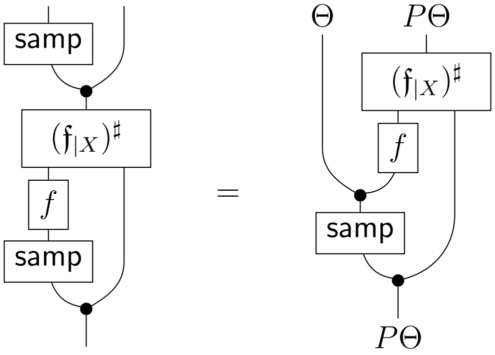
Similar to before, this (typically non-deterministic) morphism takes a hypothesis as its first argument, a prior as its second argument, and outputs the distribution over posteriors which results if the given hypothesis is true and Bayesian updating is used with respect to the given prior. Again as before, in order for $\hat{f}$ to be well defined up to prsamp-a.s. equality, we need $\mathsf{C}_{P\Theta}$ to be a.s.-compatibly representable as opposed to merely being representable. However, this is guaranteed by Assumption 5.8 and Lemma 3.25.

Next, the standard measure $\hat{f}_{\mathsf{prsamp}} := \hat{f} \circ \mathsf{prsamp}$ in $\mathsf{C}_{P\Theta}$ is represented in C by



(5.37)

and Lemma 5.11 becomes:



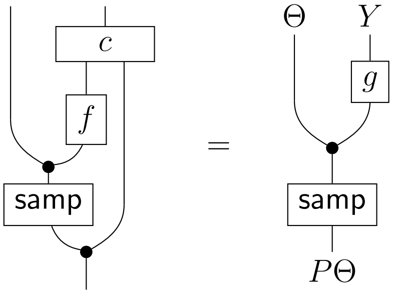
(5.38)

Consequently, the statement of Theorem 5.13 in $\mathsf{C}_{P\Theta}$ with respect to prsamp reads as follows.

Corollary 5.20. *Let C be an a.s.-compatibly representable Markov category with conditionals. Consider two morphisms $f\colon \Theta \to X$ and $g\colon \Theta \to Y$ in C with $\hat{f}_{\mathsf{prsamp}}$ and $\hat{g}_{\mathsf{prsamp}}$ given by the right hand side of equation (5.37).*

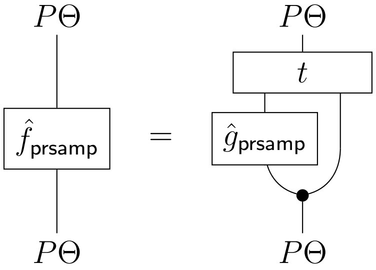
Then the following are equivalent:

1. $f \succeq_{\mathrm{Bayes}} g$, *i.e. there exists a morphism* $c \colon X \otimes P\Theta \to Y$ *satisfying*



(5.39)

2. $\hat{f}_{\mathsf{prsamp}} \sqsubseteq \hat{g}_{\mathsf{prsamp}}$, *i.e. there exists a* $\hat{g}_{\mathsf{prsamp}}$*-dilation* $t \colon P\Theta \otimes P\Theta \to P\Theta$ *satisfying*



(5.40)

Indeed, one can easily check that every $\hat{g}_{\mathsf{prsamp}}$-dilation in $\mathsf{C}_{P\Theta}$ is represented by a $\hat{g}_{\mathsf{prsamp}}$-dilation in C (and vice versa), so that the claim follows.

Instantiating this in $\mathsf{BorelStoch}$ results in the following: A statistical experiment represented by a Markov kernel f is more informative than one represented by g for *every* prior if and only if the resulting expected distribution over posteriors of the first experiment is more spread out than that of the second experiment for *every* prior, where in both parts of this statement the dependence on the prior is assumed measurable.

References

[1] Raghu Raj Bahadur. A characterization of sufficiency. *Ann. Math. Statist*, 26(2):286–293, 1955. doi:10.1214/aoms/1177728545. ↑40

[2] David Blackwell. Comparison of experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 93–102. University of California Press, 1951. projecteuclid.org/euclid.aoms/1200500222. ↑5, 37, 43

[3] David Blackwell. There are no Borel SPLIFs. *Ann. Probab.*, 8(6):1189–1190, 1980. doi:10.1214/aop/1176994581. ↑57

[4] David Blackwell and R. V. Ramamoorthi. A Bayes but not classically sufficient statistic. *Ann. Statist.*, 10(3):1025–1026, 1982. doi:10.1214/aos/1176345895. ↑55

[5] Paul Blain Levy, John Power, and Hayo Thielecke. Modelling environments in call-by-value programming languages. *Information and Computation*, 185:182–210, 2003. doi:10.1016/s0890-5401(03)00088-9. ↑4, 27

[6] Martin Brandenburg. *Tensor categorical foundations of algebraic geometry*. PhD thesis, Westfälische Wilhelms-Universität Münster, 2014. arXiv:1410.1716. ↑13, 25

[7] Francesco Buscemi. Comparison of quantum statistical models: equivalent conditions for sufficiency. *Communications in Mathematical Physics*, 310(3):625–647, 2012. doi:doi.org/10.1007/s00220-012-1421-3. ↑6

[8] Aurelio Carboni and Robert F. C. Walters. Cartesian bicategories. I. *J. Pure Appl. Algebra*, 49(1-2):11–32, 1987. doi:10.1016/0022-4049(87)90121-6. ↑8

[9] Kenta Cho and Bart Jacobs. Disintegration and Bayesian inversion via string diagrams. *Math. Structures Comput. Sci.*, 29:938–971, 2019. doi:10.1017/s0960129518000488. ↑3, 7, 8, 10

[10] Dion Coumans and Bart Jacobs. Scalars, monads and categories. In *Quantum Physics and Linguistics: A Compositional, Diagrammatic Discourse*. Oxford Academic, 2013. doi:10.1093/acprof:oso/9780199646296.003.0007. ↑14

[11] Henrique de Oliveira. Blackwell's informativeness theorem using diagrams. *Games and Economic Behavior*, 109:126–131, 2018. doi:10.1016/j.geb.2017.12.008. ↑6, 38

[12] Brendan Fong. Causal theories: A categorical perspective on Bayesian networks. Master's thesis, University of Oxford, 2012. arXiv:1301.6201. ↑7

[13] Tobias Fritz. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Adv. Math.*, 370:107239, 2020. doi:10.1016/j.aim.2020.107239. ↑3, 7, 8, 9, 10, 13, 15, 29, 40, 41, 47

[14] Tobias Fritz and Paolo Perrone. Bimonoidal structure of probability monads. *Electronic notes in theoretical computer science*, 341:121–149, 2018. doi:10.1016/j.entcs.2018.11.007. ↑15

[15] Tobias Fritz and Paolo Perrone. Monads, partial evaluations, and rewriting. *Electronic Notes in Theoretical Computer Science*, 352:129–148, 2020. doi:10.1016/j.entcs.2020.09.007. ↑34

[16] Tobias Fritz and Eigil Fjeldgren Rischel. Infinite products and zero-one laws in categorical probability. *Compositionality*, 2:3, 2020. doi:10.32408/compositionality-2-3. ↑3

[17] Carsten Führmann. Direct models for the computational lambda calculus. *Electr. Notes Theor. Comput. Sci.*, 20:245–292, 1999. doi:10.1016/S1571-0661(04)80078-1. ↑20, 21

[18] Carsten Führmann. Varieties of effects. In *Foundations of Software Science and Computation Structures*, pages 144–158, 2002. doi:10.1007/3-540-45931-6_11. ↑20

[19] Malte Gerhold, Stephanie Lachs, and Michael Schürmann. Categorial independence and lévy processes. *SIGMA. Symmetry, Integrability and Geometry: Methods and Applications*, 18:075, 2022. ↑7

[20] Peter V. Golubtsov. Monoidal Kleisli category as a background for information transformers theory. *Информационные процессы (Information Theory and Information Processing)*, 2:62–84, 2002. Translated from Russian. jip.ru/2002/GOLU1.pdf. ↑3, 7, 38

[21] Tomáš Gonda and Robert W Spekkens. Monotones in general resource theories. arXiv:1912.07085. ↑6

[22] Martin Hyland and Andrea Schalk. Abstract games for linear logic (extended abstract). In *CTCS '99: Conference on Category Theory and Computer Science*, volume 29 of *Electron. Notes Theor. Comput. Sci.*, pages Paper No. 29013, 24. Elsevier Sci. B. V., Amsterdam, 1999. doi:10.1016/s1571-0661(05)80312-3. ↑11

[23] Olav Kallenberg. *Random Measures, Theory and Applications*, volume 77 of *Probability Theory and Stochastic Modelling*. Springer, Cham, 2017. ↑9

[24] Gregory Maxwell Kelly. Doctrinal adjunction. In *Category Seminar*, volume 420 of *Lecture Notes in Mathematics*, pages 257–280. Springer, 1974. doi:10.1007/BFb0063096. ↑23

[25] Anders Kock. Commutative monads as a theory of distributions. *Theory Appl. Categ.*, 26:No. 4, 97–131, 2012. arXiv:1108.5952. ↑12

[26] Lucien Le Cam. Comparison of experiments: A short review. *Lecture Notes–Monograph Series*, pages 127–138, 1996. doi:10.1214/lnms/1215453569. ↑38, 43, 45, 52

[27] Sean Moss and Paolo Perrone. Probability monads with submonads of deterministic states. In *Proceedings of the 37th Annual ACM/IEEE Symposium on Logic in Computer Science*, pages 1–13, 2022. doi:10.1145/3531130.3533355. ↑21

[28] Arthur J. Parzygnat. Inverses, disintegrations, and bayesian inversion in quantum Markov categories, 2020. arXiv:2001.08375. ↑6

[29] Evan Patterson. *The algebra and machine representation of statistical models*. PhD thesis, Stanford University, 2020. arXiv:2006.08945. ↑3

[30] Paolo Perrone. Notes on category theory with examples from basic mathematics, 2019. arXiv:1912.10642. ↑12

[31] Emily Riehl. *Category theory in context*. Mineola, NY: Dover Publications, 2016. https://math.jhu.edu/∼eriehl/context.pdf. ↑20

[32] Erik Torgersen. *Comparison of statistical experiments*, volume 36 of *Encyclopedia of Mathematics and its Applications*. 1991. ↑40, 48