Appl Categor Struct (2014) 22:647–662
DOI 10.1007/s10485-013-9324-9

# A Categorical Foundation for Bayesian Probability

**Jared Culbertson · Kirk Sturtz**

Received: 18 June 2012 / Accepted: 11 July 2013 / Published online: 21 August 2013
© Springer Science+Business Media Dordrecht (outside the USA) 2013

**Abstract** Building on the work of Lawvere and others, we develop a categorical framework for Bayesian probability. This foundation will then allow for Bayesian representations of uncertainty to be integrated into other categorical modeling applications. The main result uses an existence theorem for regular conditional probabilities by Faden, which holds in more generality than the standard setting of Polish spaces. This more general setting is advantageous, as it allows for non-trivial decision rules (Eilenberg–Moore algebras) on finite (as well as non finite) spaces. In this way, we obtain a common framework for decision theory and Bayesian probability.

**Keywords** Bayesian probability · Decision rules · Giry monad · Regular conditional probability

**Mathematics Subject Classifications (2010)** 60A05 · 62C10

## 1 Introduction

Bayesian probability is a subject that has proven very successful in prediction, inference and model selection [3, 11, 12]. Čencov [4] gives a categorical foundation for non-Bayesian statistical inference, but a categorical framework for Bayesian probability has not been fully developed. Such a foundation is desirable in that it would allow for Bayesian methods to be tightly integrated into categorical modeling applications. The purpose of this work then, is to define the most general category

J. Culbertson (✉)
Air Force Research Laboratory, Sensors Directorate, 2241 Avionics Circle,
WPAFB, OH 45433, USA
e-mail: jared.culbertson@wpafb.af.mil

K. Sturtz
Universal Mathematics, Vandalia, OH 45377, USA
e-mail: kirksturtz@UniversalMath.com

in which it is possible to realize Bayesian probability in full. Lawvere took the first steps in this direction with the manuscript "The category of probabilistic mappings" (1962, unpublished), which gave the foundational definitions and applications of a probabilistic category. Following this, Lawvere (2011, personal communication) gave a seminar in Zurich on *Bayesian Sections*, further developing this category as a basis for Bayesian probability. The first appearance in the literature was an expansion of these ideas by Giry [10], who showed that the endofunctor on the category of measurable spaces $\mathcal{G}\colon \mathcal{M} \to \mathcal{M}$ associated to the probability adjunction given by Lawvere forms a monad, and that Lawvere's category of probabilistic mappings is the Kleisli category of that monad.

Subsequently, Meng [15], examined the category of convex sets and affine linear maps, which can be shown to be equivalent to the category of Eilenberg–Moore algebras of the Giry monad. This category can be thought of as the category of "decision rules" since the objects of that category are certain measurable functions $\mathcal{G}X \to X$ whose fibers partition the space of probability measures on a given space $X$ into positive convex measurable sets.[1] Based on the work by Giry restricting the monad to Polish spaces, Doberkat [5, 7] has since characterized the Eilenberg–Moore $\mathcal{G}$-algebras for these topological spaces. That work, however, was based upon giving the space of probability measures the $\sigma$-algebra generated by the weak topology (as used for the Polish space monad) which results in (nontrivial) finite spaces having no $\mathcal{G}$-algebras. In the final section, we show that this negative result can be circumvented by avoiding topological conditions and using the initial $\sigma$-algebra generated by the evaluation maps. Others, including Wendt [19] and Abramsky et al. [1] have also studied similar constructions.

Central to the study of Bayesian probability is the existence of regular conditional probabilities. Many textbooks restrict to Polish spaces in order to prove this existence (e.g., see [8]), though this is not a strictly necessary condition. Several more general characterizations of conditions which guarantee the existence of regular conditional probabilities have been found, either restricting the spaces involved or the joint distributions which are allowed. In [16], Pachl does not require even countably generated $\sigma$-algebras, but relies instead on a certain notion of compactness. We will prefer to follow [9], where Faden gives a necessary and sufficient condition when we restrict to countably generated spaces. Namely, the marginals of a joint distribution must give *perfect* measure spaces. The class of perfect measures is broad and includes, for example, all Radon measures. For this reason, we will only consider perfect measures and begin by defining the category of perfect probabilistic mappings. This is a subcategory of the Kleisli category of the Giry monad restricted to countably generated spaces, and is the most general category in which Bayesian probability can be realized. Note that everything prior to Section 3.2 holds without restricting to perfect probability measures (see [10]).

The main theorem (Theorem 5) states that inference maps are uniquely determined by a prior probability and a sampling distribution. This result follows from the existence of regular conditional probabilities, and we restate an existence theorem (Theorem 3) of Faden [9]. Using our characterization of Bayesian probability and the well-known fact that the Kleisli category embeds into the category of $\mathcal{G}$-algebras,

[1] Doberkat [6] refers to this process of making decision rules as *derandomization*, which in certain applications like probabilistic semantics may be a more appropriate terminology.

we can then see that the category of decision rules provides a common framework for both decision theory and Bayesian probability. Some of these ideas are similar in spirit to the general notion of distributions based on commutative monads found in the recent paper of Kock [13].

## 2 The Category of Perfect Probabilistic Mappings

We begin with an overview of the category of perfect probabilistic mappings, which is a slight modification of Lawvere's category of probabilistic mappings. The restrictions to have countably generated spaces as objects and restrict to perfect probability measures in the definition of the morphisms are required to ensure the existence of regular conditional probabilities, but are otherwise unnecessary. We include these restrictions throughout the paper to avoid overcomplicating the statements of the main theorems. See the introduction for a discussion of alternative definitions. Most of the following fundamental results are widely known without the restriction to perfect measures, and we indicate where the usual arguments must be modified due to this additional constraint. We also review the definition and foundational properties of perfect probability measures.

Let us fix the following notation. We will denote the $\sigma$-algebra of a measurable space $X$ by $\Sigma_X$, and the category of countably generated measurable spaces and measurable functions by $\mathcal{M}_{\text{cg}}$. For an object $(X, \Sigma_X)$ in $\mathcal{M}_{\text{cg}}$ we will often drop the associated $\sigma$-algebra from the notation and denote it simply by $X$ when the $\sigma$-algebra is obvious or inconsequential. We will use $(1, \Sigma_1)$ and $(2, \Sigma_2)$ for the one-element and two-element measurable sets with the discrete $\sigma$-algebras, but we will similarly just write "1" or "2" when these are used as objects in some category.

**Definition 1** A measure space $(X, \Sigma, \mu)$ is called *perfect* if for any measurable function $f\colon X \to \mathbb{R}$, there exists a Borel set $E \subset f(X)$ such that $\mu(f^{-1}(E)) = \mu(X)$. A family $\{\mu_i\}_{i\in I}$ of measures on $X$ is *equiperfect* if given $f$ as before, there exists a single Borel set $E \subset f(X)$ with $\mu_i(f^{-1}(E)) = \mu_i(X)$ for all $i \in I$.

The following theorem collects many basic results about perfect probability measures. We refer the reader to [9, 18] and the references therein for proofs and more details on perfect measures.

**Theorem 1** *Let $(X, \Sigma_X)$ and $(Y, \Sigma_Y)$ be measurable spaces. Then*

(a) *if $P$ is a {0,1}-valued probability measure, then $P$ is perfect*
(b) *if $P$ and $Q$ are probability measures on $X$ such that $P$ is perfect and $Q \ll P$, then $Q$ is perfect,*
(c) *if $\Sigma'_X \subset \Sigma_X$ and a probability measure $P$ on $X$ is perfect, then $P$ restricted to $\Sigma'_X$ is perfect,*
(d) *if $f\colon X \to Y$ is a measurable function and a probability measure $P$ on $X$ is perfect, then $f_*P$ is a perfect probability measure on $Y$,*
(e) *if $J$ is a probability measure on $(X \times Y, \Sigma_X \otimes \Sigma_Y)$ with marginals $P$ and $Q$, then $J$ is perfect if and only if $P$ and $Q$ are perfect,*
(f) *if $f\colon X \times \Sigma_Y \to [0, 1]$ is measurable for each fixed $B \in \Sigma_Y$ and $P$ is a perfect probability measure on $X$, then the probability measure on $Y$ defined by*

*$Q(B) = \int_X Q(x, B)\, dP$ is perfect if and only if $\{f(x, \cdot)\}_{x \in X}$ is an equiperfect family of probability measures on $Y$ except on a $P$-null set.*

**Definition 2** The *category of perfect probabilistic mappings* $\mathcal{P}$ has countably generated measurable spaces $(X, \Sigma_X)$ as objects and an arrow between two such objects $f\colon (X, \Sigma_X) \to (Y, \Sigma_Y)$ consists of a function $f\colon X \times \Sigma_Y \to [0, 1]$ such that

(i) for all $B \in \Sigma_Y$, the function $f(\cdot, B)\colon X \to [0, 1]$ is measurable,
(ii) the collection $\{f(x, \cdot)\colon \Sigma_Y \to [0, 1]\}_{x \in X}$ is an equiperfect family of probability measures on $Y$.

That is, morphisms in $\mathcal{P}$ can be thought of as parametrized families of perfect probability measures that vary measurably. For an arrow $f\colon (X, \Sigma_X) \to (Y, \Sigma_Y)$ we will often denote the function $f(\cdot, B)\colon X \to [0, 1]$ by $f_B$ and the function $f(x, \cdot)\colon \Sigma_Y \to [0, 1]$ by $f_x$.

Given two arrows

$$(X, \Sigma_X) \xrightarrow{f} (Y, \Sigma_Y) \xrightarrow{g} (Z, \Sigma_Z) \tag{1}$$

the composition $g \circ f\colon X \times \Sigma_Z \to [0, 1]$ is defined by

$$(g \circ f)(x, C) = \int_{y \in Y} g_C(y) df_x. \tag{2}$$

This composition is well-defined due to Theorem 1 (f), and associativity follows easily from the monotone convergence theorem. An important fact is that every measurable function $f\colon X \to Y$ may be regarded as a $\mathcal{P}$-morphism $\delta_f\colon X \to Y$, where the Dirac (or one point) measure

$$\delta_f(x, B) = \begin{cases} 1 & \text{if } f(x) \in B \\ 0 & \text{if } f(x) \notin B \end{cases} \tag{3}$$

assigns to each $x \in X$ the Dirac measure on $Y$ which is concentrated at $f(x)$. Taking the measurable function $f$ to be the identity map on a particular measurable space $X$ gives the arrow $\delta_{Id_X}\colon (X, \Sigma_X) \to (X, \Sigma_X)$, i.e., the identity arrow for $X$ in $\mathcal{P}$. In fact, it is easy to check that the association $f \mapsto \delta_f$ determines a functor $\delta\colon \mathcal{M}_{\text{cg}} \to \mathcal{P}$ taking a measurable space to itself. Note that this functor is not faithful, however, and so we do not get an embedding of $\mathcal{M}_{\text{cg}}$ into $\mathcal{P}$. We will call a $\mathcal{P}$ arrow $P\colon X \to Y$ *deterministic* if for every $B \in \Sigma_Y$ the measurable functions $P_B\colon X \to [0, 1]$ assume only the values 0 or 1. In fact, every deterministic $\mathcal{P}$-arrow $P\colon X \to Y$ is of the form $P = \delta_f$ for some measurable function $f$, provided that $X$ has cardinality below the cardinality of the set of all measurable sets.

The following lemma gives two useful properties which follow easily from standard exercises in measure theory and the definition of composition in $\mathcal{P}$.

**Lemma 1** *If $p\colon X \to Y$ is a measurable function and $f\colon Y \to Z$ a $\mathcal{P}$-morphism, then the composition*

$$X \xrightarrow{\quad\delta_p\quad} Y \xrightarrow{\quad\quad f\quad\quad} Z$$

*is given by* $(f \circ \delta_p)(x, C) = f_{p(x)}(C)$. *On the other hand, if* $q\colon Y \to Z$ *is a measurable function and* $g\colon X \to Y$ *a* $\mathcal{P}$*-morphism, then the composition*

$$X \xrightarrow{\quad g \quad} Y \xrightarrow{\quad \delta_q \quad} Z$$

*is given by* $(\delta_q \circ g)(x, C) = g_x(q^{-1}C)$.

There are several distinguished objects in $\mathcal{P}$ that play an important role in many constructions. Any set $X$ with the indiscrete $\sigma$-algebra $\Sigma_X = \{X, \emptyset\}$ is a terminal object since any arrow $P\colon Y \to X$ is completely determined by the fact that $P_y$ must be a probability measure on $X$. We denote the canonical terminal object by 1 since it is isomorphic to the one-element set. Notice that an arrow $P\colon 1 \to X$ is precisely a perfect probability measure on $X$ and that 1 is a separator for $\mathcal{P}$. In addition to having a separator, the category $\mathcal{P}$ also has a coseparator, the two-element set $2 = \{\top, \bot\}$ with $\Sigma_2$ the discrete algebra on 2. Moreover, there is a set bijection $\hom_{\mathcal{P}}(X, 2) \simeq \hom_{\mathcal{M}_{\text{cg}}}(X, [0, 1])$.

To see how the category $\mathcal{P}$ relates to the Giry monad $\mathcal{G}\colon \mathcal{M}_{\text{cg}} \to \mathcal{M}_{\text{cg}}$ and its Kleisli category $K(\mathcal{G})$, we briefly recall the construction of the monad. Let $\mathcal{P}X$ denote the set of (not necessarily perfect) probability measures on $X$. We endow $\mathcal{P}X$ with the coarsest $\sigma$-algebra such that the evaluation maps $ev_B\colon \mathcal{P}X \to [0, 1]$ given by $ev_B(P) = P(B)$ are measurable. Then we can define a functor $\mathcal{P}\colon K(\mathcal{G}) \to \mathcal{M}_{\text{cg}}$ which sends a measurable space $X$ to the space $\mathcal{P}X$ of probability measures on $X$. On arrows, $\mathcal{P}$ sends the $K(\mathcal{G})$-arrow $f\colon X \to Y$ to the measurable function $\mathcal{P}f\colon \mathcal{P}X \to \mathcal{P}Y$ defined pointwise on $\Sigma_Y$ by

$$\mathcal{P}f(P)(B) = \int_X f_B \, dP. \tag{4}$$

That is, $\mathcal{P}f(P)$ gives the probability measure on $Y$ defined by the composition

$$1 \xrightarrow{\quad P \quad} X \xrightarrow{\quad f \quad} Y \qquad \text{with } f \circ P \text{ the composite} \tag{5}$$

in $K(\mathcal{G})$.[2] Since $\mathcal{P}X = \hom_{\mathcal{P}}(1, X)$ as sets, another common notation for $\mathcal{P}X$ is $X^1$, but we will use the functor notation for clarity.

Based on the definition of the $\sigma$-algebra on $\mathcal{P}X$, we can define the evaluation morphism $\varepsilon_X\colon \mathcal{P}X \to X$ by $\varepsilon_X(P, A) = P(A)$. With this, we are able to explicitly state the relationship between $\mathcal{M}_{\text{cg}}$ and $K(\mathcal{G})$, first proved in [10] for general measurable spaces and probability measures.

**Theorem 2** *The functors* $\delta\colon \mathcal{M}_{\text{cg}} \to K(\mathcal{G})$ *and* $\mathcal{P}\colon K(\mathcal{G}) \to \mathcal{M}_{\text{cg}}$ *form an adjunction*

$$\mathcal{M}_{\text{cg}} \underset{\mathcal{P}}{\overset{\delta}{\rightleftarrows}} \mathcal{P} \qquad\qquad \delta \dashv \mathcal{P}$$

*with the unit of the adjunction* $\eta_X(x) = \delta_{\{x\}}$ *and the counit* $\varepsilon_X\colon \mathcal{P}X \to X$.

[2] Ramachandran showed in [17] that even if $P$ is perfect and all of the $f_x$ are perfect, the composition $f \circ P$ is not necessarily a perfect measure on $Y$. This is precisely why we must restrict to equiperfect families in the definition of $\mathcal{P}$.

Thus, we can realize the Giry monad as the composition $\mathcal{G} = \mathcal{P} \circ \delta$, and moreover, $K(\mathcal{G})$ is equivalent to the smallest category through which $T$ factors. Hence every $K(\mathcal{G})$ arrow $P\colon X \to Y$ corresponds uniquely to a measurable arrow $X \to TY$. From this, we can see that $\mathcal{P}$ is a subcategory of $K(\mathcal{G})$, where $\mathcal{P}$ precisely does not include the $K(\mathcal{G})$-morphisms which do not give equiperfect families of probability measures. On the other hand, since every probability measure on a Polish space is perfect, we see that $\mathcal{P}$ includes the Kleisli category of the Giry monad restricted to Polish spaces. So in general we have

$$K(\mathcal{G}_{\text{Pol}}) \hookrightarrow \mathcal{P} \hookrightarrow K(\mathcal{G}_{\text{cg}}),$$

where the subscripts indicate the respective restrictions of the Giry monad.

## 3 Joint Distributions and Conditionals

Given a family of objects $\{X_i\}_{i\in I}$ we can form the cartesian product $\prod_{i\in I} X_i$ and endow this set with the product $\sigma$-algebra generated by all the projection maps $\prod_{i\in I} X_i \xrightarrow{\pi_j} X_j$, one for each index $j \in I$. It is easy to see that

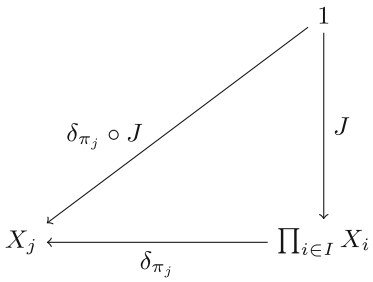
$$\left(\left(\prod_{i\in I} X_i, \bigotimes_{i\in I} \Sigma_{X_i}\right), \{\delta_{\pi_i}\}_{i\in I}\right) \tag{6}$$

does not give a categorical product. In fact, only weak products and equalizers exist in $\mathcal{P}$, as the uniqueness condition fails for both constructions. We use the terminology "product space" to denote the set product of any family $\{(X_i, \Sigma_{X_i})\}_{i\in I}$ of objects with the product $\sigma$-algebra and not to imply that that object

$$\left(\prod_{i\in I} X_i, \bigotimes_{i\in I} \Sigma_{X_i}\right)$$

with projections satisfies any universality condition. We will call a probability measure $J\colon 1 \to (\prod_{i\in I} X_i, \otimes_{i\in I}\Sigma_{X_i})$ on a product space a *joint distribution*. We do not mean to imply that these are distributions of a random variable, but rather indicate a measure on a product space which is not necessarily a product measure. These joint distributions are the main objects of study in Bayesian probability.

Given any joint distribution $J\colon 1 \to \prod_{i\in I} X_i$, for each $j \in I$ we have the diagram



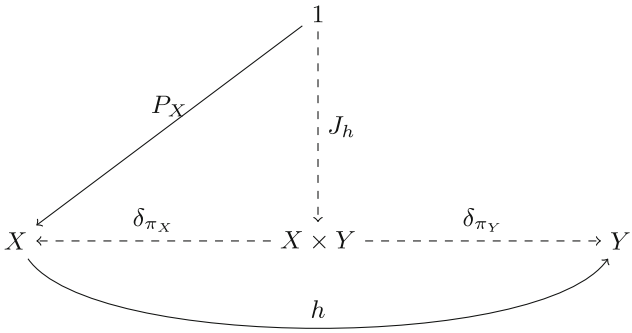

(7)

where the composite $\delta_{\pi_j} \circ J$ is called the *marginal* (distribution) of $J$ on component $X_j$ and is given by $(\delta_{\pi_j} \circ J)(A_j) = J(\pi_j^{-1} A_j)$ by Lemma 1.

Given only the probability measures on the components, $\{P_i\}_{i\in I}$, there are many joint distributions on the product space whose marginals are the given family $\{P_i\}_{i\in I}$. By using *relationships* in the form of conditionals between the components, we bring into play additional knowledge that permits the determination of the appropriate joint distribution. If the uncertainty of component $X_j$, as expressed by a probability measure $P_j$ on component $X_j$, depends conditionally on a parameter which varies over component $X_i$ then we have the $\mathcal{P}$ arrow $h\colon X_i \to X_j$. These conditionals—which are the morphisms in $\mathcal{P}$—are the key to determining a unique joint distribution. The relationship between the components $X_i$ and $X_j$ is mediated by the conditional $h$ and expresses the relationship $P_j = h \circ P_i$.

### 3.1 Constructing a Joint Distribution Given Conditionals

We now show how marginals and conditionals can be used to determine joint distributions in $\mathcal{P}$. This development follows that of [1] where we first learned of this approach (the category $\mathcal{S}toch$ of stochastic kernels in that paper is nearly what we call $\mathcal{P}$, although our restriction to countably generated spaces and perfect measures is replaced in the paper by restricting to Polish spaces). Given a $\mathcal{P}$-arrow $h\colon X \to Y$ and a perfect probability measure $P_X\colon 1 \to X$ on $X$, consider the diagram




(8)

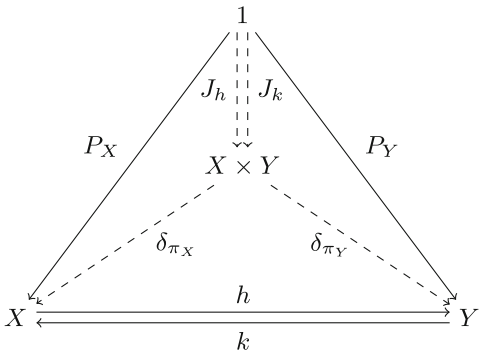
where $J_h$ is the uniquely determined joint distribution on the product space $X \times Y$ defined on the rectangles of the $\sigma$-algebra $\Sigma_X \otimes \Sigma_Y$ by

$$J_h(A \times B) = \int_A h_B \, dP_X. \tag{9}$$

Then it follows from Theorem 1 that $J_h$ is perfect. The marginal of $J_h$ with respect to $Y$ then satisfies $\delta_{\pi_Y} \circ J_h = h \circ P_X$ and the marginal of $J_h$ with respect to $X$ is $P_X$, both of which are also perfect. By a symmetric argument, if we are given a probability measure $P_Y$ and conditional probability $k\colon Y \to X$ then we obtain a unique perfect joint distribution $J_k$ on the product space $X \times Y$ given on the rectangles by

$$J_k(A \times B) = \int_B k_A \, dQ. \tag{10}$$

However, if we are given $P_X$, $P_Y$, $h$, $k$ as indicated in the diagram




(11)

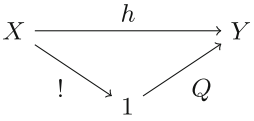
then we have that $J_h = J_k$ if and only if the compatibility conditions

$$\begin{aligned} P_X &= k \circ P_Y \\ P_Y &= h \circ P_X \end{aligned} \tag{12}$$

are satisfied. Thus if the compatibility conditions are satisfied, then we can realize the product rule of probability in $\mathcal{P}$ as

$$\int_A h_B \, dP_X = J(A \times B) = \int_B k_A \, dP_Y. \tag{13}$$

In the extreme case, suppose we have a $\mathcal{P}$-arrow $h\colon X \to Y$ which factors through the terminal object 1 as




(14)

where ! represents the unique arrow from $X \to 1$. If we are also given a perfect probability measure $P\colon 1 \to X$, then we can calculate the joint distribution determined by $P$ and $h = Q \circ !$ as

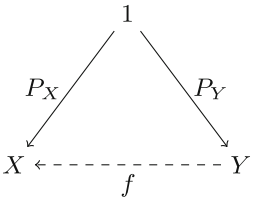
$$\begin{aligned} J(A \times B) &= \int_A (Q \circ !)_B \, dP \\ &= P(A) \cdot Q(B) \end{aligned} \tag{15}$$

so that $J = P \otimes Q$. This is precisely the situation where we say that the marginals $P$ and $Q$ are *independent*. Thus in $\mathcal{P}$ independence corresponds to a special instance of a conditional—one that factors through the terminal object.

### 3.2 Constructing Regular Conditionals Given a Joint Distribution

The following result is the basis from which the inference maps in Bayesian probability theory are subsequently constructed. The proof of the following theorem can be found in [9], where several equivalent conditions are identified.

**Theorem 3** *If $J : 1 \to X \times Y$ is a $\mathcal{P}$-morphism with marginals $P_X$ on $X$ and $P_Y$ on $Y$, then there exists a $\mathcal{P}$-arrow $f$ that makes the diagram*




(16)

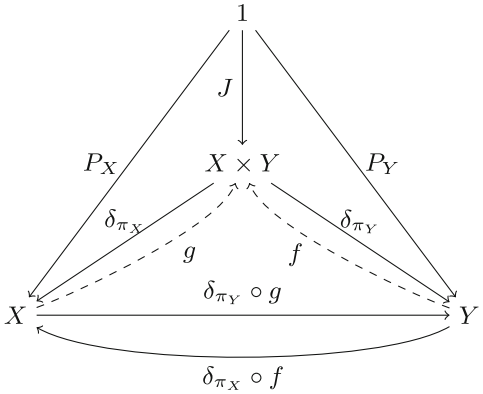
*commute and satisfies*

$$J(A \times B) = \int_B f_A \, dP_Y. \tag{17}$$

*Moreover, the morphism $f$ is the unique $\mathcal{P}$-morphism with these properties, up to a set of $P_Y$-measure zero.*

Interestingly, we can use Theorem 3 to obtain a seemingly stronger statement, i.e., that the regular conditional probability factors through the product. Though this is not difficult to prove, we will prefer this stronger statement in the sequel.

**Theorem 4** *If $J : 1 \to X \times Y$ is a $\mathcal{P}$-morphism with marginal distributions $P_X$ and $P_Y$ on $X$ and $Y$, then there exist $\mathcal{P}$ arrows $f$ and $g$ such that the diagram*




(18)

*commutes and*

$$\int_A (\delta_{\pi_Y} \circ g)_B \, dP_X = J(A \times B) = \int_B (\delta_{\pi_X} \circ f)_A \, dP_Y. \tag{19}$$

*Proof* We can apply Theorem 3 to see that there is a $\mathcal{P}$-arrow $f\colon Y \to X \times Y$ satisfying $J = f \circ P_Y$ such that

$$\int_C f_{A \times B} \, dP_Y = J\left(A \times (B \cap C)\right). \tag{20}$$

Then from Lemma 1, we know that $(\delta_{\pi_X} \circ f)(y, A) = f(y, A \times Y)$ and so

$$\int_B (\delta_{\pi_X} \circ f)_A \, dP_Y = \int_B f_{A\times Y} \, dP_Y \tag{21}$$

$$= J(A \times (Y \cap B)) \tag{22}$$

$$= J(A \times B) \tag{23}$$

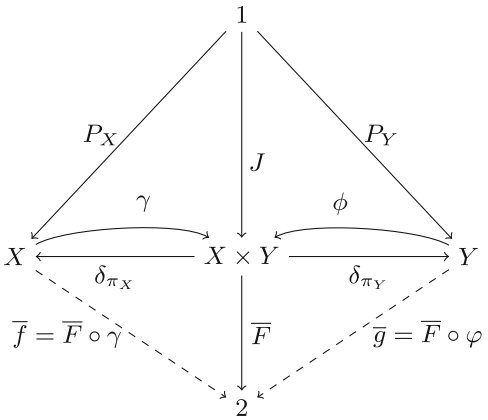
Similarly we obtain a $\mathcal{P}$-arrow $g\colon X \to X \times Y$ satisfying $J = g \circ P_X$ and

$$\int_A (\delta_{\pi_Y} \circ g)_B \, dP_X = J(A \times B). \tag{24}$$

With these facts, it is a simple exercise to check that the diagram commutes.

Note that if the joint distribution $J$ is obtained by a probability measure $P_X$ and a conditional $h\colon X \to Y$ using the method described by Diagram 8, then using the above result and notation it follows $P_X$-a.s. that $h = \delta_{\pi_Y} \circ g$.

*Remark 1* (Tonneli's Theorem) Given a $\mathcal{P}$-morphism $J : 1 \to X \times Y$, with marginals $P_X$ and $P_Y$ let $\gamma\colon X \to X \times Y$ and $\varphi\colon Y \to X \times Y$ be the $\mathcal{P}$ arrows satisfying $\gamma \circ P_X = J$ and $\varphi \circ P_Y = J$ whose existence is guaranteed by Theorem 4. Given any measurable function $F\colon X \times Y \to [0, 1]$ we have the diagram




(25)

where the top two triangles commute. Thus we can define $\overline{f} = \overline{F} \circ \gamma$ and $\overline{g} = \overline{F} \circ \varphi$ so that the entire diagram commutes. From this, it follows that

$$\int_X f \, dP_X = \int_{X\times Y} F \, dJ = \int_Y g \, dP_Y \tag{26}$$

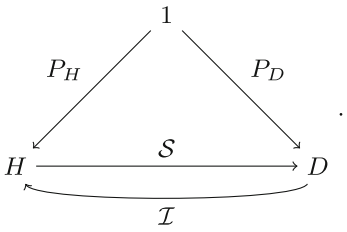
and we can realize Tonneli's Theorem as the special case with $J = P_X \otimes P_Y$.

This formulation provides the context for the following optimal transportation problem: given marginals $P_X$ and $P_Y$ that model the supply and demand constraints, and a cost function $F$ (defined up to a scalar constant) representing the unit cost to transport a product from $x \in X$ to $y \in Y$, what joint distribution $J$ on $X \times Y$ with marginals $P_X$ and $P_Y$ minimizes the objective function $\int_{X\times Y} F \, dJ$? The optimal assignment is then the conditional probability $X \to Y$ determined by the optimal

joint distribution $J$. For example, this problem is investigated and a unique solution is given in certain cases in [14].

## 4 Bayesian Probability in $\mathcal{P}$

If we replace $X$ and $Y$ in Diagram 18 by $D$(ata) and $H$(ypotheses), and the composites $\delta_{\pi_Y} \circ g$ by $\mathcal{S}$(ampling distribution) and $\delta_{\pi_X} \circ f$ by $\mathcal{I}$(nference), then we can define $P_D = \mathcal{S} \circ P_H$ to obtain




(27)

In the context of Bayesian probability the probability measure $P_H$ is often called a *prior probability*.

In this notation, the product rule in $\mathcal{P}$ given in Eq. 13 becomes

$$\int_{\mathcal{D}} \mathcal{I}_{\mathcal{H}} \, dP_D = J(\mathcal{H} \times \mathcal{D}) = \int_{\mathcal{H}} \mathcal{S}_{\mathcal{D}} \, dP_H \tag{28}$$

where $\mathcal{H} \in \Sigma_H$ and $\mathcal{D} \in \Sigma_D$. We will spend the remainder of this sections showing how this interpretation of these spaces in $\mathcal{P}$ provides a categorical foundation for Bayesian probability. First, we briefly review the fundamental concepts in Bayesian probability theory which can be found in [12] and then proceed to show how $\mathcal{P}$ is the appropriate category for this theory. Generally, a *Bayesian model* is comprised of a number of items including

(i) two measurable spaces $H$ and $D$ representing hypotheses and data, respectively,
(ii) a probability measure $P_H$ on the $H$ space called the prior probability,
(iii) a $\mathcal{P}$ arrow $\mathcal{S}\colon H \to D$ called the sampling distribution,
(iv) a $\mathcal{P}$ arrow $\mathcal{I}\colon D \to H$ called the inference map,

Note that the data space $D$ can also be thought of as the event space for some experiment, and the $\sigma$-algebra on $D$ is determined by distinguishable data. The prior probability measures are updated via the inference map as one takes measurements, which correspond to probability measures $\mu$ on $D$. These updated probability measures are then called *posterior probabilities* and are given by $\widehat{P}_H = \mathcal{I} \circ \mu$. The posterior $\widehat{P}_H$ then becomes the prior probability for the next step and the process continues as more measurements are taken. At each step in a Bayesian process, the posterior probability is a representation of knowledge about the hypotheses based on all of the data that has been accumulated up to that point.

Using the sampling distribution $\mathcal{S}$ and the prior probability $P_H$ on $H$, we can define a joint distribution $J$ on the product space $H \times D$ as in Section 3.1 by defining it on the rectangles as $J(A \times B) = \int_A \mathcal{S}_B \, dP_H$. The $D$-marginal (prior probability

on data) is then $P_D = \mathcal{S} \circ P_H = \delta_{\pi_D} \circ J$. Using Theorem 4, we have the following theorem.

**Theorem 5** *Given $\mathcal{P}$ arrows representing a prior probability $P_H \colon 1 \to H$ and a sampling distribution $\mathcal{S} \colon H \to D$, the inference map $\mathcal{I} \colon D \to H$ is determined uniquely up to a set of $P_D$-measure zero.*

*Proof* Let $\mathcal{I} \colon D \to H$ be the composition $\delta_{\pi_H} \circ f$, where $\delta_{\pi_H}$ is the projection $H \times D \to H$ and $f \colon D \to H \times D$ is the $\mathcal{P}$-arrow satisfying

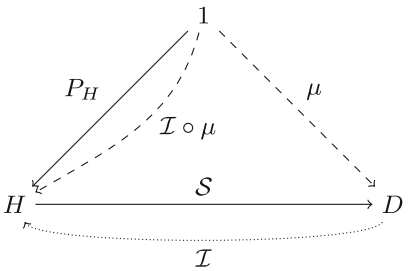
$$J(U \times V) = \int_D f_{U \times V} \, dP_D \tag{29}$$

whose existence is given by Theorem 4. Thus

$$\int_A \mathcal{S}_B \, dP_H = J(A \times B) = \int_B \mathcal{I}_A \, dP_D \tag{30}$$

and this inference arrow $\mathcal{I}$ is unique in that if $\mathcal{I}'$ also satisfies Eq. 30 then the set $\{y \in Y \mid \mathcal{I}_y \neq \mathcal{I}'_y\}$ has $P_D$-measure zero.

Thus the complete process works in the following way. A prior probability $P_H$ and sampling distribution $\mathcal{S}$ are specified, from which one determines the inference map $\mathcal{I}$. Once measurements $\mu \colon 1 \to D$ are taken, we then calculate the posterior probability by $\mathcal{I} \circ \mu$. This updating procedure can be characterized by the diagram




(31)

where the solid lines indicate arrows given a priori, the dotted line indicates the arrow determined using Theorem 4, and the dashed lines indicate the updating after a measurement. Note that if there is no uncertainty in the measurement, then $\mu = \delta_{\{x\}}$ for some $x \in D$, but in practice there is usually some uncertainty in the measurements themselves.

Following the calculation of the posterior probability, the sampling distribution is then updated, if required. The process can then repeat: using the posterior probability and the updated sampling distribution the updated joint probability distribution on the product space is determined and the corresponding (updated) inference map determined. We can then continue to iterate as long as new measurements are received. For some problems (such as with the standard urn problem with replacement of balls) the sampling distribution does not change from iterate to iterate, but the inference map is updated since the posterior probability on the hypothesis space changes with each measurement. The model selection problem (either once at the beginning of this process, or iteratively throughout) can also be modeled as a meta-Bayesian process, where the hypothesis space is the space of potential models and

the data constitutes some information that would inform on the suitability of a given model.

*Remark 2* We know from Theorem 5 that the inference map $\mathcal{I}$ is uniquely determined by $P_H$ and $\mathcal{S}$ up to a set of $P_D$-measure zero. However, there is no reason a priori that a measurement $\mu\colon 1 \to D$ is required to be absolutely continuous with respect to $P_D$. In $\mu$ is not absolutely continuous with respect to $P_D$, then a different choice of inference map $\mathcal{I}'$ could yield a different posterior probability—i.e., we could have $\mathcal{I} \circ \mu \neq \mathcal{I}' \circ \mu$. Thus we make the assumption that measurement probabilities on $D$ are absolutely continuous with respect to the prior probability $P_D$ on $D$. This is a reasonable assumption, however, since if a data event is impossible (has $P_D$-measure zero) under a certain model, then the model should not be expected to make an meaningful inference when presented with that data. On the other hand, it is easy to see that if a measurement $\mu \ll P_D$, then $\mathcal{I} \circ \mu \ll P_H$, as expected.

We emphasize that this procedure can be employed for any perfect prior probability and any regular conditional probability. For example, given a perfect prior $P\colon 1 \to \mathcal{P}X$ and the conditional $\varepsilon_X\colon \mathcal{P}X \to X$ there corresponds a unique inference map $\mathcal{I}\colon X \to \mathcal{P}X$ satisfying, for all $A \in \Sigma_X$ and for all $\mathcal{B} \in \Sigma_{\mathcal{P}X}$,

$$\int_{\mathcal{P}X} \varepsilon_X{}_A \, dP = \int_X \mathcal{I}_{\mathcal{B}} \, d(\varepsilon_X \circ P). \tag{32}$$

In the case where $X = 2 = \{\top, \bot\}$ (the two element set), these "higher order distributions" $P\colon 1 \to \mathcal{P}2$ can be used to explicate the concept of $A_p$ distributions as characterized by Jaynes [12, Chapter 18]. Using our notation, a proposition $A$, which is a morphism $A\colon 1 \to 2$ in the category $\mathcal{S}et$ of sets, has an associated probability of truth, say $Pr(A) = p$. Hence $A$ determines a $\mathcal{P}$-morphism $\overline{A}\colon 1 \to 2$, with $\overline{A}(\{\top\}) = p$. The information supplied by the arrow $\overline{A}$ consists only of the single value $p$ and fails to indicate how sensitive this proposition is to additional data. The confidence that one has in the value $p$ can be supplied by the higher order distributions which are probability measures on the space $\mathcal{P}2$ of probability measures on 2. Since $\mathcal{P}2$ consists precisely of the Bernoulli distributions $B_\theta = \theta\delta_\top + (1-\theta)\delta_\bot$, where $\theta \in [0, 1]$, it follows that $\mathcal{P}2 \simeq [0, 1]$. Consequently, any distribution

$$1 \xrightarrow{A_p} \mathcal{P}2 \tag{33}$$

has an expected value which can be calculated using the composition

$$1 \xrightarrow{A_p} \mathcal{P}2 \xrightarrow{\varepsilon_2} 2. \tag{34}$$

Thus $E(A_p) = (\varepsilon_2 \circ A_p)(\{\top\}) = p$ and any such distribution provides a more informative measure. For example, the two distributions on $\mathcal{P}2 \simeq [0, 1]$ specified by $p = \delta_{\frac{1}{2}}$ and $p'$ the uniform (Lebesque) measure both have expected value $\frac{1}{2}$. Yet clearly, the first is deterministic, expressing a (complete) confidence in the statement that the expected value of the proposition

$$1 \xrightarrow{\overline{A}} 2 \tag{35}$$

where $\overline{A}(\top) = \frac{1}{2}$ is $\frac{1}{2}$. On the other hand, the distribution $p'$ also determines $\overline{A}$, but instead expresses a maximal ignorance modeled by the uniform distribution.

## 5 The Category of Decision Rules

Recall that the Giry monad $\mathcal{G}$ factors through $K(\mathcal{G})$ via the adjunction of Theorem 2. At the other end of the spectrum of categories through which the Giry monad factors is the Eilenberg–Moore category $\mathcal{M}_{\text{cg}}^{\mathcal{G}}$, consisting of the Eilenberg–Moore algebras of the Giry monad. From the theory of monads (see [2], for example), we know that $K(\mathcal{G})$ then embeds into $\mathcal{M}_{\text{cg}}^{\mathcal{G}}$, which has additional structure that is useful for dealing with other aspects of probability theory and decision making which $\mathcal{P}$ and $K(\mathcal{G})$ are not equipped for. So we have $\mathcal{P} \hookrightarrow K(\mathcal{G}) \hookrightarrow \mathcal{M}_{\text{cg}}^{\mathcal{G}}$. Let us briefly recall the definition of a $T$-algebra for a general monad $T$.

If $(T, \eta, \mu)$ is a monad in a category $\mathcal{C}$, a *$T$-algebra* $(X, \alpha)$ is a pair consisting of an object $X$ in $\mathcal{C}$ and a $\mathcal{C}$ arrow $\alpha\colon TX \to X$ such that the diagrams

$$\begin{array}{ccc} T^2X & \xrightarrow{\mu_X} & TX \\ {\scriptstyle T\alpha}\downarrow & & \downarrow{\scriptstyle \alpha} \\ TX & \xrightarrow[\alpha]{} & X \end{array} \qquad \begin{array}{ccc} X & \xrightarrow{\eta_X} & TX \\ & {\scriptstyle Id}\searrow & \downarrow{\scriptstyle \alpha} \\ & & X \end{array} \tag{36}$$

commute; the first diagram is called the associative law and the second diagram the unit law. A morphism of $T$-algebras $(X, \alpha) \xrightarrow{f} (Y, \beta)$ is an arrow $f\colon X \to Y$ of $\mathcal{C}$ such that the diagram

$$\begin{array}{ccc} TX & \xrightarrow{Tf} & TY \\ {\scriptstyle \alpha}\downarrow & & \downarrow{\scriptstyle \beta} \\ X & \xrightarrow[f]{} & Y \end{array} \tag{37}$$

commutes.

For $\mathcal{G}$, an algebra $\mathcal{G}X \xrightarrow{\alpha} X$ consists of a measurable space $X$ with a countably generated $\sigma$-algebra, the space $\mathcal{G}X = \mathcal{P}X$ of probability measures on $X$, and a measurable map $\alpha$ satisfying the two defining properties of a $\mathcal{G}$-algebra. The measurable map $\mathcal{G}f$ in the definition of a morphism of $\mathcal{G}$-algebras is the pushforward map: given $P \in \mathcal{G}X$ the pushforward by $f$ is $\mathcal{G}f(P) = f_*P \in \mathcal{P}Y$.

The $\mathcal{G}$-algebras $(X, \alpha)$ are often called decision rules since the measurable map $\alpha$ assigns (decides) a value in $X$ to each probability measure $P$ on $X$. Alternatively, we can think of a decision rule as collapsing a probability distribution to a definite value, or *derandomizing* a probability distribution as in [6]. For this reason, we often use the descriptive characterization of Čencov [4] and call the category $\mathcal{M}_{\text{cg}}^{\mathcal{G}}$ the *category of decision rules*.[3]

[3] Čencov did not work in $\mathcal{M}_{\text{cg}}^{\mathcal{G}}$, but rather in $K(\mathcal{G})$, restricting to measurable spaces $X$ such that $\mathcal{P}X$ has a $\sigma$-algebra generated by finitely many atoms. The primary difference in the current approach and that of Čencov is that we take a Bayesian viewpoint, while he is attempting to describe the standard statistical inference perspective.

One of the primary advantages to using the existence of regular conditional probabilities guaranteed by Theorem 4 is that we only require that measurable spaces have countably generated $\sigma$-algebras and that the morphisms give equiperfect families. In contrast, much of the previous work involving the Giry monad and regular conditional probabilities requires resorting to topological arguments and restricting to Polish spaces. For example, in [5], Doberkat characterizes the $\mathcal{G}$-algebras for the Giry monad under the Polish space assumption, and proves the counter-intuitive result that there are no non-trivial decision rules for finite spaces in $\mathcal{M}_{\text{cg}}^{\mathcal{G}}$. In contrast, we exhibit a finite space having a $\mathcal{G}$-algebra when one does not require topological restrictions.

*Example 1* An important such case is the decision rule $d\colon \mathcal{G}2 \to 2$ given by

$$d(P) = \begin{cases} \top & \text{if } P(\{\top\}) = 1 \\ \bot & \text{if } P(\{\top\}) < 1. \end{cases} \tag{38}$$

The function $d$ is measurable since $d^{-1}(\{\top\}) = \{\delta_\top\} \in \Sigma_{\mathcal{G}2}$. The associativity identity

$$\begin{array}{ccc} \mathcal{G}^2 2 & \xrightarrow{\mathcal{G}d} & \mathcal{G}2 \\ {\scriptstyle \mu_2}\downarrow & & \downarrow{\scriptstyle d} \\ \mathcal{G}2 & \xrightarrow[d]{} & 2 \end{array} \qquad \begin{array}{ccc} Q & \xrightarrow{\mathcal{G}d} & Qd^{-1} \\ {\scriptstyle \mu_2}\downarrow & & \downarrow{\scriptstyle d} \\ \mu_2(Q) & \xrightarrow[d]{} & d(\mu_2(Q)) = d(Qd^{-1}) \end{array} \tag{39}$$

where $\mu$ is the monad multiplication defined by

$$\mu_2(Q)(A) = \int_{q \in \mathcal{G}(2)} ev_A(q)\, dQ, \tag{40}$$

is satisfied since both routes map the element $Q \neq \delta_{\delta_\top} \in \mathcal{G}^2(2) \mapsto \bot$ while $\delta_{\delta_\top} \mapsto \top$. The unit law $Id_2 = d \circ \eta_2$ is trivial to verify.

The decision rule $d\colon \mathcal{G}2 \to 2$ partitions the space $\mathcal{G}2$ into $\delta_\top$ and all measures on 2 whose value on $\{\top\}$ is of measure less than one. There are many other decision rules for 2, and any other finite or nonfinite space. Characterizing decision rules without the requirement for continuity is an open problem.

**Acknowledgements** The authors would like to express gratitude to F.W. Lawvere, P.F. Stiller and T. Nguyen for many fruitful conversations regarding these ideas. We also thank the reviewer for the helpful comments provided that have led to a much clearer exposition. This work was partially supported by the Air Force Office of Scientific Research, for which the authors are extremely grateful.

## References

1. Abramsky, S., Blute, R., Panangaden, P.: Nuclear and trace ideals in tensored $*$-categories. J. Pure Appl. Algebra **143**, 3–47 (1999)
2. Barr, M., Wells, C.: Toposes, triples and theories. Repr. Theory Appl. Categ. (12) (2005). Corrected reprint of the 1985 original (Springer-Verlag)

3. Berger, J.: Statistical Decision Theory and Bayesian Analysis, 2nd edn. Springer Series in Statistics. Springer, New York (1985)
4. Čencov, N.: Statistical decision rules and optimal inference. In: Translations of Mathematical Monographs, vol. 53. American Mathematical Society (1982)
5. Doberkat, E.E.: Characterizing the Eilenberg–Moore algebras for a monad of stochastic relations. Internal Memorandum No. 147 (2004)
6. Doberkat, E.E.: Derandomizing probabilistic semantics through Eilenberg–moore algebras for the Giry monad. Internal Memorandum No. 149 (2004)
7. Doberkat, E.E.: Kleisli morphisms and randomized congruences for the giry monad. J. Pure Appl. Algebra **211**(3), 638–664 (2007)
8. Dudley, R.: Real analysis and probability. Cambridge Studies in Advanced Mathematics, vol. 74. Cambridge University Press, Cambridge (2002)
9. Faden, A.M.: The existence of regular conditional probabilities: necessary and sufficient conditions. Ann. Probab. **13**(1), 288–298 (1985)
10. Giry, M.: A categorical approach to probability theory. In: Categorical Aspects of Topology and Analysis (Ottawa, Ont., 1980). Lecture Notes in Math., vol. 915, pp. 68–85. Springer, Berlin-New York (1982)
11. Howson, C., Urbach, P.: Scientific Reasoning: The Bayesian Approach, 2nd edn. Open Court Publishing, Chicago (1993)
12. Jaynes, E.T.: Probability Theory: The Logic of Science. Cambridge University Press, Cambridge (2003)
13. Kock, A.: Commutative monads as a theory of distributions. Theory Appl. Categ. **26**, 97–131 (2012)
14. Korman, J., McCann, R.: Optimal transportation with capacity constraints (2012). arXiv:1201.6404v2
15. Meng, X.: Categories of convex sets and of metric spaces, with applications to stochastic programming and related areas. Ph.D. thesis, State University of New York at Buffalo (1988)
16. Pachl, J.K.: Disintegration and compact measures. Math. Scand. **43**, 157–168 (1978)
17. Ramachandran, D.: Perfect mixtures of perfect measures. Ann. Probab. **7**, 444–452 (1979)
18. Rodine, R.H.: Perfect probability measures and regular conditional probabilities. Ann. Math. Stat. **37**, 1273–1278 (1966)
19. Wendt, M.: The category of disintegrations. Cah. Topol. Géom. Différ. Catég. **35**(4), 291–308 (1994)