A categorical characterization of relative entropy on standard Borel spaces

Nicolas Gagné[1,2]

School of Computer Science
McGill University
Montréal, Québec, Canada

Prakash Panangaden[1,3]

School of Computer Science
McGill University
Montréal, Québec, Canada

Abstract

We give a categorical treatment, in the spirit of Baez and Fritz, of relative entropy for probability distributions defined on standard Borel spaces. We define a category called **SbStat** suitable for reasoning about statistical inference on standard Borel spaces. We define relative entropy as a functor into Lawvere's category $[0, \infty]$ and we show convexity, lower semicontinuity and uniqueness.

Keywords: Entropy, Gìry Monad, Bayesian Learning, standard Borel spaces

1 Introduction

The inspiration for the present work comes from two recent developments. The first is the beginning of a categorical understanding of Bayesian inversion and learning [9,8,7] the second is a categorical reconstruction of relative entropy [3,2,15]. The present paper provides a categorical treatment of entropy in the spirit of Baez and Fritz in the setting of standard Borel spaces, thus setting the stage to explore the role of entropy in learning.

Recently there have been some exciting developments that bring some categorical insights to probability theory and specifically to learning theory. These are reported in some recent papers by Clerc, Dahlqvist, Danos and Garnier [9,8,7]. The first of

[1] This research has been supported by NSERC.

[2] Email: nicolas.gagne@mail.mcgill.ca

[3] Email: prakash@cs.mcgill.ca

This paper is electronically published in
Electronic Notes in Theoretical Computer Science
URL: www.elsevier.nl/locate/entcs

these papers showed how to view the Dirichlet distribution as a natural transformation thus opening the way to an understanding of higher-order probabilities, while the second gave a powerful framework for constructing several natural transformations. In [9] the hope was expressed that one could use these ideas to understand Bayesian inversion, a core concept in machine learning. In [7] this was realized in a remarkably novel way. These papers carry out their investigations in the setting of standard Borel spaces and are based on the Giry monad [11,13].

In [3,2] a beautiful treatment of relative entropy is given in categorical terms. The basic idea is to understand entropy in terms of the results of experiments and observations. How much does one learn about a probabilistic situation by doing experiments and observing the results? A category is set up where the morphisms capture the interplay between the original space and the space of observations. In order to interpret the relative entropy as a functor they use Lawvere's category which consists of a single object and a morphism for every extended positive real number [14].

Our contribution is to develop the theory of Baez et al. in the setting of standard Borel spaces [4]; their work is carried out with finite sets. While the work of [2] gives a firm conceptual direction, it gives little guidance in the actual development of the mathematical theory. We had to redevelop the mathematical framework and find the right analogues for the concepts appropriate to the finite case.

2 Background

In this section we review some of the background. We assume that the reader is familiar with concepts from topology and measure theory as well as basic category theory. We have found books by Ash [1], Billingsley [4] and Dudley [10] to be useful.

We will use letters like X, Y, Z for measurable spaces and capital Greek letters like Σ, Λ, Ω for σ-algebras. We will use $p, q, \ldots$ for probability measures. Given (X, Σ) and (Y, Λ) and a measurable function $f : X \longrightarrow Y$ and a probability measure p on (X, Σ) we obtain a measure on (Y, Λ) by $p \circ f^{-1}$; this is called the *pushforward* measure or the *image* measure.

2.1 The Giry monad

We denote the category of measurable spaces and measurable functions by **Mes**. We recall the Gìry [11] [5] functor $\Gamma : \mathbf{Mes} \to \mathbf{Mes}$ which maps each measurable space X to the space $\Gamma(X)$ of probability measures over X. Let $A \in \Sigma$, we define $\mathsf{ev}_A : \Gamma(X) \longrightarrow [0, 1]$ by $\mathsf{ev}_A(p) = p(A)$. We endow $\Gamma(X)$ with the smallest σ-algebra making all the ev's measurable. A morphism $f : X \longrightarrow Y$ in **Mes** is mapped to $\Gamma(f) : \Gamma(X) \longrightarrow \Gamma(Y)$ by $\Gamma(f)(p) = p \circ f^{-1}$. With the following natural transformations,

[4] In an earlier draft we were sloppy about the difference between standard Borel spaces and Polish spaces. We are really working with standard Borel spaces which are Polish spaces but with the topology forgotten and the σ-algebra retained.

[5] Gìry's name does have the accent grave on the ı; however, we omit it from now on to ease the typesetting.

this endofunctor is a monad: the Giry monad. The natural transformation $\eta : I \longrightarrow \Gamma$ is given by $\eta_X(x) = \delta_x$, the Dirac measure concentrated at x. The monad multiplication $\mu : \Gamma^2 \to \Gamma$ is given by

$$\forall A \in \mathcal{B}(X),\ \mu_X(p)(A) := \int_{\Gamma(X)} \mathsf{ev}_A \, \mathrm{d}p$$

where p is a probability measure in $\Gamma(\Gamma(X))$ and $\mathsf{ev}_A : \Gamma(X) \longrightarrow [0,1]$ is the measurable function on $\Gamma(X)$ defined by $\mathsf{ev}_A(p) = p(A)$.

Even if **Mes** is an interesting category in and of itself, the need for regular conditional probabilities forces us to restrict ourselves to a subcategory of standard Borel spaces.

2.2 Standard Borel spaces and disintegration

The Radon-Nikodym theorem is the main tool used to show the existence of conditional probability distributions, also called Markov kernels, see the discussion below. It is a very general theorem, but it does not give as strong regularity features as one might want. A stronger theorem is needed; this is the so-called *disintegration theorem*. It requires stronger hypotheses on the space on which the kernels are being defined. A category of spaces that satisfy these stronger hypotheses is the category of standard Borel spaces. In order to define standard Borel spaces, we must first define Polish spaces.

Definition 2.1 A *Polish space* is a separable, completely metrizable topological space.

Definition 2.2 A *standard Borel space* is a measurable space obtained by forgetting the topology of a Polish space but retaining its Borel algebra. The category of standard Borel spaces has measurable functions as morphisms; we denote it by **StBor**.

We can now state a version of the *disintegration theorem*. The following is also known as *Rohlin's disintegration theorem* [17].

Theorem 2.3 (Disintegration) *Let (X,p) and (Y,q) be two standard Borel spaces equipped with probability measures, where q is the pushforward measure $q := p \circ f^{-1}$ for a Borel measurable function $f : X \to Y$. Then, there exists a q-almost everywhere uniquely determined family of probability measures $\{p_y\}_{y \in Y}$ on X such that*

(i) the function $y \mapsto p_y(A)$ is a Borel-measurable function for each Borel-measurable set $A \subset X$;

(ii) p_y is a probability measure on $f^{-1}(y)$ for q-almost all $y \in Y$;

(iii) for every Borel-measurable function $h : X \to [0,\infty]$,

$$\int_X h \, dp = \int_Y \int_{f^{-1}(y)} h \, dp_y dq.$$

The objects obtained are often called *regular conditional probability distributions*. One can find a crisp categorical formulation of disintegration in [7, Theorem 1].

2.3 The Kleisli category of Γ *on* **StBor**

It is well known that the Giry monad on **Mes** restricted to **StBor** admits the same monad structure. [11]

The Kleisli category of Γ has as objects standard Borel spaces and as morphisms maps from X to $\Gamma(Y)$: $h : X \longrightarrow (\mathcal{B}_Y \longrightarrow [0,1])$ which are measurable. Here $\mathcal{B}_Y$ stands for the Borel sets of Y and $\Gamma(Y)$ has the σ-algebra described above. Now we can curry this to write it as $h : X \times \mathcal{B}_Y \longrightarrow [0,1]$ or $h(x, U)$ where x is a point in X and U is a Borel set in Y. Written this way it is called a Markov kernel and one can view it as a transition probability function or conditional probability distribution given x. Composition of morphisms $f : X \longrightarrow Y$ and $g : Y \longrightarrow Z$ in the Kleisli category is given by the formula

$$(g \circ f)(x, V \subset Z) = \int_Y g(y, V)\, \mathrm{d}f(x, \cdot).$$

For an arrow $s : Y \to \Gamma(X)$ in **StBor**, we write s_y for $s(y)$ or, in kernel form $s(y, \cdot)$. For arrows $t : Z \to \Gamma(Y)$ and $s : Y \to \Gamma(X)$ in **StBor**, we denote their Kleisli composition by $s \mathbin{\tilde{\circ}} t := \mu_X \circ \Gamma(s) \circ t$. For standard Borel spaces equipped with a probability measure p, we sometimes omit the measure in the notation, *i.e.* we sometimes write X instead of (X, p). We say a probability measure p is *absolutely continuous* with respect to another measure q on the same measurable space X, denoted by $p \ll q$, if for all measurable sets B, $q(B) = 0$ implies that $p(B) = 0$.

We note that absolute continuity is preserved by Kleisli composition; the proof is straightforward.

Proposition 2.4 *Given a standard Borel space Y with probability measures q and q' such that $q \ll q'$. Then, for arbitrary standard Borel space X and morphism s from Y to $\Gamma(X)$, we have $s \mathbin{\tilde{\circ}} q \ll s \mathbin{\tilde{\circ}} q'$.*

3 The categorical setting

In this section, following Baez and Fritz [2] (see also [3]) we describe the categories **FinStat** and **FP** which they use for their characterization of entropy on finite spaces. We then introduce the category **SbStat** which will be the arena for the generalization to standard Borel spaces.

Before doing so, we define the notion of coherence which will play an important role in what follows.

Definition 3.1 Given standard Borel spaces X and Y with probability measure p and q, respectively, a pair (f, s), $f : (X, p) \to (Y, q)$ and $s : Y \to \Gamma(X)$, is said to be *coherent* when f is measure preserving, i.e., $q = p \circ f^{-1}$, and s_y is a probability

measure on $f^{-1}(y)$ q-almost everywhere. [6] If in addition, p is absolutely continuous with respect to $s \mathbin{\tilde{\circ}} q$, then we say that (f, s) is *absolutely coherent*.

Definition 3.2 The category **FinStat** has

- **Objects** : Pairs (X, p) where X is a finite set and p a probability measure on X.
- **Morphisms** : $\mathrm{Hom}(X, Y)$ are all coherent pairs (f, s), $f : X \to Y$ and $s : Y \to \Gamma(X)$.

We compose arrows $(f, s) : (X, p) \to (Y, q)$ and $(g, t) : (Y, q) \to (Z, m)$ as follows: $(g, t) \circ (f, s) := (g \circ f, s \mathbin{\tilde{\circ}_{fin}} t)$ where $\tilde{\circ}_{fin}$ is defined as

$$(s \mathbin{\tilde{\circ}_{fin}} t)_z(x) = \sum_{y \in Y} t_z(y) s_y(x).$$

One can think of f as a measurement process from X to Y and of s as a hypothesis about X given an observation in Y. We say that a hypothesis s is *optimal* if $p = s \mathbin{\tilde{\circ}_{fin}} q$, or equivalenty, if s is a disintegration of p along f . We denote by **FP** the subcategory of **FinStat** consisting of the same objects, but with only those morphisms where the hypothesis is optimal. See [3,2] and [15] for a discussion of these ideas.

We now leave the finite world for a more general one: the category **SbStat**.

Definition 3.3 The category **SbStat** has

- **Objects** : Pairs (X, p) where X is a standard Borel space and p a probability measure on the Borel subsets of X.
- **Morphisms** : $\mathrm{Hom}(X, Y)$ are all coherent pairs (f, s), $f : X \to Y$ and $s : Y \to \Gamma(X)$.

We compose arrows $(f, s) : (X, p) \to (Y, q)$ and $(g, t) : (Y, q) \to (Z, m)$ as follows: $(g, t) \circ (f, s) := (g \circ f, s \mathbin{\tilde{\circ}} t)$.

Following the graphical representation from [2] we represent composition as follows:

$$\begin{tikzcd} (X,p) \arrow[r, bend right, "f"'] & (Y,q) \arrow[l, squiggly, bend right, "s"'] \arrow[r, bend right, "g"'] & (Z,m) \arrow[l, squiggly, bend right, "t"'] \end{tikzcd} \xRightarrow[\text{Composition}]{} \begin{tikzcd} (X,p) \arrow[r, bend right, "g \circ f"'] & (Z,m) \arrow[l, squiggly, bend right, "s \mathbin{\tilde{\circ}} t"'] \end{tikzcd} .$$

The proof of the following proposition is done in the extended version.

Proposition 3.4 *Given coherent pairs the composition is coherent. If, in addition, they are absolutely coherent, the composition is absolutely coherent.*

We end this section by defining one more category; this one is due to Lawvere [14]. It is just the set $[0, \infty]$ but endowed with categorical structure. This allows numerical values associated with morphisms to be regarded as functors.

[6] Note that (f, s) being coherent is equivalent to $\eta_Y = \Gamma(f) \circ s$.

Definition 3.5 The category $[0, \infty]$ has

- **Objects** : One single object: $\bullet$.
- **Morphisms** : For each element $r \in [0, \infty]$, one arrow $r : \bullet \to \bullet$.

Arrow composition is defined as addition in $[0, \infty]$.

This is a remarkable category with monoidal closed structure and many other interesting properties.

4 Relative entropy functor

We recapitulate the definition of the relative entropy functor on **FinStat** from Baez and Fritz [2] and then extend it to **SbStat**.

Definition 4.1 The relative entropy functor RE_{fin} is defined from **FinStat** to $[0, \infty]$ as follows:

- **On Objects** : It maps every object (X, p) to $\bullet$.
- **On Morphisms** : It maps a morphism $(f, s) : (X, p) \to (Y, q)$ to $S_{fin}(p, s \mathbin{\tilde{\circ}_{fin}} q)$, where
$$S_{fin}(p, s \mathbin{\tilde{\circ}_{fin}} q) := \sum_{x \in X} p(x) \ln \left(\frac{p(x)}{(s \mathbin{\tilde{\circ}_{fin}} q)(x)} \right).$$

The convention from now on will be that $\infty \cdot c = c \cdot \infty = \infty$ for $0 < c \leq \infty$ and $\infty \cdot 0 = 0 \cdot \infty = 0$. We extend RE_{fin} from **FinStat** to **SbStat**.

Definition 4.2 The relative entropy functor RE is defined from **SbStat** to $[0, \infty]$ as follows:

- **On Objects** : It maps every object (X, p) to $\bullet$.
- **On Morphisms** : Given a coherent morphism $(f, s) : (X, p) \to (Y, q)$, if (f, s) is absolutely coherent, then $RE((f, s)) = S(p, s \mathbin{\tilde{\circ}} q)$, where
$$S(p, s \mathbin{\tilde{\circ}} q) := \int_X \log \left(\frac{\mathrm{d}p}{\mathrm{d}(s \mathbin{\tilde{\circ}} q)} \right) \mathrm{d}p,$$
otherwise it is defined as $RE((f, s)) = \infty$.

This quantity is also known as the *Kullback-Leibler divergence*.

We could have defined our category to have only absolutely coherent morphisms but it would make the comparison with the finite case more awkward as the finite case does not assume the morphisms to be absolutely coherent. The present definition leads to slightly awkward proofs where we have to consider absolutely coherent pairs and ordinary coherent pairs separately; most of which have been omitted in this abridged version but can be found in the extended version.

Clearly, RE restricts to RE_{fin} on **FinStat**. If (f, s) is absolutely coherent, then p is absolutely continuous with respect to $(s \tilde{\circ} q)$ and the Radon-Nikodym derivative is defined. The relative entropy is always non-negative [12]; this is an easy consequence of Jensen's inequality. This shows that RE is defined everywhere in **SbStat**.

We will use the following notation occasionally:

$$RE\Big((X,p) \underset{f}{\overset{s}{\rightleftarrows}} (Y,q)\Big) := RE((f,s)).$$

It remains to show that RE is indeed a functor. That is, we want to show that

$$RE\Big((X,p) \underset{f}{\overset{s}{\rightleftarrows}} (Y,q) \underset{g}{\overset{t}{\rightleftarrows}} (Z,m)\Big) = RE\Big((X,p) \underset{f}{\overset{s}{\rightleftarrows}} (Y,q)\Big) + RE\Big((Y,q) \underset{g}{\overset{t}{\rightleftarrows}} (Z,m)\Big).$$

In order to do so, we will need the following lemma.

Lemma 4.3 *The relative entropy is preserved under pre-composition by optimal hypotheses. In other words, we always have*

$$RE\Big((Y,q) \underset{g}{\overset{t}{\rightleftarrows}} (Z,m)\Big) = RE\Big((X, s \tilde{\circ} q) \underset{f}{\overset{s}{\rightleftarrows}} (Y,q) \underset{g}{\overset{t}{\rightleftarrows}} (Z,m)\Big).$$

Proof.

Case I : (g,t) is absolutely coherent. Since (g,t) is absolutely coherent, so is $(g \circ f, s \tilde{\circ} t)$ by Proposition 2.4. Hence, to show $RE(g,t) = RE(g \circ f, s \tilde{\circ} t)$ is to show

$$\int_Y \log\left(\frac{\mathrm{d}q}{\mathrm{d}(t \tilde{\circ} m)}\right) \mathrm{d}q = \int_X \log\left(\frac{\mathrm{d}(s \tilde{\circ} q)}{\mathrm{d}(s \tilde{\circ} t \tilde{\circ} m)}\right) \mathrm{d}(s \tilde{\circ} q).$$

Because f is measure preserving, it is sufficient to show that the following functions on X

$$\frac{\mathrm{d}q}{\mathrm{d}(t \tilde{\circ} m)} \circ f = \frac{\mathrm{d}(s \tilde{\circ} q)}{\mathrm{d}(s \tilde{\circ} t \tilde{\circ} m)} \ s \tilde{\circ} t \tilde{\circ} m\text{-almost everywhere.}$$

By the Radon-Nikodym theorem, it is sufficient to show that for any $E \subset X$ measurable set, we have

$$(s \tilde{\circ} q)(E) = \int_E \frac{\mathrm{d}q}{\mathrm{d}(t \tilde{\circ} m)} \circ f \ \mathrm{d}(s \tilde{\circ} t \tilde{\circ} m).$$

The following calculation establishes the above.

$$
\begin{aligned}
&\int_E \frac{\mathrm{d}q}{\mathrm{d}(t \mathbin{\tilde{\circ}} m)} \circ f \; \mathrm{d}(s \mathbin{\tilde{\circ}} t \mathbin{\tilde{\circ}} m) && (1)\\
&\quad = \int_Y \left(\int_{x \in f^{-1}(y) \cap E} \left(\frac{\mathrm{d}q}{\mathrm{d}(t \mathbin{\tilde{\circ}} m)} \circ f \right)(x) \; \mathrm{d}(s \mathbin{\tilde{\circ}} t \mathbin{\tilde{\circ}} m)_y \right) \mathrm{d}(t \mathbin{\tilde{\circ}} m) && (2)\\
&\quad = \int_Y \frac{\mathrm{d}q}{\mathrm{d}(t \mathbin{\tilde{\circ}} m)}(y) \left(\int_{f^{-1}(y) \cap E} \mathrm{d}(s \mathbin{\tilde{\circ}} t \mathbin{\tilde{\circ}} m)_y \right) \mathrm{d}(t \mathbin{\tilde{\circ}} m) && (3)\\
&\quad = \int_Y \frac{\mathrm{d}q}{\mathrm{d}(t \mathbin{\tilde{\circ}} m)}(y) \left(\int_{f^{-1}(y) \cap E} \mathrm{d}s_y \right) \mathrm{d}(t \mathbin{\tilde{\circ}} m) && (4)\\
&\quad = \int_Y \frac{\mathrm{d}q}{\mathrm{d}(t \mathbin{\tilde{\circ}} m)}(y) s_y(E \cap f^{-1}(y)) \; \mathrm{d}(t \mathbin{\tilde{\circ}} m) && (5)\\
&\quad = \int_Y \frac{\mathrm{d}q}{\mathrm{d}(t \mathbin{\tilde{\circ}} m)}(y) s_y(E) \; \mathrm{d}(t \mathbin{\tilde{\circ}} m) && (6)\\
&\quad = \int_Y s_y(E) \; \mathrm{d}q && (7)\\
&\quad = (s \mathbin{\tilde{\circ}} q)(E) && (8)
\end{aligned}
$$

We get (2) by applying the disintegration theorem to $f : (X, s \mathbin{\tilde{\circ}} t \mathbin{\tilde{\circ}} m) \to (Y, t \mathbin{\tilde{\circ}} m)$. The equation (3) follows by using the fact that $\frac{\mathrm{d}q}{\mathrm{d}(t \mathbin{\tilde{\circ}} m)} \circ f$ is constant on $f^{-1}(y)$ for every y. To obtain (4) we apply Lemma **??**. To show (6) we use the fact that s_y is a probability measure on $f^{-1}(y)$. We get (7) by the definition of the Radon-Nikodym derivative and we finally establish (8) by the definition of Kleisli composition.

Case II : (g, t) is not absolutely coherent. The proof is simple but slightly tedious. It can be found in the extended version. ∎

Theorem 4.4 (Functoriality) *Given arrows $(f, s) : (X, p) \to (Y, q)$ and $(g, t) : (Y, q) \to (Z, m)$, we have*

$$RE\left((g, t) \circ (f, s)\right) = RE((f, s)) + RE((g, t)).$$

Proof. Note that by definition, $RE\left((g, t) \circ (f, s)\right) = RE\left((g \circ f, s \mathbin{\tilde{\circ}} t)\right)$.

Case I : (f, s) and (g, t) are absolutely coherent. By Proposition 3.4, we have that $(g \circ f, s \mathbin{\tilde{\circ}} t)$ is absolutely coherent.

$$
\begin{aligned}
RE\left((g \circ f, s \mathbin{\tilde{\circ}} t)\right) &= \int_X \log\left(\frac{\mathrm{d}p}{\mathrm{d}(s \mathbin{\tilde{\circ}} t \mathbin{\tilde{\circ}} m)}\right) \mathrm{d}p && (9)\\
&= \int_X \log\left(\frac{\mathrm{d}p}{\mathrm{d}(s \mathbin{\tilde{\circ}} q)} \frac{\mathrm{d}(s \mathbin{\tilde{\circ}} q)}{\mathrm{d}(s \mathbin{\tilde{\circ}} t \mathbin{\tilde{\circ}} m)}\right) \mathrm{d}p && (10)\\
&= \int_X \log\left(\frac{\mathrm{d}p}{\mathrm{d}(s \mathbin{\tilde{\circ}} q)}\right) \mathrm{d}p + \int_X \log\left(\frac{\mathrm{d}(s \mathbin{\tilde{\circ}} q)}{\mathrm{d}(s \mathbin{\tilde{\circ}} t \mathbin{\tilde{\circ}} m)}\right) \mathrm{d}p && (11)\\
&= RE((f,s)) + \int_X \log\left(\frac{\mathrm{d}(s \mathbin{\tilde{\circ}} q)}{\mathrm{d}(s \mathbin{\tilde{\circ}} t \mathbin{\tilde{\circ}} m)}\right) \mathrm{d}p && (12)\\
&= RE((f,s)) + RE((g,t)) && (13)
\end{aligned}
$$

We get (10) by the chain rule for Radon-Nikodym derivatives and (13) by applying Lemma 4.3.

Case II : (g,t) is not absolutely coherent. The proof is very similar to the second case of the previous lemma. It can be found in the extended version.

Case III : (f,s) is not absolutely coherent.

Although we relegated the proof of case III to the extended version, it is neither trivial nor boring. [7] For both of the above cases, we deduce that

$$RE\left((g,t) \circ (f,s)\right) = \infty = RE((f,s)) + RE((g,t)).$$

∎

We have thus shown that RE is a well-defined functor from **SbStat** to $[0,\infty]$.

4.1 Convex linearity

We show below that the relative entropy functor satisfies a convex linearity property. In [2] convexity looks familiar; here since we are performing "large" sums we have to express it as an integral. First we define a localized version of the relative entropy.

Note that Lemma **??** in the appendix says that $s_y = (s \mathbin{\tilde{\circ}} q)_y$ q-almost everywhere. Thus, in the following there is no notational clash between the kernel s_y and $(s \mathbin{\tilde{\circ}} q)_y$, the later being the disintegration of $(s \mathbin{\tilde{\circ}} q)$ along f.

Given an arrow $(f,s) : (X,p) \to (Y,q)$ in **StBor** and a point $y \in Y$, we denote by $(f,s)_y$, the morphism (f,s) restricted to the pair of standard Borel spaces $f^{-1}(y)$ and $\{y\}$. Explicitly,

$$(f,s)_y := (f|_{f^{-1}(y)}, s_y) : (f^{-1}(y), p_y) \longrightarrow (\{y\}, \delta_y),$$

where δ_y is the one and only probability measure on $\{y\}$.

[7] It is not analogous to the previous case since the existence of a measurable set $A \subset X$ such that $(s \mathbin{\tilde{\circ}} q)(A) = 0$ and $p(A) > 0$ is surprisingly not enough to conclude that $(s \mathbin{\tilde{\circ}} t \mathbin{\tilde{\circ}} m)(A) = 0$.

Definition 4.5 A functor F from **SbStat** to $[0, \infty]$ is *convex linear* if for every arrow $(f, s) : (X, p) \to (Y, q)$, we have

$$F\left((f, s)\right) = \int_Y F\left((f, s)_y\right) \, \mathrm{d}q.$$

We will sometimes refer to the relative entropy of $(f, s)_y$ as the *local relative entropy of* (f, s) *at* y.

Theorem 4.6 (Convex Linearity) *The functor RE is convex linear, i.e., for every arrow* $(f, s) : (X, p) \to (Y, q)$*, we have*

$$RE((f, s)) = \int_Y RE\left((f, s)_y\right) \, \mathrm{d}q.$$

Proof.

Case I : (f, s) **is absolutely coherent.**

Note that by Lemma **??**, $p_y \ll (s \mathbin{\tilde{\circ}} q)_y$ almost everywhere. So we have

$$\begin{aligned}
RE((f, s)) &= \int_X \log\left(\frac{\mathrm{d}p}{\mathrm{d}(s \mathbin{\tilde{\circ}} q)}\right) \mathrm{d}p && (14)\\
&= \int_Y \int_{f^{-1}(y)} \log\left(\frac{\mathrm{d}p}{\mathrm{d}(s \mathbin{\tilde{\circ}} q)}\right) \mathrm{d}p_y \, \mathrm{d}q && (15)\\
&= \int_Y \int_{f^{-1}(y)} \log\left(\frac{\mathrm{d}p_y}{\mathrm{d}(s \mathbin{\tilde{\circ}} q)_y}\right) \, \mathrm{d}p_y \, \mathrm{d}q && (16)\\
&= \int_Y RE((f, s)_y) \, \mathrm{d}q. && (17)
\end{aligned}$$

We get (15) by the disintegration theorem and (16) by applying Lemma **??**.

Case II : (f, s) **is not absolutely coherent.** The proof is not hard and can be found in the extended version.

∎

4.2 Lower-semi-continuity

Recall that a sequence of probability measures p_n converges strongly to p, denoted by $p_n \to p$, if for all measurable set E, one has $\lim_{n\to\infty} p_n(E) = p(E)$.

Definition 4.7 A functor F from **SbStat** to $[0, \infty]$ is *lower semi-continuous* if for every arrow $(f, s) : (X, p) \to (Y, q)$, whenever $p_n \to p$, then

$$F\Bigg(\begin{tikzcd}(X,p) \arrow[rr, bend right, "f"'] & & (Y,q) \arrow[ll, squiggly, bend right, "s"']\end{tikzcd}\Bigg) \leq \liminf_{n\to\infty} F\Bigg(\begin{tikzcd}(X,p_n) \arrow[rr, bend right, "f"'] & & (Y,q_n) \arrow[ll, squiggly, bend right, "s"']\end{tikzcd}\Bigg).$$

Note that a lower semi-continuous functor F on **SbStat** restricts to a lower semi-continuous (as defined slightly differently in [2]) functor on **FinStat**.

Theorem 4.8 (Lower semi-continuity) *The functor RE is lower semi-continuous.*

Proof. This is a direct consequence of Pinsker [16, Section 2.4]. ∎

5 Uniqueness

We now show that the relative entropy is, up to a multiplicative constant, the unique functor satisfying the conditions established so far. We first prove a crucial lemma.

Lemma 5.1 *Let X be a Borel space equipped with probability measures p and q, if $p \ll q$, then we can find a sequence of simple functions p_n^* on X such that for the sequence of probability measures $p_n(E) := \int_E p_n^* \, \mathrm{d}q$, we have that p_n and p agree on the elements of the partition on X induced by p_n^* and moreover, $p_n \to p$ strongly.*

Proof. We write $I_{n,k}$ for the interval $[k2^{-n}, (k+1)2^{-n})$ and $I_{n,\leq}$ for the interval $[n, \infty)$. Denote by K_n the index set $\{0, 1, \ldots, n2^n - 1, \leq\}$ of k. We fix a version $\frac{\mathrm{d}p}{\mathrm{d}q}$ of the Radon-Nikodym such that $\frac{\mathrm{d}p}{\mathrm{d}q} < \infty$ everywhere. We define a family of partitions and a family of simple functions as follows:

$$X_{n,k} := \left\{ x' \in X \mid \frac{\mathrm{d}p}{\mathrm{d}q}(x') \in I_{n,k} \right\}, \qquad p_n^*(x) := \frac{p\left(X_{n,k}\right)}{q\left(X_{n,k}\right)} \text{ on } x \in X_{n,k}.$$

Every function induces a partition on the domain; if moreover the function is simple, the induced partition is finite.

We first note that p_n and p agree on the elements of the partition induced by p_n^*:

$$p_n(X_{n,k}) = \int_{X_{n,k}} p_n^* \, \mathrm{d}q = \int_{X_{n,k}} \frac{p\left(X_{n,k}\right)}{q\left(X_{n,k}\right)} \, \mathrm{d}q = \frac{p\left(X_{n,k}\right)}{q\left(X_{n,k}\right)} q(X_{n,k}) = p(X_{n,k}).$$

Next, we prove the strong convergence of $p_n \to p$. We first show $p_n^* \to \frac{\mathrm{d}p}{\mathrm{d}q}$ pointwise. Let $x \in X$. Pick N large enough such that $\frac{\mathrm{d}p}{\mathrm{d}q}(x) \leq N$. For a fixed integer $n \geq N$, there is exactly one k_n for which $x \in X_{n,k_n}$. On the one hand, we have $k_n 2^{-n} \leq \frac{\mathrm{d}p}{\mathrm{d}q}(x) \leq (k_n+1)2^{-n}$ on X_{n,k_n}. But on the other hand, by integrating over X_{n,k_n} and dividing everything by $q(X_{n,k_n})$, we also have $k_n 2^{-n} \leq \frac{p\left(X_{n,k_n}\right)}{q\left(X_{n,k_n}\right)} \leq (k_n+1)2^{-n}$ on X_{n,k_n}. We thus get pointwise convergence since we have

$$\left| p_n^*(x) - \frac{\mathrm{d}p}{\mathrm{d}q}(x) \right| = \left| \frac{p\left(X_{n,k_n}\right)}{q\left(X_{n,k_n}\right)} - \frac{\mathrm{d}p}{\mathrm{d}q}(x) \right| \leq 2^{-n} \text{ for any } n \geq N.$$

From the above inequality and the choice of N, we note the following

$$p_n^*(x) \leq \frac{\mathrm{d}p}{\mathrm{d}q}(x) + 2^{-n} \leq \frac{\mathrm{d}p}{\mathrm{d}q}(x) + 1, \quad \text{for } x \text{ with } \frac{\mathrm{d}p}{\mathrm{d}q}(x) < n,$$

$$p_n^*(x) = p(X_{n,\leq}) \leq 1 \leq \frac{\mathrm{d}p}{\mathrm{d}q}(x) + 1, \quad \text{for } x \text{ with } \frac{\mathrm{d}p}{\mathrm{d}q}(x) \geq n.$$

So for all n, we can bound $p_n^*(x)$ everywhere by the integrable function $g(x) := \frac{\mathrm{d}p}{\mathrm{d}q}(x) + 1$. Given a measurable set $E \subset X$, we can thus apply Lebesgue's dominated convergence theorem. We get

$$\lim_{n\to\infty} p_n(E) = \lim_{n\to\infty} \int_E p_n^* \, \mathrm{d}q = \int_E \lim_{n\to\infty} p_n^* \, \mathrm{d}q = \int_E \frac{\mathrm{d}p}{\mathrm{d}q} \, \mathrm{d}q = p(E).$$

∎

Before proving uniqueness, we recall the main theorem of Baez and Fritz [2] on **FinStat**.

Theorem 5.2 *Suppose that a functor*

$$F : \mathbf{FinStat} \to [0, \infty]$$

is lower semicontinuous, convex linear and vanishes on **FP**. *Then for some* $0 \leq c \leq \infty$ *we have* $F(f, s) = cRE_{fin}(f, s)$ *for all morphisms* (f, s) *in* **FinStat**.

We are now ready to extend this characterization to **SbStat**.

Theorem 5.3 *Suppose that a functor*

$$F : \mathbf{SbStat} \to [0, \infty]$$

is lower semicontinuous, convex linear and vanishes on **FP**. *Then for some* $0 \leq c \leq \infty$ *we have* $F(f, s) = cRE(f, s)$ *for all morphisms.*

Proof. Since F satisfies all the above properties on **FinStat**, we can apply Theorem 5.2 in order to establish that $F = cRE_{fin} = cRE$ for all morphisms in the subcategory **FinStat**. We show that F extends uniquely to cRE on all morphisms in **SbStat**.

By convex linearity of F, for an arbitrary morphism (f, s) from (X, p) to (Y, q), we have

$$F((f, s)) = \int_Y F((f, s)_y) \, \mathrm{d}q,$$

so F is totally described by its local relative entropies. It is thus sufficient to show $F = cRE$ on an arbitrary morphism $(f, s) : (X, p) \to (\{y\}, \delta_y)$. The case where p is not absolutely continuous with respect to s is straightforward, so let us assume $p \ll s$. [8]

[8] See the extended version for details.

We apply Lemma 5.1 with p and s to get the family of simple functions p_n^* and the corresponding family of partitions $\{X_{n,k}\}$. We define π_n as the function that maps $x \in X_{n,k'}$ to the element $X_{n,k'} \in \{X_{n,k}\}_{k \in K_n}$. Denote by s_{π_n} the disintegration of s along π_n and by s_n the corresponding marginal. Note that since p_n and p agree on every $X_{n,k}$, p_n is indeed the push-forward of p along π_n, so we can identify p_n to the corresponding marginal of p. Presented as diagrams, we have

$$\begin{tikzcd} (X,p) \arrow[r, bend right, "\pi_n"'] & (\{X_{n,k}\}, p_n) \arrow[l, squiggly, bend right, "s_{\pi_n}"'] \arrow[r, bend right, "f_n"'] & (\{y\}, \delta_y) \arrow[l, squiggly, bend right, "s_n"'] \end{tikzcd} \xRightarrow[\text{Composition}]{} \begin{tikzcd} (X,p) \arrow[r, bend right, "f"'] & (\{y\}, \delta_y) \arrow[l, squiggly, bend right, "s"'] \end{tikzcd}.$$

From the above diagram and the hypothesis that F is a functor, we have the following inequality

$$F((f_n, s_n)) \leq F((f, s)), \text{ for all } n \in \mathbb{N}. \tag{18}$$

Note that, on the one hand the disintegration of p_n along π_n at the point $X_{n,k'} \in \{X_{n,k}\}$ is given by $p_{n,\pi} := p_n(\cdot)/p_n(X_{n,k'})$, but on the other hand, for any measurable set $E \subset X$, we also have

$$\sum_{k \in K_n} \left(\int_{X_{n,k}} \mathbb{1}_E \, \mathrm{d}p_{n,\pi} \right) s_n(X_{n,k}) = \sum_{k \in K_n} \left(\frac{p_n(E \cap X_{n,k})}{p_n(X_{n,k})} \right) s_n(X_{n,k})$$

$$= \sum_{k \in K_n} \left(\frac{s(E \cap X_{n,k})}{s(X_{n,k})} \right) s_n(X_{n,k}) = \sum_{k \in K_n} s(E \cap X_{n,k}) = s(E).$$

This means that $p_{n,\pi}$ is the disintegration of s along π_n. Presented as diagrams, where we use f^{p_n} instead of f to indicate that the arrow leaves from the object (X, p_n) as opposed to (X, p), we have

$$\begin{tikzcd} (X,p_n) \arrow[r, bend right, "\pi_n"'] & (\{X_{n,k}\}, p_n) \arrow[l, squiggly, bend right, "p_{n,\pi}"'] \arrow[r, bend right, "f_n"'] & (\{y\}, \delta_y) \arrow[l, squiggly, bend right, "s_n"'] \end{tikzcd} \xRightarrow[\text{Composition}]{} \begin{tikzcd} (X,p_n) \arrow[r, bend right, "f^{p_n}"'] & (\{y\}, \delta_y) \arrow[l, squiggly, bend right, "s"'] \end{tikzcd}.$$

But since F vanishes on **FP**, we have $F\left((\pi_n, p_{n,\pi})\right) = 0$. Combined with the fact that F is a functor, we get

$$F\left((f^{p_n}, s)\right) = F\left((\pi_n, p_{n,\pi})\right) + F\left((f_n, s_n)\right) = F\left((f_n, s_n)\right). \tag{19}$$

By Lemma 5.1, we know that $p_n \to p$, in terms of our diagrams we have

$$\begin{tikzcd} (X,p_n) \arrow[r, bend right, "f^{p_n}"'] & (\{y\}, \delta_y) \arrow[l, squiggly, bend right, "s"'] \end{tikzcd} \xRightarrow[\text{Strong Convergence}]{} \begin{tikzcd} (X,p) \arrow[r, bend right, "f"'] & (\{y\}, \delta_y) \arrow[l, squiggly, bend right, "s"'] \end{tikzcd}.$$

Hence, combining (19) with the lower semicontinuity of F, we also have the inequality

$$F((f,s)) \leq \liminf_{n\to\infty} F((f^{p_n},s)) = \liminf_{n\to\infty} F((f_n,s_n)). \tag{20}$$

Since (f_n, s_n) is in **FinStat**, we must have $F((f_n,s_n)) = cRE((f_n,s_n))$. Thus, combining (18) and (20), we get that $F((f,s))$ must satisfy

$$\limsup_{n\to\infty} cRE((f_n,s_n)) \leq F((f,s)) \leq \liminf_{n\to\infty} cRE((f_n,s_n)),$$

but so does $cRE((f,s))$. We also have

$$\limsup_{n\to\infty} cRE((f_n,s_n)) \leq cRE((f,s)) \leq \liminf_{n\to\infty} cRE((f_n,s_n)).$$

Therefore $F((f,s)) = cRE((f,s))$, as desired. ∎

6 Conclusions and Further Directions

As promised, we have given a categorial characterization of relative entropy on standard Borel spaces. This greatly broadens the scope of the original work by Baez et al. [3,2]. However, the main motivation is to study the role of entropy arguments in machine learning. These appear in various ad-hoc ways in machine learning but with the appearance of the recent work by Danos and his co-workers [9,7,8] we feel that we have the prospect of a mathematically well-defined framework on which to understand Bayesian inversion and its interplay with entropy. The most recent paper in this series [7] adopts a point-free approach introduced in [5,6]. It would be interesting to extend our definitions to a point-free situation. There are also many interesting questions with regard to understanding the "algebra of entropy"; see the book by Yeung [18] for a taste of these ideas.

Acknowledgments

We have benefitted from discussions with Florence Clerc, Vincent Danos, Tobias Fritz, Renaud Raquépas and Ilias Garnier. We are grateful to the anonymous referees for very helpful comments. This research was supported by a grant from Google and from NSERC.

References

[1] Ash, R. B., "Real Analysis and Probability," Academic Press, 1972.

[2] Baez, J. C. and T. Fritz, *A bayesian characterization of relative entropy*, Theory and Applications of Categories **29** (2014), pp. 422–456.

[3] Baez, J. C., T. Fritz and T. Leinster, *A characterization of entropy in terms of information loss*, Entropy **13** (2011), pp. 1945–1957.

[4] Billingsley, P., "Probability and Measure," Wiley-Interscience, 1995.

[5] Chaput, P., V. Danos, P. Panangaden and G. Plotkin, *Approximating Markov processes by averaging*, in: *Proceedings of the 37th International Colloquium On Automata Languages And Programming (ICALP)*, Lecture Notes In Computer Science **5556**, 2009, pp. 127–138.

[6] Chaput, P., V. Danos, P. Panangaden and G. Plotkin, *Approximating Markov processes by averaging*, J. ACM **61** (2014), pp. 5:1–5:45.
URL http://doi.acm.org/10.1145/2537948

[7] Clerc, F., V. Danos, F. Dahlqvist and I. Garnier, *Pointless learning*, in: *Proceedings of FoSSaCS 2017*, 2017.

[8] Dahlqvist, F., V. Danos and I. Garnier, *Giry and the machine*, Electronic Notes in Theoretical Computer Science **325** (2016), pp. 85–110.

[9] Danos, V. and I. Garnier, *Dirichlet is natural*, Electronic Notes in Theoretical Computer Science **319** (2015), pp. 137–164.

[10] Dudley, R. M., "Real Analysis and Probability," Wadsworth and Brookes/Cole, 1989.

[11] Giry, M., *A categorical approach to probability theory*, in: B. Banaschewski, editor, *Categorical Aspects of Topology and Analysis*, number 915 in Lecture Notes In Mathematics (1981), pp. 68–85.

[12] Kullback, S. and R. A. Leibler, *On information and sufficiency*, The annals of mathematical statistics **22** (1951), pp. 79–86.

[13] Lawvere, F. W., *The category of probabilistic mappings* (1964), unpublished typescript.

[14] Lawvere, F. W., *Metric spaces, generalized logic and closed categories*, Rend. Sem. Mat. Fis. Milano **43** (1973), pp. 135–166.

[15] Leinster, T., *An operadic introduction to entropy*, n-category cafe.

[16] Pinsker, M. S., *Information and information stability of random variables and processes* (1960).

[17] Rokhlin, V. A., *On the fundamental ideas of measure theory*, Matematicheskii Sbornik **67** (1949), pp. 107–150.

[18] Yeung, R. W., "Information theory and network coding," Springer Science & Business Media, 2008.

A Lemmas

Lemma A.1 *Given an arrow $(f, s) : (X, p) \to (Y, q)$ in* **SbStat**. *Let $\{(s \mathbin{\tilde{\circ}} q)_y\}_{y \in Y}$ denote the collection of conditional probability measures of $(s \mathbin{\tilde{\circ}} q)$ conditioned by Y, then $(s \mathbin{\tilde{\circ}} q)_y = s_y$ q-almost everywhere.*

Lemma A.2 *Given*

$$(X, p) \xrightarrow{f} (Y, q) \xleftarrow{f} (X, p')$$

where f is a continuous function preserving the measure of both Borel probability measures p and p'. If $p \ll p'$, then $\frac{\mathrm{d}p_y}{\mathrm{d}p'_y}(x) = \frac{\mathrm{d}p}{\mathrm{d}p'}(x)$ p'-almost everywhere.