Logical Methods in Computer Science
Volume 19, Issue 4, 2023, pp. 10:1–10:18
https://lmcs.episciences.org/

Submitted Jun. 25, 2021
Published Nov. 09, 2023

A CATEGORICAL CHARACTERIZATION OF RELATIVE ENTROPY ON STANDARD BOREL SPACES

NICOLAS GAGNÉ [a] AND PRAKASH PANANGADEN [b]

[a] Département d'informatique et de recherche opérationnelle, Université de Montréal, Montréal, Québec, Canada
e-mail address: nicolas.gagne.5@umontreal.ca

[b] School of Computer Science, McGill University, Montréal, Québec, Canada
e-mail address: prakash@cs.mcgill.ca

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.

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 [DG15, DDG16, CDDG17, DSDG18]. The second is a categorical reconstruction of relative entropy [BFL11, BF14, Lei]. 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 [DG15, DDG16, CDDG17]. The first of 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 [DG15] the hope was expressed that one could use these ideas to understand Bayesian inversion, a core concept in machine learning. In [CDDG17] this was realized in a remarkably novel way.

Key words and phrases: Entropy, Giry Monad, Bayesian Learning, standard Borel spaces, scoring rules.
* An extended abstract of this work appeared in the 33^{rd} Mathematical Foundations of Programming Semantics Conference.

This research has been supported by NSERC and by Google.

LOGICAL METHODS IN COMPUTER SCIENCE
DOI:10.46298/LMCS-19(4:10)2023

These papers carry out their investigations in the setting of standard Borel spaces and are based on the Giry monad [Gir81, Law64].

In [BFL11, BF14] 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 [Law73].

Our contribution is to develop the theory of Baez et al. in the setting of standard Borel spaces; their work is carried out with finite sets. While the work of [BF14] 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 [Ash72], Billingsley [Bil95] and Dudley [Dud89] 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 $f : X \to Y$ 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 Giry [Gir81] 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 $\mathsf{ev}_A : \Gamma(X) \to [0,1]$ by $\mathsf{ev}_A(p) = p(A)$. We endow $\Gamma(X)$ with the smallest σ-algebra making all the ev's measurable. A morphism $f : X \to Y$ in **Mes** is mapped to $\Gamma(f) : \Gamma(X) \to \Gamma(Y)$ by $\Gamma(f)(p) = p \circ f^{-1}$. With the following natural transformations, this endofunctor is a monad: the Giry monad. The natural transformation $\eta : I \to \Gamma$ 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 $\mathsf{ev}_A : \Gamma(X) \to [0,1]$ 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.*

Theorem 2.3 [Rok49]. *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*

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

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

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

$$\int_X h \, \mathrm{d}p = \int_Y \int_{f^{-1}(y)} h \, \mathrm{d}p_y \mathrm{d}q.$$

The objects obtained are often called *regular conditional probability distributions.* One can find a crisp categorical formulation of disintegration in [CDDG17, 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. [Gir81]

The Kleisli category of Γ has as objects standard Borel spaces and as morphisms maps from X to $\Gamma(Y)$: $h : X \to (\mathcal{B}_Y \to [0, 1])$ 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 $h : X \times \mathcal{B}_Y \to [0, 1]$ 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 $f : X \to Y$ and $g : Y \to Z$ in the Kleisli category is given by the formula

$$(g \circ f)(x, V \in \mathcal{B}_Z) = \int_Y g(y, V) \, \mathrm{d}f(x, \cdot).$$

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 $s \,\tilde{\circ}\, t := \mu_X \circ \Gamma(s) \circ t$. 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 $s \,\tilde{\circ}\, q \ll s \,\tilde{\circ}\, q'$.*

3. The categorical setting

In this section, following Baez and Fritz [BF14] (see also [BFL11]) we describe the category **FinStat** 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), with $f : (X, p) \to (Y, q)$ and $s : Y \to \Gamma(X)$ measurable, is said to be *coherent*[1] 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. [2] If in addition, p is absolutely continuous with respect to $s \,\tilde{\circ}\, q$, 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: $(g, t) \circ (f, s) := (g \circ f, s \,\tilde{\circ}_{fin}\, t)$ where $\tilde{\circ}_{fin}$ is defined as

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

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

[1] Note that a coherent pair (f, s) by definition satisfies condition (1) and condition (2) of Theorem (2.3) but is not required to satisfy condition (3).

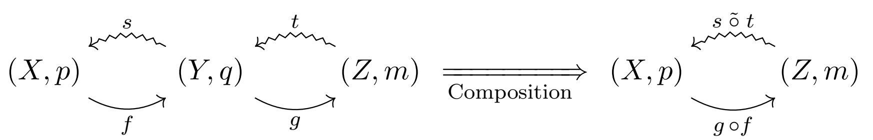
[2] Note that (f, s) being coherent is equivalent to $\eta_Y = \Gamma(f) \circ s$.

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: $(g, t) \circ (f, s) := (g \circ f, s \,\tilde{\circ}\, t)$.

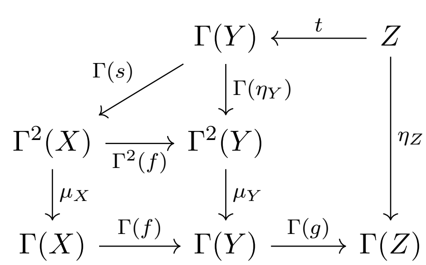
Note that the identity arrow on object (X, p) is (id_X, η_X) where id_X is the identity function on X. Following the graphical representation from [BF14] we represent composition as follows:



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*[3] if $p = s \,\tilde{\circ}\, q$. We denote by **FP** the subcategory of **SbStat** consisting of the same objects, but with only those morphisms where the hypothesis is optimal. See [BFL11, BF14] and [Lei] for a discussion of these ideas in the finite case.

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

Proof. We first show that the composition is coherent, i.e., $\eta_Z = (\Gamma(g) \circ \Gamma(f)) \circ (s \,\tilde{\circ}\, t)$. It is sufficient to show that the following diagram commutes:



Using the hypothesis that $\eta_Z = \Gamma(g) \circ t$ and the fact that $\mathrm{Id} = \mu \circ \Gamma(\eta)$, we get that the right-hand square commutes. The triangle commutes since it is the application of Γ to our hypothesis $\eta_Y = \Gamma(f) \circ s$ and the left-hand square commutes because μ is a natural transformation. Therefore, the whole diagram commutes and we have thus shown the composition of coherent morphisms is also coherent.

Next, in addition, assume the pairs (f, s) and (g, t) are absolutely coherent. We show $p \ll (s \,\tilde{\circ}\, t \,\tilde{\circ}\, m)$. By hypotheses, $p \ll s \,\tilde{\circ}\, q$ and $q \ll t \,\tilde{\circ}\, m$. Using Proposition 2.4 on $q \ll t \,\tilde{\circ}\, m$, we get $s \,\tilde{\circ}\, q \ll s \,\tilde{\circ}\, t \,\tilde{\circ}\, m$. By transitivity of $\ll$, we conclude $p \ll (s \,\tilde{\circ}\, t \,\tilde{\circ}\, m)$. □

[3]For a coherent pair (f, s), asking s to be optimal is equivalent to asking that (f, s) satisfies condition (3) in Theorem (2.3) as will be shown in Lemma (4.3).

We end this section by defining one more category; this one is due to Lawvere [Law73]. It is just the set $[0,\infty]$ but endowed with categorical structure. This allows numerical values associated with morphisms to be regarded as functors.

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 $[0,\infty]$. Consequently, 0 is the identity arrow.

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 [BF14] 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 $S_{fin}(p, s\,\tilde{\circ}_{fin}\,q)$, where
$$S_{fin}(p, s\,\tilde{\circ}_{fin}\,q) := \sum_{x\in X} p(x)\ln\left(\frac{p(x)}{(s\,\tilde{\circ}_{fin}\,q)(x)}\right).$$

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 $RE((f,s)) = S(p, s\,\tilde{\circ}\,q)$, where
$$S(p, s\,\tilde{\circ}\,q) := \int_X \log\left(\frac{\mathrm{d}p}{\mathrm{d}(s\,\tilde{\circ}\,q)}\right)\,\mathrm{d}p,$$
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.

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 [KL51]; 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:

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

It's easy to see that RE sends the identity arrows of **SbStat** to 0—the identity arrow of the unique object $\bullet$ of $[0,\infty]$. Hence, in order to show that RE is indeed a functor, it suffices to show that

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

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

Lemma 4.3. *Given an arrow $(f,s) : (X,p) \to (Y,q)$ in* **SbStat**. *Let $\{(s \,\tilde{\circ}\, q)_y\}_{y\in Y}$ be a disintegration of $(s \,\tilde{\circ}\, q)$ along f, then*

$$(s \,\tilde{\circ}\, q)_y = s_y \ q\text{-almost everywhere.}$$

We just have to show that $\{s_y\}_{y\in Y}$ satisfies the three properties implied by the disintegration theorem. We prove the third one; the first two being obvious.

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

$$\int_X h \, \mathrm{d}(s \,\tilde{\circ}\, q) = \int_{y\in Y} \int_{f^{-1}(y)} h \, \mathrm{d}s_y \, \mathrm{d}q.$$

Proof. Let's assume as a special case that h is the indicator function for a measurable set $E \subset X$. Then, we have

$$\int_X h \, \mathrm{d}(s \,\tilde{\circ}\, q) = \int_E \mathrm{d}(s \,\tilde{\circ}\, q) = (s \,\tilde{\circ}\, q)(E) = \int_{y\in Y} s_y(E) \, \mathrm{d}q = \int_{y\in Y} \int_{f^{-1}(y)} h \, \mathrm{d}s_y \, \mathrm{d}q.$$

We have shown that it is true for any indicator function. By linearity, it is true for any simple function and then, by the monotone convergence theorem, it is true for all Borel-measurable functions $h : X \to [0,\infty]$. □

Lemma 4.4. *The relative entropy is preserved under pre-composition by optimal hypotheses, i.e., for any $(g,t) : (Y,q) \to (Z,m)$ and $(f,s) : (X, s \,\tilde{\circ}\, q) \to (Y,q)$, we have*

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

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

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

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

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

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

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

The following calculation establishes the above.

$$\begin{aligned}
&\int_E \frac{\mathrm{d}q}{\mathrm{d}(t \mathbin{\tilde{\circ}} m)} \circ f\ \mathrm{d}(s \mathbin{\tilde{\circ}} t \mathbin{\tilde{\circ}} m) \\
&\quad = \int_Y \left(\int_{x \in f^{-1}(y) \cap E} \left(\frac{\mathrm{d}q}{\mathrm{d}(t \mathbin{\tilde{\circ}} m)} \circ f\right)(x)\ \mathrm{d}(s \mathbin{\tilde{\circ}} t \mathbin{\tilde{\circ}} m)_y\right)\ \mathrm{d}(t \mathbin{\tilde{\circ}} m) && (4.1)\\
&\quad = \int_Y \frac{\mathrm{d}q}{\mathrm{d}(t \mathbin{\tilde{\circ}} m)}(y) \left(\int_{f^{-1}(y) \cap E} \mathrm{d}(s \mathbin{\tilde{\circ}} t \mathbin{\tilde{\circ}} m)_y\right)\ \mathrm{d}(t \mathbin{\tilde{\circ}} m) && (4.2)\\
&\quad = \int_Y \frac{\mathrm{d}q}{\mathrm{d}(t \mathbin{\tilde{\circ}} m)}(y) \left(\int_{f^{-1}(y) \cap E} \mathrm{d}s_y\right)\ \mathrm{d}(t \mathbin{\tilde{\circ}} m) && (4.3)\\
&\quad = \int_Y \frac{\mathrm{d}q}{\mathrm{d}(t \mathbin{\tilde{\circ}} m)}(y) s_y(E \cap f^{-1}(y))\ \mathrm{d}(t \mathbin{\tilde{\circ}} m) \\
&\quad = \int_Y \frac{\mathrm{d}q}{\mathrm{d}(t \mathbin{\tilde{\circ}} m)}(y) s_y(E)\ \mathrm{d}(t \mathbin{\tilde{\circ}} m) && (4.4)\\
&\quad = \int_Y s_y(E)\ \mathrm{d}q && (4.5)\\
&\quad = (s \mathbin{\tilde{\circ}} q)(E) && (4.6)
\end{aligned}$$

We get (4.1) 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 (4.2) 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 (4.3) we apply Lemma 4.3. To show (4.4) we use the fact that s_y is a probability measure on $f^{-1}(y)$. We get (4.5) by the definition of the Radon-Nikodym derivative and we finally establish (4.6) by the definition of Kleisli composition.

Case II : (g, t) is not absolutely coherent. We have $RE((g, t)) = \infty$. We show that $(g \circ f, s \mathbin{\tilde{\circ}} t)$ is not absolutely coherent, i.e., $s \mathbin{\tilde{\circ}} q$ is not absolutely continuous with respect to $s \mathbin{\tilde{\circ}} t \mathbin{\tilde{\circ}} m$.

Since, by hypothesis, $q \ll t \mathbin{\tilde{\circ}} m$ doesn't hold, there exists a measurable set $B \subset Y$ such that $(t \mathbin{\tilde{\circ}} m)(B) = 0$ but $q(B) > 0$. We argue that $(s \mathbin{\tilde{\circ}} t \mathbin{\tilde{\circ}} m)(f^{-1}(B)) = 0$ and

$(s \mathbin{\tilde\circ} q)(f^{-1}(B)) > 0$. On one hand, we have

$$(s \mathbin{\tilde\circ} t \mathbin{\tilde\circ} m)(f^{-1}(B)) = \int_B s_y(f^{-1}(B))\mathrm{d}(t \mathbin{\tilde\circ} m) \leq (t \mathbin{\tilde\circ} m)(B) = 0.$$

But on the other hand, since f is a measure preserving map from $(X, s \mathbin{\tilde\circ} q)$ to (Y, q), we have $(s \mathbin{\tilde\circ} q)(f^{-1}(B)) = q(B) > 0$.

Therefore,

$$RE((g,t)) = \infty = RE((g \circ f, s \mathbin{\tilde\circ} t)). \qquad \square$$

Theorem 4.5 (Functoriality). *Given arrows $(f,s) : (X,p) \to (Y,q)$ and $(g,t) : (Y,q) \to (Z,m)$, we have*

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

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.

$$\begin{aligned}
RE\left((g \circ f, s \mathbin{\tilde\circ} t)\right) &= \int_X \log\left(\frac{\mathrm{d}p}{\mathrm{d}(s \mathbin{\tilde\circ} t \mathbin{\tilde\circ} m)}\right)\,\mathrm{d}p \\
&= \int_X \log\left(\frac{\mathrm{d}p}{\mathrm{d}(s \mathbin{\tilde\circ} q)}\frac{\mathrm{d}(s \mathbin{\tilde\circ} q)}{\mathrm{d}(s \mathbin{\tilde\circ} t \mathbin{\tilde\circ} m)}\right)\,\mathrm{d}p \qquad (4.7)\\
&= \int_X \log\left(\frac{\mathrm{d}p}{\mathrm{d}(s \mathbin{\tilde\circ} q)}\right)\,\mathrm{d}p + \int_X \log\left(\frac{\mathrm{d}(s \mathbin{\tilde\circ} q)}{\mathrm{d}(s \mathbin{\tilde\circ} t \mathbin{\tilde\circ} m)}\right)\,\mathrm{d}p \\
&= RE((f,s)) + \int_X \log\left(\frac{\mathrm{d}(s \mathbin{\tilde\circ} q)}{\mathrm{d}(s \mathbin{\tilde\circ} t \mathbin{\tilde\circ} m)}\right)\,\mathrm{d}p \\
&= RE((f,s)) + RE((g,t)) \qquad (4.8)
\end{aligned}$$

We get (4.7) by the chain rule for Radon-Nikodym derivatives and (4.8) by applying Lemma 4.4.

Case II : (g,t) is not absolutely coherent. We argue that $(g \circ f, s \mathbin{\tilde\circ} t)$ is not absolutely coherent. By hypothesis, $q \ll t \mathbin{\tilde\circ} m$ doesn't hold, so there is a measurable set $B \subset Y$ such that $(t \mathbin{\tilde\circ} m)(B) = 0$ and $q(B) > 0$. We show that $(s \mathbin{\tilde\circ} t \mathbin{\tilde\circ} m)\left(f^{-1}(B)\right) = 0$ and $p(f^{-1}(B)) > 0$. On one hand, we have

$$(s \mathbin{\tilde\circ} t \mathbin{\tilde\circ} m)\left(f^{-1}(B)\right) = \int_B s_y\left(f^{-1}(B)\right)\,\mathrm{d}(t \mathbin{\tilde\circ} m) \leq (t \mathbin{\tilde\circ} m)(B) = 0,$$

but on the other hand, we have $p(f^{-1}(B)) = q(B) > 0$. Therefore

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

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

This case 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 $(s \mathbin{\tilde\circ} t \mathbin{\tilde\circ} m)(A) = 0$.

By the hypothesis of (f,s) not being absolutely coherent, $p \ll s \mathbin{\tilde{\circ}} q$ doesn't hold, so there is a measurable set $A \subset X$ such that $(s \mathbin{\tilde{\circ}} q)(A) = 0$ and $p(A) > 0$.

We partition A into

$$A_\epsilon := \{x \in A \mid s_{f(x)}(A) > 0\} \text{ and } A_0 := \{x \in A \mid s_{f(x)}(A) = 0\}$$

and we partition Y into

$$B_\epsilon := \{y \in Y \mid s_y(A) > 0\} \text{ and } B_0 := \{y \in Y \mid s_y(A) = 0\}.$$

We argue that $(s \mathbin{\tilde{\circ}} t \mathbin{\tilde{\circ}} m)(A_0) = 0$ and $p(A_0) > 0$.

Since $A_0 \subset f^{-1}(B_0)$, $f^{-1}(B_\epsilon)$ is disjoint from A_0, so for all $y \in B_\epsilon$ we have $s_y(A_0) = 0$ because their support is disjoint from A_0. On one hand, we thus have

$$\begin{aligned}
(s \mathbin{\tilde{\circ}} t \mathbin{\tilde{\circ}} m)(A_0) &= \int_Y s_y(A_0)\, \mathrm{d}(t \mathbin{\tilde{\circ}} m) \\
&= \int_{B_0} s_y(A_0)\, \mathrm{d}(t \mathbin{\tilde{\circ}} m) + \int_{B_\epsilon} s_y(A_0)\, \mathrm{d}(t \mathbin{\tilde{\circ}} m) \\
&= \int_{B_0} s_y(A_0)\, \mathrm{d}(t \mathbin{\tilde{\circ}} m) \\
&\leq \int_{B_0} s_y(A)\, \mathrm{d}(t \mathbin{\tilde{\circ}} m) \\
&= 0.
\end{aligned}$$

On the other hand, since we have $p(A_0) + p(A_\epsilon) = p(A) > 0$ and $A_\epsilon \subset f^{-1}(B_\epsilon)$, it suffices to show $p(f^{-1}(B_\epsilon)) = 0$ to conclude $p(A_0) > 0$.

By hypothesis, we have

$$(s \mathbin{\tilde{\circ}} q)(A) = \int_{B_0} s_y(A)\, \mathrm{d}q + \int_{B_\epsilon} s_y(A)\, \mathrm{d}q = 0,$$

so $q(B_\epsilon) = 0$ and because f is measure preserving, we have $p(f^{-1}(B_\epsilon)) = q(B_\epsilon) = 0$ as desired.

So $(g \circ f, s \mathbin{\tilde{\circ}} t)$ is not absolutely coherent, hence

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

This completes the proof of this case. □

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 [BF14] 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 4.3 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 $\{y\}$.

Definition 4.6. 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. Before proving that RE is convex linear, we first prove the following lemma.

Lemma 4.7. *Given*

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

where f is a measurable map preserving the measure of both Borel probability measures p and p'. If $p \ll p'$, then $\frac{\mathrm{d}p_y}{\mathrm{d}p'_y}$ is defined for q-almost every y and

$$\frac{\mathrm{d}p_y}{\mathrm{d}p'_y}(x) = \frac{\mathrm{d}p}{\mathrm{d}p'}(x) \ \ p'\text{-almost everywhere.}$$

Proof. For an arbitrary measurable function $h : X \to [0, \infty]$, by first applying the Radon-Nikodym theorem and then the disintegration theorem on the measurable function $h\frac{\mathrm{d}p}{\mathrm{d}p'}$, we get

$$\int_X h \,\mathrm{d}p = \int_X h\frac{\mathrm{d}p}{\mathrm{d}p'} \,\mathrm{d}p' = \int_Y \int_{f^{-1}(y)} h\frac{\mathrm{d}p}{\mathrm{d}p'} \,\mathrm{d}p'_y \,\mathrm{d}q.$$

Hence, for q-almost every y, we must have $\frac{\mathrm{d}p_y}{\mathrm{d}p'_y}(x) = \frac{\mathrm{d}p}{\mathrm{d}p'}(x)$ p'-almost everywhere. □

Theorem 4.8 (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.**

We have

$$\begin{aligned} RE((f,s)) &= \int_X \log\left(\frac{\mathrm{d}p}{\mathrm{d}(s \,\tilde{\circ}\, q)}\right) \mathrm{d}p \\ &= \int_Y \int_{f^{-1}(y)} \log\left(\frac{\mathrm{d}p}{\mathrm{d}(s \,\tilde{\circ}\, q)}\right) \mathrm{d}p_y \, \mathrm{d}q \qquad (4.9) \\ &= \int_Y \int_{f^{-1}(y)} \log\left(\frac{\mathrm{d}p_y}{\mathrm{d}(s \,\tilde{\circ}\, q)_y}\right) \mathrm{d}p_y \, \mathrm{d}q \qquad (4.10) \\ &= \int_Y RE((f,s)_y) \, \mathrm{d}q. \end{aligned}$$

We get (4.9) by the disintegration theorem and (4.10) by applying Lemma 4.7.

Case II : (f, s) is not absolutely coherent. By the hypothesis of (f, s) not being absolutely coherent, there is a measurable set $A \subset X$ such that $(s \,\tilde{\circ}\, q)(A) = 0$ and $p(A) > 0$. Applying lemma 4.3, on one hand we have

$$\int_Y (s \,\tilde{\circ}\, q)_y(A) \, \mathrm{d}q = \int_Y s_y(A) \, \mathrm{d}q = (s \,\tilde{\circ}\, q)(A) = 0,$$

but on the other hand we have

$$\int_Y p_y(A) \, \mathrm{d}q = p(A) > 0.$$

Hence, the subset of Y on which $p_y \ll (s \,\tilde{\circ}\, q)_y$ doesn't hold contains a set of measure strictly greater than 0. Therefore,

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

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 $\lim_{n\to\infty} p_n(E) = p(E)$.

The singleton set equipped with the trivial measure, which we will denote by $(1, \delta)$, is a weakly terminal object of **SbStat**, it is weakly terminal in the sense that for every (X, p) there exist a non-unique arrow $(f, s) : (X, p) \to (1, \delta)$ in **SbStat**.

Definition 4.9. A functor F from **SbStat** to $[0, \infty]$ is *lower semi-continuous* if for every arrow $(f, s) : (X, p) \to (1, \delta)$, whenever $p_n \to p$ and $s_n \to s$, then

$$F\Big((X,p) \underset{f}{\overset{s}{\rightleftarrows}} (1,\delta)\Big) \leq \liminf_{n\to\infty} F\Big((X,p_n) \underset{f}{\overset{s_n}{\rightleftarrows}} (1,\delta)\Big).$$

Recall that in [BF14], lower semicontinuity was defined on **FinStat** as the following.

Definition 4.10 (Baez and Fritz)**.** A functor $F : \mathbf{FinStat} \to [0, \infty]$ is *lower semicontinuous* if for any sequence of morphisms $(f, s_i) : (X, p_i) \to (Y, q_i)$ that converges[4] to a morphism $(f, s) : (X, p) \to (Y, q)$, we have

$$F(f, s) \leq \liminf_{i\to\infty} F(f, s_i).$$

Recalling that **FP** stands for the subcategory of **SbStat** consisting of the same objects, but with only those morphisms where the hypothesis is optimal. We claim that a lower semi-continuous (as defined in Definition 4.9) functor F that vanishes on **FP** restricts to a lower semi-continuous functor on **FinStat** (as defined in Definition 4.10). To see this, note that, given a sequence of morphisms $(f, s_i) : (X, p_i) \to (Y, q_i)$ that converges pointwise to a morphism $(f, s) : (X, p) \to (Y, q)$, we can recover

$$F(f, s) \leq \liminf_{i\to\infty} F(f, s_i)$$

from

$$(X, p_i) \underset{f}{\overset{s_i}{\rightleftarrows}} (Y, q_i) \underset{g}{\overset{q_i}{\rightleftarrows}} (1, \delta)$$

and

$$\begin{aligned} F(f, s) &= F(f, s) + \overbrace{F(g, q)}^{0} = F(g \circ f, s \,\tilde{\circ}\, q) \\ &\leq \liminf_{i\to\infty} F(g \circ f, s_i \,\tilde{\circ}\, q_i) = \liminf_{i\to\infty} F(f, s_i) + \liminf_{i\to\infty} \underbrace{F(g, q_i)}_{0} \\ &= \liminf_{i\to\infty} F(f, s_i). \end{aligned}$$

Note that, on finite sets, converging pointwise is equivalent to strong convergence.

Theorem 4.11 (Lower semi-continuity)**.** *The functor RE is lower semi-continuous.*

Proof. Let us denote

$$a := \liminf_{n\to\infty} RE\left((X, p_n) \underset{f}{\overset{s_n}{\rightleftarrows}} (1, \delta) \right).$$

If $a = \infty$, then the statement holds automatically, so we assume that $a < \infty$.

By virtue of a being a limit inferior, we can pick a subsequence $\{n_i\}_{i\in\mathbb{N}}$ such that for all $i \in \mathbb{N}$, we have both

$$RE\left((X, p_{n_i}) \underset{f}{\overset{s_{n_i}}{\rightleftarrows}} (1, \delta) \right) = \int_X \log \left(\frac{\mathrm{d}p_{n_i}}{\mathrm{d}s_{n_i}} \right) \mathrm{d}p_{n_i} < \infty$$

[4]Where convergence is just pointwise convergence.

and

$$\lim_{i\to\infty} RE\left((X, p_{n_i}) \underset{f}{\overset{s_{n_i}}{\rightleftarrows}} (1,\delta) \right) = a.$$

Now, instantiating statements (2.4.7) and (2.4.9) from Pinsker [Pin60, Section 2.4][5] in our setting, we have

$$RE\left((X, p) \underset{f}{\overset{s}{\rightleftarrows}} (1,\delta) \right) \leq \lim_{i\to\infty} F\left((X, p_{n_i}) \underset{f}{\overset{s_{n_i}}{\rightleftarrows}} (1,\delta) \right) = a,$$

as desired. □

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 $\{0, 1, \dots, n2^n - 1, \leq\}$ 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

[5] Or equivalently, and perhaps a more accessible reference, Theorem 1 from [Pos75].

$q(X_{n,k_n})$, we also have $k_n 2^{-n} \leq \frac{p(X_{n,k_n})}{q(X_{n,k_n})} \leq (k_n+1)2^{-n}$ 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

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

Before proving uniqueness, we recall the main theorem of Baez and Fritz [BF14] 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**.

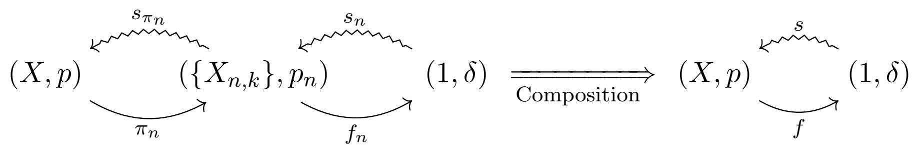
By convex linearity of F, for an arbitrary morphism (f, s) from (X, p) to (Y, q), we have

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

so F is totally described by its local relative entropies. It is thus sufficient to show $F = cRE$ on an arbitrary morphism $(f, s) : (X, p) \to (1, \delta)$. The case where p is not absolutely continuous with respect to s is straightforward, so let us assume $p \ll s$.

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. Presented as diagrams, we have



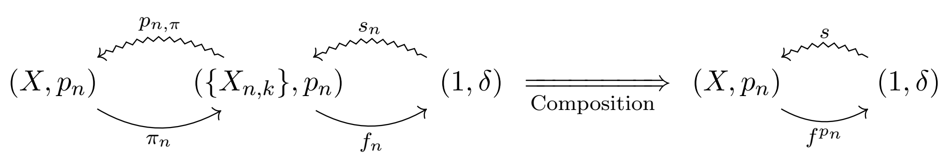
From the above diagram and the hypothesis that F is a functor, we have the following inequality

$$F((f_n, s_n)) \leq F((f, s)), \text{ for all } n \in \mathbb{N}. \tag{5.1}$$

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

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

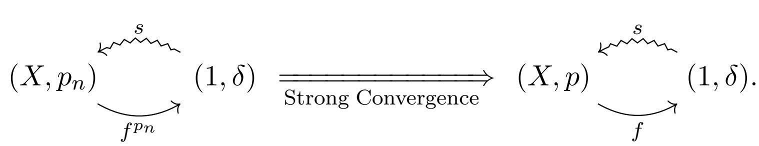
This means that $p_{n,\pi}$ is the disintegration of s along π_n. Presented as diagrams, where we use f^{p_n} instead of f to indicate that the arrow leaves from the object (X, p_n) as opposed to (X, p), we have



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

$$F((f^{p_n}, s)) = F((\pi_n, p_{n,\pi})) + F((f_n, s_n)) = F((f_n, s_n)). \tag{5.2}$$

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



Hence, combining (5.2) with the lower semicontinuity of F, we also have the inequality

$$F((f, s)) \leq \liminf_{n\to\infty} F((f^{p_n}, s)) = \liminf_{n\to\infty} F((f_n, s_n)). \tag{5.3}$$

Since (f_n, s_n) is in **FinStat**, we must have $F((f_n, s_n)) = cRE((f_n, s_n))$. Thus, combining (5.1) and (5.3), 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. [BFL11, BF14]. 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 [DG15, CDDG17, DDG16] 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 [CDDG17] adopts a point-free approach introduced in [CDPP09, CDPP14]. It would be interesting to extend our definitions to a point-free situation.

References

[Ash72] R. B. Ash. *Real Analysis and Probability*. Academic Press, 1972.

[BF14] John C. Baez and Tobias Fritz. A bayesian characterization of relative entropy. *Theory and Applications of Categories*, 29(16):422–456, 2014.

[BFL11] John C. Baez, Tobias Fritz, and Tom Leinster. A characterization of entropy in terms of information loss. *Entropy*, 13(11):1945–1957, 2011.

[Bil95] P. Billingsley. *Probability and Measure*. Wiley-Interscience, 1995.

[CDDG17] Florence Clerc, Vincent Danos, Fredrik Dahlqvist, and Ilias Garnier. Pointless learning. In *Proceedings of FoSSaCS 2017*, 2017.

[CDPP09] Philippe Chaput, Vincent Danos, Prakash Panangaden, and Gordon Plotkin. Approximating Markov processes by averaging. In *Proceedings of the 37th International Colloquium On Automata Languages And Programming (ICALP)*, volume 5556 of *Lecture Notes In Computer Science*, pages 127–138, 2009.

[CDPP14] Philippe Chaput, Vincent Danos, Prakash Panangaden, and Gordon Plotkin. Approximating Markov processes by averaging. *J. ACM*, 61(1):5:1–5:45, January 2014. `doi:10.1145/2537948`.

[DDG16] Fredrik Dahlqvist, Vincent Danos, and Ilias Garnier. Giry and the machine. *Electronic Notes in Theoretical Computer Science*, 325:85–110, 2016.

[DG15] Vincent Danos and Ilias Garnier. Dirichlet is natural. *Electronic Notes in Theoretical Computer Science*, 319:137–164, 2015.

[DSDG18] Fredrik Dahlqvist, Alexandra Silva, Vincent Danos, and Ilias Garnier. Borel kernels and their approximation, categorically. *Electronic Notes in Theoretical Computer Science*, 341:91–119, 2018.

[Dud89] R. M. Dudley. *Real Analysis and Probability*. Wadsworth and Brookes/Cole, 1989.

[Gir81] M. Giry. A categorical approach to probability theory. In B. Banaschewski, editor, *Categorical Aspects of Topology and Analysis*, number 915 in Lecture Notes In Mathematics, pages 68–85. Springer-Verlag, 1981.

[KL51] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[Law64] F. W. Lawvere. The category of probabilistic mappings. Unpublished typescript., 1964.

[Law73] F. William Lawvere. Metric spaces, generalized logic and closed categories. *Rend. Sem. Mat. Fis. Milano*, 43(1):135–166, 1973.

[Lei] Tom Leinster. An operadic introduction to entropy. n-category cafe.

[Pin60] Mark S Pinsker. Information and information stability of random variables and processes. *Holden-Day 1964*, 1960.

[Pos75] Edward Posner. Random coding strategies for minimum entropy. *IEEE Transactions on Information Theory*, 21(4):388–391, 1975.

[Rok49] Vladimir Abramovich Rokhlin. On the fundamental ideas of measure theory. *Matematicheskii Sbornik*, 67(1):107–150, 1949.