A BAYESIAN CHARACTERIZATION
OF RELATIVE ENTROPY

JOHN C. BAEZ AND TOBIAS FRITZ

ABSTRACT. We give a new characterization of relative entropy, also known as the Kullback–Leibler divergence. We use a number of interesting categories related to probability theory. In particular, we consider a category `FinStat` where an object is a finite set equipped with a probability distribution, while a morphism is a measure-preserving function $f \colon X \to Y$ together with a stochastic right inverse $s \colon Y \to X$. The function f can be thought of as a measurement process, while s provides a hypothesis about the state of the measured system given the result of a measurement. Given this data we can define the entropy of the probability distribution on X relative to the 'prior' given by pushing the probability distribution on Y forwards along s. We say that s is 'optimal' if these distributions agree. We show that any convex linear, lower semicontinuous functor from `FinStat` to the additive monoid $[0, \infty]$ which vanishes when s is optimal must be a scalar multiple of this relative entropy. Our proof is independent of all earlier characterizations, but inspired by the work of Petz.

1. Introduction

This paper gives a new characterization of the concept of relative entropy, also known as 'relative information', 'information gain' or 'Kullback-Leibler divergence'. Whenever we have two probability distributions p and q on the same finite set X, we define the information of q relative to p as:

$$S(q, p) = \sum_{x \in X} q_x \ln \left(\frac{q_x}{p_x} \right)$$

Here we set $q_x \ln(q_x/p_x)$ equal to ∞ when $p_x = 0$, unless q_x is also zero, in which case we set it equal to 0. Relative entropy thus takes values in $[0, \infty]$.

Intuitively speaking, $S(q, p)$ is the expected amount of information gained when we discover the probability distribution is really q, when we had thought it was p. We should

think of p as a 'prior'. When we take p to be the uniform distribution on X, relative entropy reduces to the ordinary Shannon entropy, up to a sign and an additive constant. The advantage of relative entropy is that it makes the role of the prior explicit.

Since Bayesian probability theory emphasizes the role of the prior, relative entropy naturally lends itself to a Bayesian interpretation [3]. Our goal here is to make this precise in a mathematical characterization of relative entropy. We do this using a category `FinStat` where:

- an object (X, q) consists of a finite set X and a probability distribution $x \mapsto q_x$ on that set;

- a morphism $(f, s) \colon (X, q) \to (Y, r)$ consists of a measure-preserving function f from X to Y, together with a probability distribution $x \mapsto s_{xy}$ on X for each element $y \in Y$ with the property that $s_{xy} = 0$ unless $f(x) = y$.

We can think of an object of `FinStat` as a system with some finite set of **states** together with a probability distribution on its states. A morphism $(f, s) \colon (X, q) \to (Y, r)$ then consists of two parts. First, there is a deterministic 'measurement process' $f \colon X \to Y$ mapping states of some system being measured to states of a 'measurement apparatus'. The condition that f be measure-preserving says that the probability that the apparatus winds up in some state $y \in Y$ is the sum of the probabilities of states of X leading to that outcome:

$$r_y = \sum_{x \colon f(x) = y} q_x.$$

Second, there is a 'hypothesis' s: an assumption about the probability s_{xy} that the system being measured is in the state x given any measurement outcome $y \in Y$. We assume that this probability vanishes unless $f(x) = y$, as we would expect from a hypothesis made by someone who knew the behavior of the measurement apparatus.

Suppose we have any morphism $(f, s) \colon (X, q) \to (Y, r)$ in `FinStat`. From this we obtain two probability distributions on the states of the system being measured. First, we have the probability distribution $p \colon X \to \mathbb{R}$ given by

$$p_x = s_{x\,f(x)} r_{f(x)}. \tag{1}$$

This is our 'prior', given our hypothesis and the probability distribution of measurement outcomes. Second, we have the 'true' probability distribution $q \colon X \to \mathbb{R}$. It follows that any morphism in `FinStat` has a relative entropy $S(q, p)$ associated to it. This is the expected amount of information we gain when we update our prior p to q.

In fact, this way of assigning relative entropies to morphisms defines a functor

$$\mathrm{RE} \colon \mathtt{FinStat} \to [0, \infty]$$

where we use $[0, \infty]$ to denote the category with one object, the nonnegative real numbers together with ∞ as morphisms, and addition as composition. More precisely, if $(f, s) \colon (X, q) \to (Y, r)$ is any morphism in `FinStat`, we define

$$\mathrm{RE}(f, s) = S(q, p)$$

where the prior p is defined as in Equation (1). The fact that RE is a functor is nontrivial and rather interesting. It says that given any composable pair of measurement processes:

$$(X, q) \xrightarrow{(f,s)} (Y, r) \xrightarrow{(g,t)} (Z, u)$$

the relative entropy of their composite is the sum of the relative entropies of the two parts:

$$\mathrm{RE}((g, t) \circ (f, s)) = \mathrm{RE}(g, t) + \mathrm{RE}(f, s).$$

We prove that RE is a functor in Section 3. However, we go much further: we characterize relative entropy by saying that up to a constant multiple, RE is the unique functor from `FinStat` to $[0, \infty]$ obeying three reasonable conditions.

The first condition is that RE vanishes on morphisms $(f, s) \colon (X, q) \to (Y, r)$ where the hypothesis s is 'optimal'. By this, we mean that Equation (1) gives a prior p equal to the 'true' probability distribution q on the states of the system being measured.

The second condition is that RE is lower semicontinuous. The set $P(X)$ of probability distributions on a finite set X naturally has the topology of an $(n-1)$-simplex when X has n elements. The set $[0, \infty]$ can be given the topology induced by the usual order on this set, and it is then homeomorphic to a closed interval. However, with these topologies, the relative entropy does not define a continuous function

$$\begin{aligned} S \colon P(X) \times P(X) &\to [0, \infty] \\ (q, p) &\mapsto S(q, p). \end{aligned}$$

The problem is that

$$S(q, p) = \sum_{x \in X} q_x \ln \left(\frac{q_x}{p_x} \right)$$

and $q_x \ln(q_x/p_x)$ equals ∞ when $p_x = 0$ and $q_x > 0$, but 0 when $p_x = q_x = 0$. So, it turns out that S is only lower semicontinuous, meaning that it can suddenly jump down, but not up. More precisely, if $p^i, q^i \in P(X)$ are sequences with $p^i \to p$, $q^i \to q$, then

$$S(q, p) \le \liminf_{i \to \infty} S(q^i, p^i).$$

In Section 3 we give the set of morphisms in `FinStat` a topology, and show that with this topology, RE maps morphisms to morphisms in a lower semicontinuous way.

The third condition is that RE is convex linear. In Section 3 we describe how to take convex linear combinations of morphisms in `FinStat`. The functor RE is convex linear in the sense that it maps any convex linear combination of morphisms in `FinStat` to the corresponding convex linear combination of numbers in $[0, \infty]$. Intuitively, this means that if we flip a probability-λ coin to decide whether to perform one measurement process or another, the expected information gained is λ times the expected information gain of the first process plus $(1 - \lambda)$ times the expected information gain of the second.

Our main result is Theorem 3.1: any lower semicontinuous, convex linear functor

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

that vanishes on morphisms with an optimal hypothesis must equal some constant times the relative entropy. In other words, there exists some constant $c \in [0, \infty]$ such that

$$F(f, s) = c \, \mathrm{RE}(f, s)$$

for any morphism $(f, s) \colon (X, p) \to (Y, q)$ in FinStat.

This theorem, and its proof, was inspired by results of Petz [8], who sought to characterize relative entropy both in the 'classical' case discussed here and in the more general 'quantum' setting. Our original intent was merely to express his results in a more category-theoretic framework. Unfortunately his work contained a flaw, which we had to repair. As a result, our proof is now self-contained. For details, see the remarks after Theorem 5.2.

Our characterization of relative entropy implicitly relies on topological categories and on the operad whose operations are convex linear combinations. However, since these structures are not strictly necessary for stating or proving our result, and they may be unfamiliar to some readers, we discuss them only in Appendix A and Appendix B.

2. The categories in question

FINSTOCH. To describe the categories used in this paper, we need to start with a word on the category of finite sets and stochastic maps. A stochastic map $f \colon X \rightsquigarrow Y$ is different from an ordinary function, because instead of assigning a unique element of Y to each element of X, it assigns a *probability distribution* on Y to each element of X. Thus $f(x)$ is not a specific element of Y, but instead has a probability of taking on different values. This is why we use a wiggly arrow to denote a stochastic map.

More formally:

2.1. DEFINITION. *Given finite sets X and Y, a* **stochastic map** *$f \colon X \rightsquigarrow Y$ assigns a real number f_{yx} to each pair $x \in X, y \in Y$ in such a way that fixing any element x, the numbers f_{yx} form a probability distribution on Y. We call f_{yx}* **the probability of y given x.**

In more detail, we require that the numbers f_{yx} obey:

- $f_{yx} \geq 0$ for all $x \in X$, $y \in Y$,

- $\displaystyle\sum_{y \in Y} f_{yx} = 1$ for all $x \in X$.

Note that we can think of $f \colon X \rightsquigarrow Y$ as a $Y \times X$-shaped matrix of numbers. A matrix obeying the two properties above is called **stochastic**. This viewpoint is nice because it reduces the problem of composing stochastic maps to matrix multiplication. It is easy to

check that multiplying two stochastic matrices gives a stochastic matrix. So, we define the composite of stochastic maps $f\colon X \rightsquigarrow Y$ and $g\colon Y \rightsquigarrow Z$ by

$$(g \circ f)_{zx} = \sum_{y \in Y} g_{zy} f_{yx}.$$

Since matrix multiplication is associative and identity matrices are stochastic, this construction gives a category:

2.2. DEFINITION. *Let* FinStoch *be the category of finite sets and stochastic maps between them.*

We are restricting attention to finite sets merely to keep the discussion simple and avoid issues of convergence. It would be interesting to generalize all our work to more general probability spaces.

FINPROB. Choose any 1-element set and call it **1**. A function $f\colon \mathbf{1} \to X$ is just a point of X. But a stochastic map $q\colon \mathbf{1} \rightsquigarrow X$ is something more interesting: it is a probability distribution on X.

We use the term **finite probability measure space** to mean a finite set with a probability distribution on it. As we have just seen, there is a very quick way to describe such a thing within FinStoch:

$$
\begin{array}{c}
\mathbf{1} \\
{\scriptstyle q} \Big\downarrow \\
X
\end{array}
$$

This gives a quick way to think about a measure-preserving function between finite probability measure spaces! It is simply a commutative triangle like this:

$$
\begin{array}{ccc}
 & \mathbf{1} & \\
{\scriptstyle q} \swarrow & & \searrow {\scriptstyle r} \\
X & \xrightarrow{\ \ f\ \ } & Y
\end{array}
$$

Note that the horizontal arrow $f\colon X \to Y$ is not wiggly. The straight arrow means it is an honest function, not a stochastic map. But a function can be seen as a special case of a stochastic map. So it makes sense to compose a straight arrow with a wiggly arrow—and the result is, in general, a wiggly arrow. If we then demand that the above triangle commute, this says that the function $f\colon X \to Y$ is measure-preserving.

We now work through the details. First: how can we see a function as a special case of a stochastic map? A function $f\colon X \to Y$ gives a matrix of numbers

$$f_{yx} = \delta_{y\, f(x)}$$

where δ is the Kronecker delta. This matrix is stochastic, and it defines a stochastic map sending each point $x \in X$ to the probability distribution supported at $f(x)$.

Given this, we can see what the commutativity of the above triangle means. If we use q_x to stand for the probability that $q \colon 1 \rightsquigarrow X$ assigns to each element $x \in X$, and similarly for r_y, then the triangle commutes if and only if

$$r_y = \sum_{x \in X} \delta_{y\, f(x)} q_x$$
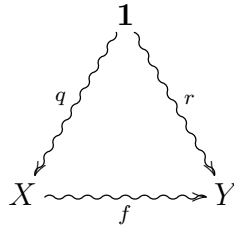
or in other words:

$$r_y = \sum_{x \colon f(x) = y} q_x$$

In this situation we say p is q **pushed forward along** f, and that f is a **measure-preserving function**.

So, we have used `FinStoch` to describe another important category:

2.3. DEFINITION. *Let* `FinProb` *be the category of finite probability measure spaces and measure-preserving functions between them.*

Another variation may be useful at times:

A commuting triangle like this is a **measure-preserving stochastic map**. In other words, q gives a probability measure on X, r gives a probability measure on Y, and $f \colon X \rightsquigarrow Y$ is a stochastic map that is measure-preserving in the following sense:

$$r_y = \sum_{x \in X} f_{yx} q_x.$$

FINSTAT. The category we need for our characterization of relative entropy is a bit more subtle. In this category, an object is a finite probability measure space:
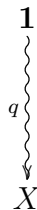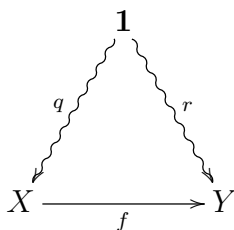
but a morphism looks like this:

$$\begin{array}{c} \mathbf{1} \\[2em] \end{array}$$
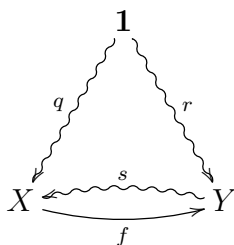
$$
\begin{aligned}
f \circ q &= r \\
f \circ s &= 1_Y
\end{aligned}
$$

The diagram need not commute, but the two equations shown must hold. The first equation says that $f \colon X \to Y$ is a measure-preserving function. In other words, this triangle, which we have seen before, commutes:

The second equation says that $f \circ s$ is the identity, or in other words, s is a 'section' for f. This requires a bit of discussion.

We can think of X as the set of 'states' of some system, while Y is a set of possible states of some other system: a 'measuring apparatus'. The function f is a 'measurement process'. One 'measures' the system using f, and if the system is in any state $x \in X$ the measuring apparatus goes into the state $f(x)$. The probability distribution q gives the probability that the system is in any given state, while r gives the probability that the measuring apparatus ends up in any given state after a measurement is made.

Under this interpretation, we think of the stochastic map s as a 'hypothesis' about the system's state given the state of the measuring apparatus. If one measures the system and the apparatus goes into the state $y \in Y$, this hypothesis asserts that the system is in the state x with probability s_{xy}.

The equation $f \circ s = 1_Y$ says that if the measuring apparatus ends up in some state $y \in Y$, our hypothesis assigns a nonzero probability only to states of the measured system for which a measurement actually leads to this state y:

2.4. LEMMA. *If $f \colon X \to Y$ is a function between finite sets and $s \colon Y \rightsquigarrow X$ is a stochastic map, then $f \circ s = 1_Y$ if and only for all $y \in Y$, $s_{xy} = 0$ unless $f(x) = y$.*

PROOF. The condition $f \circ s = 1_Y$ says that for any fixed $y, y' \in Y$,

$$\sum_{x:\, f(x)=y'} s_{xy} = \sum_{x \in X} \delta_{y'\, f(x)} s_{xy} = \delta_{y'y}.$$

It follows that the sum at left vanishes if $y' \neq y$. If s is stochastic, the terms in this sum are nonnegative. So, s_{xy} must be zero if $f(x) = y'$ and $y' \neq y$.

Conversely, suppose we have a stochastic map $s \colon Y \rightsquigarrow X$ such that $s_{xy} = 0$ unless $f(x) = y$. Then for any $y \in Y$ we have

$$1 = \sum_{x \in X} s_{xy} = \sum_{x:\, f(x)=y} s_{xy} = \sum_{x \in X} \delta_{y\, f(x)} s_{xy}$$

while for $y' \neq y$ we have

$$0 = \sum_{x:\, f(x)=y'} s_{xy} = \sum_{x \in X} \delta_{y'\, f(x)} s_{xy},$$

so for all $y, y' \in Y$

$$\sum_{x \in X} \delta_{y'\, f(x)} s_{xy} = \delta_{y'y},$$
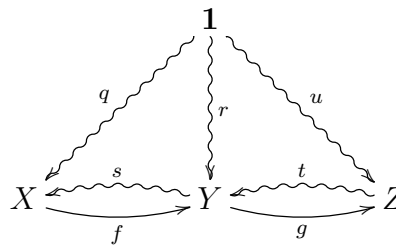
which says that $f \circ s = 1_Y$.   ∎

It is also worth noting that $f \circ s = 1_Y$ implies that f is onto: if $y \in Y$ were not in the image of f, we could not have

$$\sum_{x \in X} s_{xy} = 1$$

as required, since $s_{xy} = 0$ unless $f(x) = y$. So, the equation $f \circ s = 1_Y$ also rules out the possibility that our measuring apparatus has 'extraneous' states that never arise when we make a measurement.

This is how we compose morphisms of the above sort:

$$\begin{aligned}
f \circ q &= r & g \circ r &= u \\
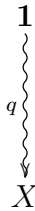f \circ s &= 1_Y & g \circ t &= 1_Z
\end{aligned}$$

We get a measure-preserving function $g \circ f \colon X \to Z$ and a stochastic map going back, $s \circ t \colon Z \to X$. It is easy to check that these obey the required equations:
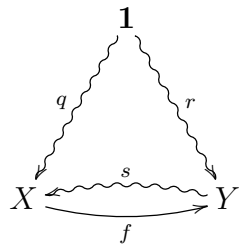
$$g \circ f \circ q = u$$

$$g \circ f \circ s \circ t = 1_Z$$

So, this way of composing morphisms gives a category, which we call `FinStat`, to allude to its role in statistical reasoning:

2.5. DEFINITION. *Let* `FinStat` *be the category where an object is a finite probability measure space:*

$$
\begin{array}{c}
\mathbf{1} \\
\Big\downarrow q \\
X
\end{array}
$$

a morphism is a diagram

$$
\begin{array}{c}
\mathbf{1} \\
{}^{q}\swarrow \quad \searrow^{r} \\
X \underset{f}{\overset{s}{\rightleftarrows}} Y
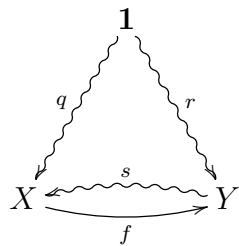\end{array}
$$

obeying these equations:

$$
\begin{aligned}
f \circ q &= r \\
f \circ s &= 1_Y
\end{aligned}
$$

and composition is defined as above.

FP. We have described how to think of a morphism in `FinStat` as consisting of a 'measurement process' f and a 'hypothesis' s, obeying two equations:

$$
\begin{array}{c}
\mathbf{1} \\
{}^{q}\swarrow \quad \searrow^{r} \\
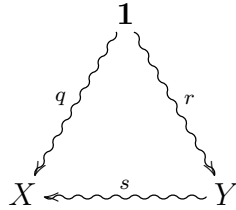X \underset{f}{\overset{s}{\rightleftarrows}} Y
\end{array}
$$

$$
\begin{aligned}
f \circ q &= r \\
f \circ s &= 1_Y
\end{aligned}
$$

We say the hypothesis is **optimal** if also

$$s \circ r = q.$$

Conceptually, this says that if we take the probability distribution r on our observations and use it to infer a probability distribution for the system's state using our hypothesis s, we get the correct answer: q. Mathematically, it says that this diagram commutes:


In other words, s is a measure-preserving stochastic map.

It is easy to check that this optimality property is preserved by composition of morphisms. Hence there is a subcategory of FinStat with all the same objects, but only morphisms where the hypothesis is optimal:

2.6. DEFINITION. *Let* FP *be the subcategory of* FinStat *where an object is a finite probability measure space*
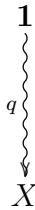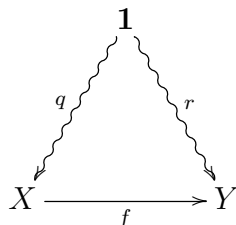

and a morphism is a diagram
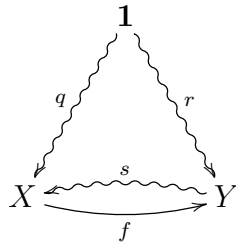
obeying these equations:

$$
\begin{aligned}
f \circ q &= r \\
f \circ s &= 1_Y \\
s \circ r &= q
\end{aligned}
$$

The category FP was introduced by Leinster [5]. He gave it this name for two reasons. First, it is a close relative of FinProb, where a morphism looks like this:

We now explain the similarities and differences between FP and FinProb by studying the properties of the forgetful functor FP → FinProb, which sends every morphism (f, s) to its underlying measure-preserving function f.

For a morphism in FP, the conditions on s are so strong that they completely determine it, unless there are states of the measurement apparatus that happen with probability zero: that is, unless there are $y \in Y$ with $r_y = 0$. To see this, note that

$$s \circ r = q$$

says that

$$\sum_{y \in Y} s_{xy} r_y = q_x$$

for any choice of $x \in X$. But we have already seen in Lemma 2.4 that $s_{xy} = 0$ unless $f(x) = y$, so the sum has just one term, and the equation says

$$s_{xy} r_y = q_x$$

where $y = f(x)$. We can solve this for s_{xy} unless $r_y = 0$. Furthermore, we have already seen that every $y \in Y$ is of the form $f(x)$ for some $x \in X$.

Thus, for a morphism $(f, s) \colon (X, q) \to (Y, r)$ in FP, we can solve for s in terms of the other data unless there exists $y \in Y$ with $r_y = 0$. Except for this special case, a morphism in FP is just a morphism in FinProb. But in this special case, a morphism in FP has a little extra information: an arbitrary probability distribution on the inverse image of each point y with $r_y = 0$. The point is that in FinStat, and thus FP, a 'hypothesis' must provide a probability for each state of the system given a state of the measurement apparatus, even for states of the measurement apparatus that occur with probability zero.

A more mathematical way to describe the situation is that our functor FP → FinProb is 'generically' full and faithful: the function

$$\begin{array}{ccc} \mathrm{FP}((X, q), (Y, r)) & \longrightarrow & \mathrm{FinProb}((X, q), (Y, r)) \\ (f, s) & \mapsto & f \end{array}$$

is a bijection if the support of r is the whole set Y, which is the generic situation.

The second reason Leinster called this category FP is that it is freely formed from an operad called P. This is a topological operad whose n-ary operations are probability distributions on the set $\{1, \dots, n\}$. These operations describe convex linear combinations, so algebras of this operad include convex subsets of $\mathbb{R}^n$, more general convex spaces [2], and even more. As Leinster explains [5], the category FP (or more precisely, an equivalent one) is the free P-algebra among categories containing an internal P-algebra. We will not need this fact here, but it is worth mentioning that Leinster used this fact to characterize entropy as a functor from FP to $[0, \infty)$. He and the authors then rephrased this in simpler language [1], obtaining a characterization of entropy as a functor from FinProb to $[0, \infty)$. The characterization of relative entropy in the current paper is a closely related result. However, the proof is completely different.

3. Characterizing entropy

THE THEOREM. We begin by stating our main result. Then we clarify some of the terms involved and begin the proof.

3.1. THEOREM. *Relative entropy determines a functor*

$$\mathrm{RE}\colon \mathtt{FinStat} \quad \to \quad [0,\infty]$$

$$\left((X,q) \overset{s}{\underset{f}{\rightsquigarrow}} (Y,r) \right) \quad \mapsto \quad S(q, s \circ r) \tag{2}$$

that is lower semicontinuous, convex linear, and vanishes on morphisms in the subcategory FP.

Conversely, these properties characterize the functor RE *up to a scalar multiple. In other words, if F is another functor with these properties, then for some $0 \le c \le \infty$ we have $F(f,s) = c\,\mathrm{RE}(f,s)$ for all morphisms (f,s) in* FinStat. *(Here we define $\infty \cdot a = a \cdot \infty = \infty$ for $0 < a \le \infty$, but $\infty \cdot 0 = 0 \cdot \infty = 0$.)*

In the rest of this section we begin by describing $[0,\infty]$ as a category and checking that RE is a functor. Then we describe what it means for the functor RE to be lower semicontinuous and convex linear, and check these properties. We postpone the hard part of the proof, in which we characterize RE up to a scalar multiple by these properties, to Section 4.

In what follows, it will be useful to have an explicit formula for $S(q, s \circ r)$. By definition,

$$S(q, s \circ r) = \sum_{x \in X} q_x \ln \left(\frac{q_x}{(s \circ r)_x} \right)$$

We have

$$(s \circ r)_x = \sum_{y \in Y} s_{xy} r_y,$$

but by Lemma 2.4, $s_{xy} = 0$ unless $f(x) = y$, so the sum has just one term:

$$(s \circ r)_x = s_{x\,f(x)} r_{f(x)}$$

and we obtain

$$S(q, s \circ r) = \sum_{x \in X} q_x \ln \left(\frac{q_x}{s_{x\,f(x)} r_{f(x)}} \right). \tag{3}$$

FUNCTORIALITY. We make $[0,\infty]$ into a monoid using addition, where we define addition in the usual way for numbers in $[0,\infty)$ and set

$$\infty + a = a + \infty = \infty$$

for all $a \in [0,\infty]$. There is thus a category with one object and elements of $[0,\infty]$ as endomorphisms of this object, with composition of morphisms given by addition. With a slight abuse of language we also use $[0,\infty]$ to denote this category.

3.2. LEMMA. *The map* RE: `FinStat` $\to [0, \infty]$ *described in Theorem 3.1 is a functor.*

PROOF. Let

$$(X, q) \underset{f}{\overset{s}{\rightleftarrows}} (Y, r) \underset{g}{\overset{t}{\rightleftarrows}} (Z, u)$$

be a composable pair of morphisms in `FinStat`. Then the functoriality of RE can be shown by repeated use of Equation (3):

$$\mathrm{RE}\,(g \circ f, s \circ t) \;=\; S\,(q, s \circ t \circ u)$$

$$= \;\sum_{x \in X} q_x \ln \left(\frac{q_x}{s_{x\,f(x)} t_{f(x)\,g(f(x))} u_{g(f(x))}} \right)$$

$$\overset{(*)}{=} \;\sum_{x \in X} q_x \ln \left(\frac{q_x}{s_{x\,f(x)} r_{f(x)}} \right) + \sum_{x \in X} q_x \ln \left(\frac{r_{f(x)}}{t_{f(x)\,g(f(x))} u_{g(f(x))}} \right)$$

$$= \;S(q, s \circ r) + \sum_{y \in Y} r_y \ln \left(\frac{r_y}{t_{y\,g(y)} u_{g(y)}} \right)$$

$$= \;S(q, s \circ r) + S(r, t \circ u)$$

$$= \;\mathrm{RE}(f, s) + \mathrm{RE}(g, t).$$

Here the main step is $(*)$, where we have simply inserted

$$0 = \sum_x q_x \ln \frac{1}{r_{f(x)}} + \sum_x q_x \ln r_{f(x)}.$$

This is unproblematic as long as $r_{f(x)} > 0$ for all x. When there are x with $r_{f(x)} = 0$, then we necessarily have $q_x = 0$ as well, and both $q_x \ln \frac{1}{r_{f(x)}}$ and $q_x \ln r_{f(x)}$ actually vanish, so this case is also fine. In the step after $(*)$, we use the fact that for each $y \in Y$, r_y is the sum of q_x over all x with $f(x) = y$. ∎

LOWER SEMICONTINUITY. Next we explain what it means for a functor to be lower semicontinuous, and prove that RE has this property. There is a way to think about semicontinuous functors in terms of topological categories, but this is not really necessary for our work, so we postpone it to Appendix A. Here we take a more simple-minded approach.

If we fix two finite sets X and Y, the set of all morphisms

$$(f, s) \colon (X, q) \to (Y, p)$$

in `FinStat` forms a topological space in a natural way. To see this, let

$$P(X) = \{q \colon X \to [0, 1] \colon \sum_{x \in X} q_x = 1\}$$

be the set of probability distributions on a finite set X. This is a subset of a finite-dimensional real vector space, so we give it the subspace topology. With this topology, $P(X)$ is homeomorphic to a simplex. The set of stochastic maps $s\colon Y \rightsquigarrow X$ is also a subspace of a finite-dimensional real vector space, namely the space of matrices $\mathbb{R}^{X \times Y}$, so we also give it the subspace topology. We then give $P(X) \times P(Y) \times \mathbb{R}^{X \times Y}$ the product topology. The set of morphisms $(f, s)\colon (X, q) \to (Y, p)$ in FinStat can be seen as a subspace of this, and we give it the subspace topology. We then say:

3.3. DEFINITION. *A functor* $F\colon \mathtt{FinStat} \to [0, \infty]$ *is* **lower semicontinuous** *if for any sequence of morphisms* $(f, s^i)\colon (X, q^i) \to (Y, r^i)$ *that converges to a morphism* $(f, s)\colon (X, q)$ $\to (Y, r)$, *we have*

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

We could use nets instead of sequences here, but it would make no difference. We can then check another part of our main theorem:

3.4. LEMMA. *The functor* RE$\colon \mathtt{FinStat} \to [0, \infty]$ *described in Theorem 3.1 is lower semicontinuous.*

PROOF. Suppose that $(f, s^i)\colon (X, q^i) \to (Y, r^i)$ is a sequence of morphisms in FinStat that converges to $(f, s)\colon (X, q) \to (Y, r)$. We need to show that

$$S(q, s \circ r) \leq \liminf_{i \to \infty} S(q^i, s^i \circ r^i).$$

If there is no $x \in X$ with $s_{x\,f(x)} r_{f(x)} = 0$ then this is clear, since all the elementary functions involved in the definition of relative entropy are continuous away from 0. If all $x \in X$ with $s_{x\,f(x)} = 0$ also satisfy $q_x = 0$, then $S(q, s \circ r)$ is still finite since none of these x contribute to the sum for S. In this case $S(q^i, s^i \circ r^i)$ may remain arbitrarily large, even infinite as $i \to \infty$. But the inequality

$$S(q, s \circ r) \leq \liminf_{i \to \infty} S(q^i, s^i \circ r^i)$$

remains true. The same argument applies if there are $x \in X$ with $r_{f(x)} = 0$, which implies $q_x = 0$. Finally, if there are $x \in X$ with $s_{x\,f(x)} = 0$ but $r_{f(x)} \geq q_x > 0$, then $S(q, s \circ r) = \infty$. The above inequality is still valid in this case. ∎

That lower semicontinuity of relative entropy is an important property was already known to Petz; see the closing remark in [8].

CONVEX LINEARITY. Next we explain what it means to say that relative entropy gives a convex linear functor from FinProb to $[0, \infty]$, and we prove this is true. In general, convex linear functors go between convex categories. These are topological categories equipped with an action of the operad P discussed by Leinster [5]. Since we do not need the general theory here, we postpone it to Appendix B.

First, note that there is a way to take convex linear combinations of objects and morphisms in `FinProb`. Let (X, p) and (Y, q) be finite sets equipped with probability measures, and let $\lambda \in [0, 1]$. Then there is a probability measure

$$\lambda p \oplus (1 - \lambda) q$$

on the disjoint union $X + Y$, whose value at a point x is given by

$$(\lambda p \oplus (1 - \lambda) q)_x = \begin{cases} \lambda p_x & \text{if } x \in X \\ (1 - \lambda) q_x & \text{if } x \in Y. \end{cases}$$

Given a pair of morphisms

$$f \colon (X, p) \to (X', p'), \qquad g \colon (Y, q) \to (Y', q')$$

in `FinProb`, there is a unique morphism

$$\lambda f \oplus (1 - \lambda) g \colon (X + Y, \lambda p \oplus (1 - \lambda) q) \quad \to \quad (X' + Y', \lambda p' \oplus (1 - \lambda) q')$$

that restricts to f on X and to g on Y.

A similar construction applies to `FinStat`. Given a pair of morphisms

$$(X, p) \underset{f}{\overset{s}{\rightleftarrows}} (X', p') \qquad (Y, q) \underset{g}{\overset{t}{\rightleftarrows}} (Y', q')$$

in `FinStat`, we define their convex linear combination to be

$$(X + Y, \lambda p \oplus (1 - \lambda) q) \underset{\lambda f \oplus (1-\lambda) g}{\overset{s \oplus t}{\rightleftarrows}} (X' + Y', \lambda p' \oplus (1 - \lambda) q')$$

where $s \oplus t \colon X' + Y' \rightsquigarrow X + Y$ is the stochastic map which restricts to s on X' and t on Y'. As a stochastic matrix, it is of block-diagonal form. It is right inverse to $\lambda f \oplus (1 - \lambda) g$ by construction.

We may also define convex linear combinations of objects and morphisms in the category $[0, \infty]$. Since this category has only one object, there is only one way to define convex linear combinations of objects. Morphisms in this category are elements of the set $[0, \infty]$. We have already made this set into a monoid using addition. We can also introduce multiplication, defined in the usual way for numbers in $[0, \infty)$, and with

$$0a = a0 = 0$$

for all $a \in [0, \infty]$. This gives meaning to the convex linear combination $\lambda a + (1 - \lambda) b$ of two morphisms a, b in $[0, \infty]$. For more details, see Appendices A and B.

3.5. DEFINITION. *A functor $F\colon \mathtt{FinStat} \to [0,\infty]$ is* **convex linear** *if it preserves convex combinations of objects and morphisms.*

For objects this requirement is trivial, so all this really means is that for any pair of morphisms (f,s) and (g,t) in $\mathtt{FinStat}$ and any $\lambda \in [0,1]$, we have

$$F\left(\lambda(f,s) \oplus (1-\lambda)(g,t)\right) = \lambda F(f,s) + (1-\lambda)F(g,t).$$

3.6. LEMMA. *The functor* RE$\colon \mathtt{FinStat} \to [0,\infty]$ *described in Theorem 3.1 is convex linear.*

PROOF. This follows from a direct computation:

$$\mathrm{RE}((\lambda(f,s)\oplus(1-\lambda)(g,t)) = S(\lambda p \oplus (1-\lambda)q, \lambda s \circ p' \oplus (1-\lambda)t \circ q')$$

$$= \sum_{x\in X} \lambda p_x \ln\left(\frac{\lambda p_x}{s_{x\,f(x)} \cdot \lambda p'_{f(x)}}\right) + \sum_{y\in Y}(1-\lambda)q_y \ln\left(\frac{(1-\lambda)q_y}{t_{y\,g(y)} \cdot (1-\lambda)q'_{g(y)}}\right)$$

$$= \lambda \sum_{x\in X} p_x \ln\left(\frac{p_x}{s_{x\,f(x)}p'_{f(x)}}\right) + (1-\lambda)\sum_{y\in Y}\ln\left(\frac{q_y}{t_{y\,g(y)}q'_y}\right)$$

$$= \lambda S(p, s \circ p') + (1-\lambda)S(q, t \circ q')$$

$$= \lambda\,\mathrm{RE}(f,s) + (1-\lambda)\,\mathrm{RE}(g,t)$$

∎

4. Proof of the theorem
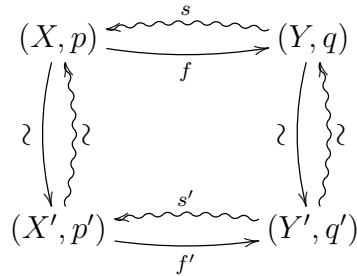
Now we prove the main part of Theorem 3.1.

4.1. LEMMA. *Suppose that a functor*

$$F\colon \mathtt{FinStat} \to [0,\infty]$$

is lower semicontinuous, convex linear, and vanishes on morphisms in the subcategory $\mathtt{FP}$. *Then for some $0 \le c \le \infty$ we have $F(f,s) = c\,\mathrm{RE}(f,s)$ for all morphisms (f,s) in* $\mathtt{FinStat}$.

PROOF. Let $F\colon \mathtt{FinStat} \to [0,\infty]$ be any functor satisfying these hypotheses. By functoriality and the fact that 0 is the only morphism in $[0,\infty]$ with an inverse, F vanishes on isomorphisms. Thus, given any commutative square in $\mathtt{FinStat}$ where the vertical

morphisms are isomorphisms:

$$
\begin{array}{ccc}
(X,p) & \underset{f}{\overset{s}{\rightleftharpoons}} & (Y,q) \\
\wr \big\downarrow \wr & & \wr \big\downarrow \wr \\
(X',p') & \underset{f'}{\overset{s'}{\rightleftharpoons}} & (Y',q')
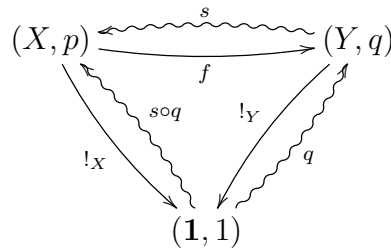\end{array}
$$

functoriality implies that F takes the same value on the top and bottom morphisms:

$$F(f,s) = F(f',s').$$

So, in what follows, we can replace an object by an isomorphic object without changing the value of F on morphisms from or to this object.

Given any morphism in $\mathtt{FinStat}$, complete it to a diagram of this form:

$$
\begin{array}{ccc}
(X,p) & \underset{f}{\overset{s}{\rightleftharpoons}} & (Y,q) \\
& {}_{!_X}\searrow \quad {}^{s\circ q} \downarrow \quad {}^{!_Y} \nearrow \ \downarrow {}_{q} & \\
& (\mathbf{1},1) &
\end{array}
$$

Here $\mathbf{1}$ denotes any one-element set equipped with the unique probability measure 1, and $!_X\colon X \to \mathbf{1}$ is the unique function, which is automatically measure-preserving since p is assumed to be normalized. Since this diagram commutes, and the morphism on the lower right lies in $\mathtt{FP}$, we obtain

$$
F\left((X,p) \underset{f}{\overset{s}{\rightleftharpoons}} (Y,q) \right) = F\left((X,p) \underset{!_X}{\overset{s\circ q}{\rightleftharpoons}} (\mathbf{1},1) \right).
$$

In other words: the value of F on a morphism depends only on the two distributions p and $s \circ q$ living on the domain of the morphism. For this reason, it is enough to prove the claim only for those morphisms whose codomain is $(\mathbf{1},1)$.

We now consider the family of distributions

$$q(\alpha) = (\alpha, 1-\alpha),$$

on a two-element set $\mathbf{2} = \{0,1\}$, and consider the function

$$
g(\alpha) = F\left((\mathbf{2},q(1)) \underset{!_\mathbf{2}}{\overset{q(\alpha)}{\rightleftharpoons}} (\mathbf{1},1) \right) \tag{4}
$$

for $\alpha \in [0,1]$. Note that for all $\beta \in [0,1)$, this square in `FinStat` commutes:

$$
\begin{array}{ccc}
(\mathbf{3},(1,0,0)) & \xrightarrow{\;\; q(\beta)\oplus 1 \;\;} & (\mathbf{2},(1,0)) \\
& \begin{array}{c} 0,1\mapsto 0 \\ 2\mapsto 1 \end{array} & \\
\begin{array}{c} 0\mapsto 0 \\ 1,2\mapsto 1 \end{array} \Big\downarrow \quad 1\oplus q\!\left(\frac{\alpha(1-\beta)}{1-\alpha\beta}\right) \quad !_{\mathbf{2}} & & \Big\downarrow q(\alpha) \\
& q(\alpha\beta) & \\
(\mathbf{2},(1,0)) & \xrightarrow{\;\;\;\;\;\;\;\;} & (\mathbf{1},1) \\
& !_{\mathbf{2}} &
\end{array}
$$

where the left vertical morphism is in `FP`, while the top horizontal morphism is the convex linear combination

$$
1\left((\mathbf{2},q(1)) \xrightarrow[\;!_{\mathbf{2}}\;]{\;q(\beta)\;} (\mathbf{1},1) \right) \oplus 0 \left(1_{(\mathbf{1},1)} \right).
$$

Applying the functoriality and convex linearity of F to this square, we thus obtain the equation

$$
g(\alpha\beta) = g(\alpha) + g(\beta). \tag{5}
$$

We claim that all solutions of this equation are of the form $g(\alpha) = -c\ln\alpha$ for some $c \in [0,\infty]$. First we show this for $\alpha \in (0,1]$.

If $g(\alpha) < \infty$ for all $\alpha \in (0,1]$, this equation is Cauchy's functional equation in its multiplicative-to-additive form, and it is known [7] that any solution with g measurable is of the desired form for some $c < \infty$. By our hypotheses on F, g is lower semicontinuous, hence measurable. Thus, for some $c < \infty$ we have $g(\alpha) = -c\ln\alpha$ for all $\alpha \in (0,1]$.

If $g(\alpha) = \infty$ for some $\alpha \in (0,1]$, then Equation (5) implies that $g(\beta) = \infty$ for all $\beta < \alpha$. Since it also implies that $g(2\beta) = \frac{1}{2}g(\beta)$, we conclude that then $g(\beta) = \infty$ for all $\beta \in (0,1)$. Thus, if we take $c = \infty$ we again have $g(\alpha) = -c\ln\alpha$ for all $\alpha \in (0,1]$.

Next consider $\alpha = 0$. If $c > 0$, then $g(0) = g(0) + g(\frac{1}{2})$ shows that we necessarily have $g(0) = \infty$. If $c = 0$, then lower semicontinuity implies $g(0) = 0$. In both cases, the equation $g(\alpha) = -c\ln\alpha$ also holds for $\alpha = 0$.

In what follows, choosing the value of c that makes $g(\alpha) = -c\ln\alpha$, we shall prove that the equation

$$
F\left((X,p) \xrightarrow[\;!_X\;]{\;r\;} (\mathbf{1},1) \right) = c\,S(p,r)
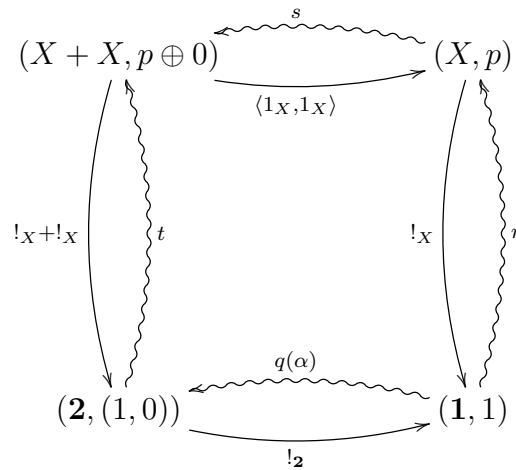$$

holds for any two probability distributions p and r on any finite set X. Using Equation (3), it suffices to show that

$$
F\left((X,p) \xrightarrow[\;!_X\;]{\;r\;} (\mathbf{1},1) \right) = c\sum_{x\in X} p_x \ln\left(\frac{p_x}{r_x}\right). \tag{6}
$$

We prove this for more and more general cases in the following series of lemmas. We start with the generic case, where $c < \infty$ and the probability distribution r has full support. In Lemma 4.4 we treat all cases with $0 < c < \infty$. In Lemma 4.5 we treat the case $c = 0$, and in Lemma 4.12 we treat the case $c = \infty$, which seems much harder than the rest. ∎

4.2. LEMMA. *Equation* (6) *holds if* $c < \infty$ *and the support of* r *is all of* X.

PROOF. Choose $\alpha \in (0,1)$ such that $\alpha < r_x$ for all $x \in X$. The decisive step is to consider the commutative square

$$
\begin{array}{ccc}
(X+X, p \oplus 0) & \underset{\langle 1_X, 1_X \rangle}{\overset{s}{\rightleftarrows}} & (X,p) \\
{\scriptstyle !_X + !_X} \Big\updownarrow {\scriptstyle t} & & {\scriptstyle !_X} \Big\updownarrow {\scriptstyle r} \\
(\mathbf{2}, (1,0)) & \underset{!_2}{\overset{q(\alpha)}{\rightleftarrows}} & (\mathbf{1},1)
\end{array}
$$

where the stochastic matrices s and t are given by

$$
s = \begin{pmatrix} \alpha \frac{p_1}{r_1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \alpha \frac{p_n}{r_n} \\ 1 - \alpha \frac{p_1}{r_1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & 1 - \alpha \frac{p_n}{r_n} \end{pmatrix}, \qquad t = \begin{pmatrix} p_1 & \frac{r_1 - \alpha p_1}{1-\alpha} \\ \vdots & \vdots \\ p_n & \frac{r_n - \alpha p_n}{1-\alpha} \end{pmatrix}.
$$

The second column of t is only relevant for commutativity. The left vertical morphism is in FP, while we already know that the lower horizontal morphism evaluates to $g(\alpha) = -c \ln \alpha$ under the functor F. Hence the diagonal of the square gets assigned the value $-c \ln \alpha$ under F. On the other hand, the upper horizontal morphism is actually a convex linear combination of morphisms

$$
(\mathbf{2}, (1,0)) \underset{!_2}{\overset{q\left(\alpha \frac{p_x}{r_x}\right)}{\rightleftarrows}} (\mathbf{1},1) \,,
$$

one for each $x \in X$, with the probabilities p_x as coefficients. Thus, composing this with the right vertical morphism we get a morphism to which F assigns the value

$$-c \sum_{x \in X} p_x \ln \left(\alpha \frac{p_x}{r_x} \right) + F \left((X, p) \underset{!_X}{\overset{r}{\rightleftarrows}} (\mathbf{1}, 1) \right).$$

Thus, we obtain

$$-c \sum_{x \in X} p_x \ln \left(\alpha \frac{p_x}{r_x} \right) + F \left((X, p) \underset{!_X}{\overset{r}{\rightleftarrows}} (\mathbf{1}, 1) \right) = -c \ln \alpha$$

and because $c < \infty$, we can simplify this to

$$F \left((X, p) \underset{!_X}{\overset{r}{\rightleftarrows}} (\mathbf{1}, 1) \right) = c \sum_{x \in X} p_x \ln \left(\frac{p_x}{r_x} \right)$$

This is the desired result, Equation (6). ∎

4.3. LEMMA. *Equation* (6) *holds if* $c < \infty$ *and* $\mathrm{supp}(p) \subseteq \mathrm{supp}(r)$.

PROOF. This can be reduced to the previous case by considering the commutative triangle
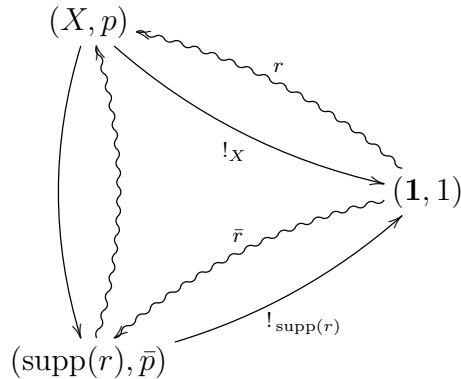
in which $\bar{p} = p|_{\mathrm{supp}(r)}$ and $\bar{r} = r|_{\mathrm{supp}(r)}$, and the vertical morphism consists of any map $X \to \mathrm{supp}(r)$ that restricts to the identity on $\mathrm{supp}(r)$ and, as its stochastic right inverse, the inclusion $\mathrm{supp}(r) \hookrightarrow X$. This morphism lies in FP. ∎

4.4. LEMMA. *Equation* (6) *holds if* $0 < c < \infty$.

PROOF. We already know by Lemma 4.3 that this holds when $\mathrm{supp}(p) \subseteq \mathrm{supp}(r)$, so assume otherwise. Our task is then show that

$$F \left((X, p) \underset{!_X}{\overset{r}{\rightleftarrows}} (\mathbf{1}, 1) \right) = \infty.$$

To do this, choose $x \in X$ with $p_x > 0 = r_x$, and consider the commutative triangle
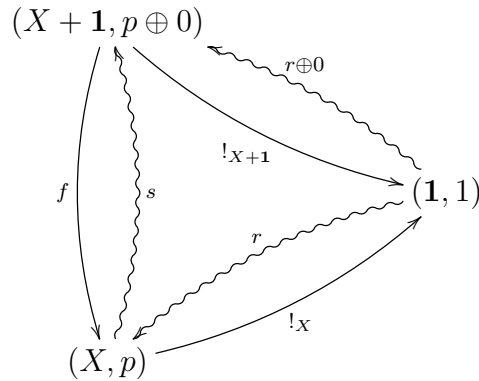
in which f maps X to itself by the identity and sends the unique element of $\mathbf{1}$ to x. This function has a one-parameter family of stochastic right inverses, and we take the arrow $s \colon X \rightsquigarrow X + \mathbf{1}$ to be any element of this family.

To construct these stochastic right inverses, let $Y = X - \{x\}$. This set is nonempty because the probability distribution r is supported on it. If $p_x < 1$ let q be the probability distribution on Y given by

$$q = \frac{1}{1 - p_x} p|_Y,$$

while if $p_x = 1$ let q be an arbitrary probability distribution on Y. For any $\alpha \in [0, 1]$, the convex linear combination

$$(1 - p_x)\left((Y, q) \underset{1_Y}{\overset{1_Y}{\rightleftharpoons}} (Y, q)\right) \oplus p_x\left((\mathbf{2}, (1, 0)) \underset{!_{\mathbf{2}}}{\overset{q(\alpha)}{\rightleftharpoons}} (\mathbf{1}, 1)\right) \qquad (7)$$

is a morphism in $\mathtt{FinStat}$. There is a natural isomorphism from its domain to that of the desired morphism (f, s):

$$(1 - p_x)(Y, q) \ \oplus \ p_x(\mathbf{2}, (1, 0)) \ \cong \ (X + \mathbf{1}, p \oplus 0)$$

and similarly for its codomain:

$$(1 - p_x)(Y, q) \ \oplus \ p_x(\mathbf{1}, 1) \ \cong \ (X, p).$$

Composing (7) with these fore and aft, we obtain the desired morphism

$$(X + \mathbf{1}, p \oplus 0) \underset{s}{\overset{f}{\rightleftharpoons}} (X, p) \,.$$

Using convex linearity and the fact that F vanishes on isomorphisms, (7) implies that $F(f, s) = -p_x c \ln \alpha$. Applying F to our commutative triangle, we thus obtain

$$F\left((X + \mathbf{1}, p \oplus 0) \underset{!_{X+\mathbf{1}}}{\overset{r \oplus 0}{\rightleftharpoons}} (\mathbf{1}, 1)\right) = -p_x c \ln \alpha + F\left((X, p) \underset{!_X}{\overset{r}{\rightleftharpoons}} (\mathbf{1}, 1)\right).$$

Since $p_x, c > 0$, the first term on the right-hand side depends on α, but no other terms do. This is only possible if both other terms are infinite. This proves

$$F\left((X,p) \xleftarrow[\;!_X\;]{\;r\;} (\mathbf{1},1)\right) = \infty,$$

as was to be shown. ∎

4.5. LEMMA. *Equation* (6) *holds if $c = 0$.*

PROOF. That (6) holds in this case is a simple consequence of lower semicontinuity: approximate r by a family of probability distributions whose support is all of X. By Lemma 4.4, F maps all the resulting morphisms to 0. Thus, the same must be true for the original r. ∎

To conclude the proof of Lemma 4.1, we need to show Equation (6) holds if $c = \infty$. To do this, it suffices to assume $c = \infty$ and show that

$$F\left((X,p) \xleftarrow[\;!_X\;]{\;r\;} (\mathbf{1},1)\right) = \infty$$

whenever $p \neq r$. The reasoning in the previous lemmas will not help us now, since in Lemma 4.2 we needed $c < \infty$. As we shall see in Proposition 5.1, the proof for $c = \infty$ must use lower semicontinuity. However, since lower semicontinuity only produces an *upper* bound on the value of F at a limit point, it will have to be used in proving the contrapositive statement: if F is finite on some morphism of the above form with $p \neq r$, then it is finite on some morphism of the form (4). Now in order to infer that the value of F at the limit point of a converging family of distributions is finite, it is not enough to know that the value of F is finite at each element of the family: one needs a *uniform* bound. The need to derive such a uniform bound is the reason for the complexity of the following argument.

In what follows we assume that p and r are probability distributions on X with $p \neq r$ and

$$F\left((X,p) \xleftarrow[\;!_X\;]{\;r\;} (\mathbf{1},1)\right) < \infty.$$

We develop a series of consequences culminating in Lemma 4.12, in which we see that $g(\alpha)$ is finite for some $\alpha < 1$. This implies $c < \infty$, thus demonstrating the contrapositive of our claim that Equation (6) holds if $c = \infty$.

4.6. LEMMA. *There exist $\alpha, \beta \in [0,1]$ with $\alpha \neq \beta$ such that*

$$h(\alpha,\beta) = F\left((\mathbf{2},q(\alpha)) \xleftarrow[\;!_\mathbf{2}\;]{\;q(\beta)\;} (\mathbf{1},1)\right) \tag{8}$$
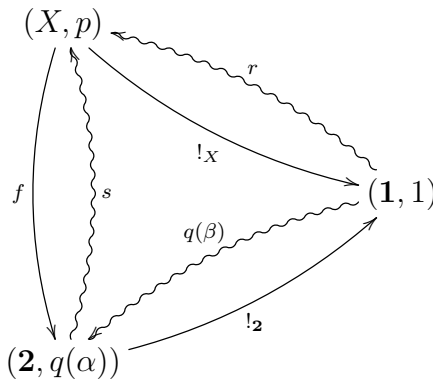
is finite.

PROOF. Choose some $y \in X$ with $p_y \neq r_y$, and define $f : X \to \mathbf{2}$ by

$$f(x) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y. \end{cases}$$

Put $\beta = 1 - r_y$. Then f has a stochastic right inverse s given by

$$s_{xj} = \begin{cases} \dfrac{r_x}{\beta}(1 - \delta_{xy}) & \text{if } j = 0 \\ \delta_{xy} & \text{if } j = 1 \end{cases}$$
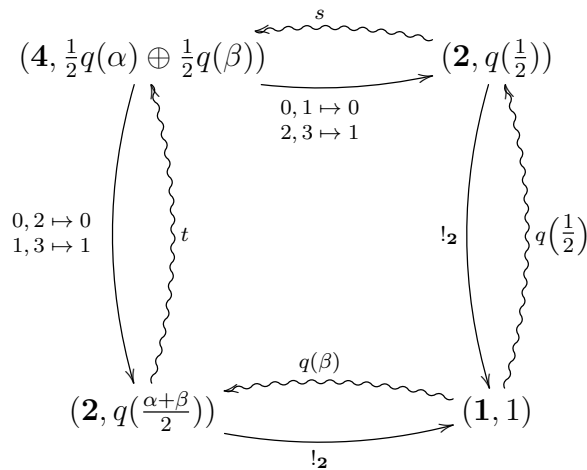
where, if $\beta = 0$, we interpret the fractions as forming an arbitrarily chosen probability distribution on $X - \{y\}$. Setting $\alpha = 1 - p_y$, we have a commutative triangle

and the claim follows from functoriality.   ∎

4.7. LEMMA. $h(\alpha', \frac{1}{2})$ is finite for some $\alpha' < \frac{1}{2}$.

PROOF. Choose α, β as in Lemma 4.6. Consider the commutative square

with the stochastic matrices

$$
s = \begin{pmatrix} \beta & 0 \\ 1-\beta & 0 \\ 0 & \beta \\ 0 & 1-\beta \end{pmatrix} = q(\beta) \oplus q(\beta), \qquad t = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}.
$$

The right vertical morphism in this square lies in FP, so F vanishes on this. The top horizontal morphism is a convex linear combination

$$
\frac{1}{2}\left((\mathbf{2}, q(\alpha)) \xleftarrow[!_{\mathbf{2}}]{q(\beta)} (\mathbf{1}, 1) \right) \oplus \frac{1}{2}\left((\mathbf{2}, q(\beta)) \xleftarrow[!_{\mathbf{2}}]{q(\beta)} (\mathbf{1}, 1) \right),
$$

where the second term is in FP. Thus, by convex linearity and Lemma 4.6, F of the top horizontal morphism equals $\frac{1}{2}h(\alpha, \beta) < \infty$. By functoriality, F is $\frac{1}{2}h(\alpha, \beta)$ on the composite of the top and right morphisms.

This implies that the value of F on the other two morphisms in the square must also be finite. Let us compute F of their composite in another way. By definition, F of the bottom horizontal morphism is $h(\frac{\alpha+\beta}{2}, \beta)$. The left vertical morphism is a convex linear combination

$$
\frac{\alpha+\beta}{2}\left((\mathbf{2}, q(\tfrac{\alpha}{\alpha+\beta})) \xleftarrow[!_{\mathbf{2}}]{q(\frac{1}{2})} (\mathbf{1}, 1) \right) \oplus \frac{2-\alpha-\beta}{2}\left((\mathbf{2}, q(\tfrac{1-\alpha}{2-\alpha-\beta})) \xleftarrow[!_{\mathbf{2}}]{q(\frac{1}{2})} (\mathbf{1}, 1) \right).
$$

By functoriality and convex linearity, F on the composite of these two morphisms is thus

$$
\frac{\alpha+\beta}{2} \cdot h\left(\frac{\alpha}{\alpha+\beta}, \frac{1}{2} \right) + \frac{2-\alpha-\beta}{2} \cdot h\left(\frac{1-\alpha}{2-\alpha-\beta}, \frac{1}{2} \right) + h\left(\frac{\alpha+\beta}{2}, \beta \right).
$$

Comparing these computations, we obtain

$$
\begin{aligned}
h(\alpha, \beta) = {} & (\alpha+\beta) \cdot h\left(\frac{\alpha}{\alpha+\beta}, \frac{1}{2} \right) \\
& + (2-\alpha-\beta) \cdot h\left(\frac{1-\alpha}{2-\alpha-\beta}, \frac{1}{2} \right) + 2 \cdot h\left(\frac{\alpha+\beta}{2}, \beta \right).
\end{aligned}
\tag{9}
$$

This shows that each term on the right-hand side must be finite. Note that the coefficients in front of these terms do not vanish, since $\alpha \neq \beta$. If $\alpha < \beta$ then we can take $\alpha' = \frac{\alpha}{\alpha+\beta}$, so that $\alpha' < \frac{1}{2}$, and the first term on the right-hand side gives $h(\alpha', \frac{1}{2}) < \infty$. If $\alpha > \beta$ we can take $\alpha' = \frac{1-\alpha}{2-\alpha-\beta}$, so that $\alpha' < \frac{1}{2}$, and the second term on the right-hand side gives that $h(\alpha', \frac{1}{2}) < \infty$. ∎

4.8. LEMMA. *For $\alpha \leq \beta \leq \frac{1}{2}$, we have $h(\beta, \frac{1}{2}) \leq h(\alpha, \frac{1}{2})$.*

PROOF. By the intermediate value theorem, there exists $\gamma \in [0, 1]$ with

$$\gamma\alpha + (1 - \gamma)(1 - \alpha) = \beta.$$

Now let $q(\alpha) \otimes q(\gamma)$ stand for the distribution on **4** with weights $(\alpha\gamma, \alpha(1-\gamma), (1-\alpha)\gamma, (1-\alpha)(1-\gamma))$. The equation above guarantees that the left vertical morphism in this square is well-defined:
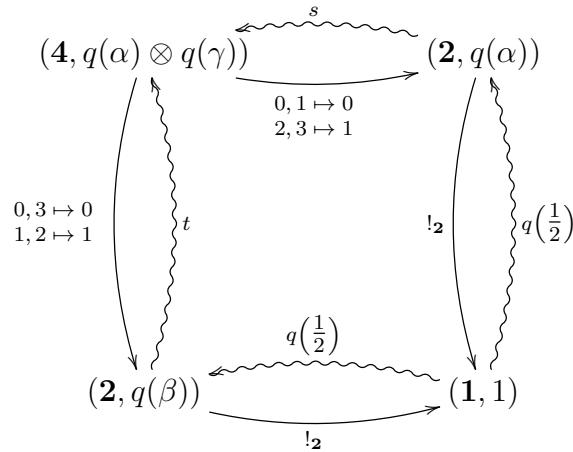

where we take:

$$s = \begin{pmatrix} \gamma & 0 \\ 1 - \gamma & 0 \\ 0 & \gamma \\ 0 & 1 - \gamma \end{pmatrix}, \qquad t = \begin{pmatrix} \gamma & 0 \\ 0 & 1 - \gamma \\ 0 & \gamma \\ 1 - \gamma & 0 \end{pmatrix}$$

The square commutes and the upper horizontal morphism is in FP, so the value of F on the bottom horizontal morphism is bounded by the value of F on the right vertical one, as was to be shown.   ∎

In the preceding lemma we are not yet claiming that $h(\alpha, \frac{1}{2})$ is finite. We show this for $\alpha = \frac{1}{4}$ in Lemma 4.10, and for all $\alpha \in (0, 1)$ in Lemma 4.11, where we actually obtain a uniform bound.

4.9. LEMMA. $h(\alpha, \frac{1}{2}) = h(1 - \alpha, \frac{1}{2})$ *for all $\alpha \in [0, 1]$.*

PROOF. Apply functoriality to the commutative triangle
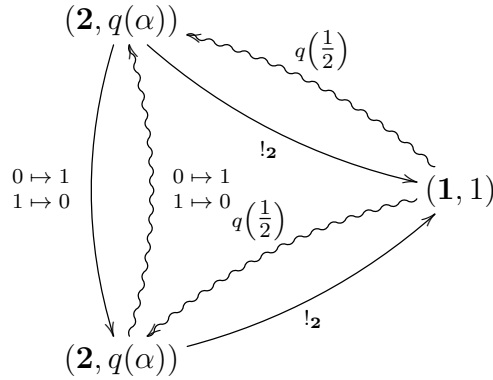
where the vertical morphism is in FP.   ∎

4.10. LEMMA. $h(\frac{1}{4}, \frac{1}{2}) < \infty$.

PROOF. We use (9) with $\beta = \frac{1}{2}$:

$$
\begin{aligned}
h\left(\alpha, \frac{1}{2}\right) &= \left(\alpha + \frac{1}{2}\right) h\left(\frac{2\alpha}{1 + 2\alpha}, \frac{1}{2}\right) \\
&+ \left(\frac{3}{2} - \alpha\right) h\left(\frac{2 - 2\alpha}{3 - 2\alpha}, \frac{1}{2}\right) + 2h\left(\frac{1 + 2\alpha}{4}, \frac{1}{2}\right),
\end{aligned}
\tag{10}
$$

which we will apply for $\alpha < \frac{1}{2}$. On the right-hand side here, the first argument of h in the second term can be replaced by $\frac{1}{3 - 2\alpha}$, thanks to Lemma 4.9. Then the first arguments in all three terms on the right-hand side are in $[0, \frac{1}{2}]$, with the smallest in the first term, so Lemma 4.8 tells us that

$$
h\left(\alpha, \frac{1}{2}\right) \le 4h\left(\frac{2\alpha}{1 + 2\alpha}, \frac{1}{2}\right).
$$

Now with $\alpha_0 = \frac{1}{4}$, the sequence recursively defined by $\alpha_{n+1} = \frac{2\alpha_n}{1 + 2\alpha_n}$ increases and converges to $\frac{1}{2}$. In particular we can find n with $\alpha' < \alpha_n < \frac{1}{2}$, where α' is chosen as in Lemma 4.7. Using that result together with Lemma 4.8, we obtain

$$
h\left(\frac{1}{4}, \frac{1}{2}\right) \le 4^n\, h\left(\alpha_n, \frac{1}{2}\right) \le 4^n\, h\left(\alpha', \frac{1}{2}\right) < \infty.
$$

∎

4.11. LEMMA. *There is a constant $B < \infty$ such that $h(\alpha, \frac{1}{2}) \le B\, h(\frac{1}{4}, \frac{1}{2})$ for all $\alpha \in (0, 1)$.*

PROOF. By the symmetry in Lemma 4.9, it is sufficient to consider $\alpha \in (0, \frac{1}{2}]$. By Lemma 4.8, we may use the bound $B = 1$ for all $\alpha \in [\frac{1}{4}, \frac{1}{2}]$. It thus remains to find a choice of B that works for all $\alpha \in (0, \frac{1}{4})$, and we assume α to lie in this interval from now on.

We reuse Equation (10). Both the second and the third term on the right-hand side have their first argument of h in the interval $[\frac{1}{4}, \frac{3}{4}]$, so we can apply Lemmas 4.8 and 4.9 to obtain

$$h\left(\alpha, \frac{1}{2}\right) \leq \left(\alpha + \frac{1}{2}\right) h\left(\frac{2\alpha}{1 + 2\alpha}, \frac{1}{2}\right) + \left(\frac{7}{2} - \alpha\right) h\left(\frac{1}{4}, \frac{1}{2}\right).$$

To find a simpler-looking upper bound, we bound the right-hand side from above by applying Lemma 4.8 in order to replace the $\frac{2\alpha}{1+2\alpha}$ argument by just 2α, and at the same time use $\alpha \in (0, \frac{1}{4})$ in order to bound the coefficients of both terms by $\alpha + \frac{1}{2} \leq \frac{3}{4}$ and $\frac{7}{2} - \alpha \leq \frac{7}{2}$:

$$h\left(\alpha, \frac{1}{2}\right) \leq \frac{3}{4} h\left(2\alpha, \frac{1}{2}\right) + \frac{7}{2} h\left(\frac{1}{4}, \frac{1}{2}\right).$$

If we put $\alpha = 2^{-n}$ for $n \geq 2$, then we can apply this inequality repeatedly until only terms of the form $h(\frac{1}{4}, \frac{1}{2})$ are left. This results in a geometric series:

$$h\left(2^{-n}, \frac{1}{2}\right) \leq \left(\left(\frac{3}{4}\right)^{n-2} + \sum_{k=0}^{n-3} \left(\frac{3}{4}\right)^k \cdot \frac{7}{2}\right) h\left(\frac{1}{4}, \frac{1}{2}\right).$$

whose convergence (as $n \to \infty$) implies the existence of a constant $B < \infty$ with

$$h(2^{-n}, \tfrac{1}{2}) \leq B\, h(\tfrac{1}{4}, \tfrac{1}{2})$$

for all $n \geq 2$. The present lemma then follows with the help of Lemma 4.8.   ∎

4.12. LEMMA. *Equation* (6) *holds if* $c = \infty$.

PROOF. By Lemma 4.11 and the lower semicontinuity of h, we see that

$$g(\tfrac{1}{2}) = h(0, \tfrac{1}{2}) < \infty$$

This implies that the constant c with $g(\alpha) = -c \ln \alpha$ has $c < \infty$. Recall that we have shown this under the assumption that there exist probability distributions p and r on a finite set X with $p \neq r$ and

$$F\left((X,p) \underset{!_X}{\overset{r}{\rightleftharpoons}} (\mathbf{1}, 1)\right) < \infty.$$

So, taking the contrapositive, we see that if $c = \infty$, then

$$F\left((X,p) \underset{!_X}{\overset{r}{\rightleftharpoons}} (\mathbf{1}, 1)\right) = \infty$$

whenever p and r are distinct probability distributions on X. This proves Equation (6) except in the case where $p = r$. But in that case, both sides vanish, since on the left we are taking F of a morphism in FP, and on the right we obtain $\infty \cdot 0 = 0$.   ∎

5. Counterexamples and subtleties

One might be tempted to think that our Theorem 3.1 also holds if one relaxes the lower semicontinuity assumption to measurability, upon equipping the hom-spaces of both FinStat and $[0, \infty]$ with their σ-algebras of Borel sets. For $[0, \infty]$, this σ-algebra is the usual Borel σ-algebra: the sets of the form (a, ∞) are open and hence measurable, the sets of the form $[0, b]$ are closed and hence measurable, and therefore all half-open intervals $(a, b]$ are measurable, and these generate the standard Borel σ-algebra. However, for Theorem 3.1, mere measurability of the functor F is not enough:

5.1. PROPOSITION. *There is a functor* FinStat $\to [0, \infty]$ *that is convex linear, measurable on hom-spaces, and vanishes on* FP, *but is not a scalar multiple of relative entropy.*

PROOF. We claim that one such functor $G\colon$ FinStat $\to [0, \infty]$ is given by

$$G\left((X, p) \xleftarrow[\;\;f\;\;]{\;\;s\;\;} (Y, q) \right) = \begin{cases} 0 & \text{if } \operatorname{supp}(p) = \operatorname{supp}(s \circ q), \\ \infty & \text{if } \operatorname{supp}(p) \neq \operatorname{supp}(s \circ q). \end{cases}$$

This G clearly vanishes on FP. Since taking the support of a probability distribution is a lower semicontinuous and hence measurable function, the set of all morphisms obeying $\operatorname{supp}(p) = \operatorname{supp}(s \circ q)$ is also measurable, and hence G is measurable.

Concerning functoriality, for a composable pair of morphisms

$$(X, p) \xleftarrow[\;\;f\;\;]{\;\;s\;\;} (Y, q) \xleftarrow[\;\;g\;\;]{\;\;t\;\;} (Z, r),$$

we have

$$\operatorname{supp}(p) = \operatorname{supp}(s \circ q), \quad \operatorname{supp}(q) = \operatorname{supp}(t \circ r) \quad \Longleftrightarrow \quad \operatorname{supp}(p) = \operatorname{supp}(s \circ t \circ r).$$

This proves functoriality. A similar argument proves convex linearity. ∎

As a measure of information gain, this functor G is not hard to understand intuitively: we gain no information whenever the set of *possible* outcomes is precisely the set that we expected; otherwise, we gain an infinite amount information.

Since the collection of all functors satisfying our hypotheses is closed under sums and scalar multiples and also contains the relative entropy functor, we actually obtain a whole family of such functors. For example, another one of these functors is $G'\colon$ FinStat $\to [0, \infty]$ given by

$$G'\left((X, p) \xleftarrow[\;\;f\;\;]{\;\;s\;\;} (Y, q) \right) = \begin{cases} S(p, s \circ q) & \text{if } \operatorname{supp}(p) = \operatorname{supp}(s \circ q), \\ \infty & \text{if } \operatorname{supp}(p) \neq \operatorname{supp}(s \circ q). \end{cases}$$

Our original idea was to use the work of Petz [8, 9] to prove Theorem 3.1. However, as it turned out, there is a gap in Petz's argument. Although his purported characterization concerns the quantum version of relative entropy, the first part of his proof in [8] treats the classical case. If his proof were correct, it would prove this:

5.2. UNPROVED "THEOREM". *The relative entropy $S(p,r)$ for pairs of probability measures on the same finite set such that r has full support is characterized up to a multiplicative constant by these properties:*

  (a) **Conditional expectation law**. *Suppose $f\colon X \to Y$ is a function and $s\colon Y \rightsquigarrow X$ a stochastic map with $f \circ s = 1_Y$. Given probability distributions p and r on X, and assuming that r has full support and $r = s \circ f \circ r$, we have*

$$S(p,r) = S(f \circ p, f \circ r) + S(p, s \circ f \circ p). \tag{11}$$

  (b) **Invariance.** *Given any bijection $f\colon X \to Y$ and probability distributions p, r on X such that r has **full support** (i.e. its support is all of X), we have*

$$S(f \circ p, f \circ r) = S(p,r).$$

  (c) **Convex linearity.** *Given probability distributions p, r on X and p', r' on Y such that r and r' have full support, and given $\lambda \in [0,1]$, we have*

$$S(\lambda p \oplus (1-\lambda)p', \lambda r \oplus (1-\lambda)r') = \lambda S(p,r) + (1-\lambda)S(p',r').$$

  (d) **Nilpotence.** *For any probability distribution p with full support on a finite set, $S(p,p) = 0$.*

  (e) **Measurability property.** *The function*

$$(p,r) \mapsto S(p,r)$$

is measurable on the space of pairs of probability distributions on X such that r has full support.

Note that [8] uses the opposite ordering for the two arguments of S.

    The problem with this "theorem" is the range of applicability of Equation (11): what is this formula supposed to mean when $s \circ f \circ p$ does not have full support? After all, $S(p,r)$ is assumed to be defined only when the second argument has full support, but this need not be the case for $s \circ f \circ p$, given the assumptions made in the statement of the conditional expectation property. (Note that $f \circ r$ has full support, so the term $S(f \circ p, f \circ r)$ is fine.)

    One can try to correct this problem by assuming that the conditional expectation property holds only if $s \circ f \circ p$ has full support as well. However, this means that the proof of Petz's Lemma 1 is valid only when (using his notation) $p_3 > 0$, which implies that his Equation (5) is known to hold only for $p_2 > 0$ and $p_3 > 0$. Upon following the thread of Petz's argument, one finds that his Equation (6) has been proven to follow from his assumptions only for $x \in (0,1)$ and $u \in (0,1)$. However, the solution of that functional equation in the references he points to crucially uses the assumption that the functional equation also holds in case that $x = 0$ or $u = 0$. This is the gap in Petz's proof.

In fact, if one allows S to take on infinite values, then the above classical version of Petz's theorem is not even correct, if one uses the interpretation that (11) is to be applied only when $s \circ f \circ p$ has full support. The counterexample is similar to our functor G' from above:

$$S'(p,r) = \begin{cases} S(p,r) & \text{if } p \text{ has full support,} \\ \infty & \text{otherwise.} \end{cases}$$

6. Conclusions

The theorem here, and our earlier characterization of entropy [1], can be seen as part of a program of demonstrating that mathematical structures that are "socially important" are also "categorically natural". Tom Leinster, whose words we quote here, has carried this forward to a categorical explanation of Lebesgue integration [6]. It would be interesting to generalize our results on entropy and relative entropy from finite sets to general measure spaces, where integrals replace sums. It would be even more interesting to do this using a category-theoretic approach to integration.

It would also be good to express our theorem more concisely. As noted in Appendix B, convex linear combinations are operations in a topological operad P. We can define 'convex algebras', that is, algebras of P, in any symmetric monoidal topological category. The category $[0, \infty]$ with the upper topology on its set of morphisms is a convex algebra in `TopCat`, the (large) topological category of topological categories. We believe, but have not proved, that `FinStat` is a 'weak' convex algebra in `TopCat`. This would mean that the axioms for a convex algebra hold up to coherent natural isomorphism [4]. If this is true, the relative entropy

$$\text{RE}\colon \texttt{FinStat} \to [0, \infty]$$

should be, up to a constant factor, the unique map of weak convex algebras that vanishes on morphisms in `FP`. Leinster [5] has shown that `FP` is also a weak convex algebra in `Cat(Top)`. In fact, it is the free such thing on an internal convex algebra. So, it seems that both entropy and relative entropy emerge naturally from a category-theoretic examination of convex linearity.

A. Semicontinuous functors

In Section 3 we explained what it meant for relative entropy to be a semicontinuous functor. A more sophisticated way to think about semicontinuous functors uses topological categories. This requires that we put a nonstandard topology on $[0, \infty]$, the so-called 'upper topology'.

A topological category is a category internal to `Top`, and a continuous functor is a functor internal to `Top`. In other words:

A.1. DEFINITION. *A* **topological category** *C is a small category where the set of objects C_0 and the set of morphisms C_1 are equipped with the structure of topological spaces, and the maps assigning to each morphism its source and target:*

$$s, t \colon C_1 \to C_0$$

the map assigning to each object its identity morphism

$$i \colon C_0 \to C_1$$

and the map sending each pair of composable morphisms to their composite

$$\circ \colon C_1 \times_{C_0} C_1 \to C_1$$

are continuous. Given topological categories C and D, a **continuous functor** *is a functor $F \colon C \to D$ such that the map on objects $F_0 \colon C_0 \to D_0$ and the map on morphisms $F_1 \colon C_1 \to D_1$ are continuous.*

We now explain how `FinStoch` and `FinStat` are topological categories. Strictly speaking, in order for this to work, we need to deal with size issues. One approach is to let the objects of `Top` be 'large' sets living in a higher Grothendieck universe, which allows us to talk about the set of all objects or morphisms of `FinStat` or `FinStoch`. Another is to replace each of these categories by its skeleton, which is an equivalent small category. From now on, we assume that one of these things has been done.

For `FinStoch`, we put the discrete topology on its set of objects $\mathtt{FinStoch}_0$. Each hom-set $\mathtt{FinStoch}(X, Y)$ is a subset of the Euclidean space $\mathbb{R}^{|X| \times |Y|}$, and we put the subspace topology on this hom-set; for example, $\mathtt{FinStoch}(1, Y)$, the set of all probability distributions on Y, is topologized as a simplex. In this way, `FinStoch` becomes a category *enriched* over `Top`, and in particular internal to `Top`.

As for `FinStat`, the identification

$$\mathtt{FinStat}_0 = \{(X, p) \mid X \in \mathtt{FinStoch}_0, \; p \in \mathtt{FinStoch}(1, X)\} \subseteq \mathtt{FinStoch}_0 \times \mathtt{FinStoch}_1$$

induces a topology on $\mathtt{FinStat}_0$. In this topology, a net $(X^\lambda, p^\lambda)_{\lambda \in \Lambda}$ converges to (X, p) if and only if eventually $X^\lambda = X$, and $p^\lambda \to p$ for those λ with $X^\lambda = X$. Similarly, every morphism in `FinStat` consists of a pair of morphisms in `FinStoch` satisfying certain conditions, and the resulting inclusion

$$\mathtt{FinStat}_1 \subseteq \mathtt{FinStoch}_1 \times \mathtt{FinStoch}_1$$

can be used to define a topology on $\mathtt{FinStat}_1$. We omit the verification that these topologies make `FinStat` into a topological category.

There is a topology on $[0, \infty]$ where the open sets are those of the form $(a, \infty]$, together with the whole space and the empty set. This is called the **upper topology**. With this

topology, a function $\psi \colon A \to [0, \infty]$ from any topological space A is continuous if and only ψ is lower semicontinuous, meaning

$$\psi(a) \leq \liminf_{\lambda \to \infty} \psi(a^\lambda)$$

for every convergent net $a^\lambda \in A$. It is easy to check that this topology on $[0, \infty]$ makes addition continuous.

In short, $[0, \infty]$ with its upper topology is a topological monoid under addition. We thus obtain a topological category with one object and $[0, \infty]$ as its topological monoid of endomorphisms. By abuse of notation we also call this topological category simply $[0, \infty]$. This lets us state Lemma 3.4 in a different way:

A.2. Lemma. *If $[0, \infty]$ is viewed as a topological category using the upper topology, the functor* RE: FinStat $\to [0, \infty]$ *is continuous.*

On the other hand, if we give the monoid $[0, \infty]$ the less exotic topology where it is homeomorphic to a closed interval, then this functor is *not* continuous.

Having gone this far, we cannot resist pointing out that $[0, \infty]$ with its upper topology is also a topological rig. Recall that a **rig** is a 'ring without negatives': a set equipped with an addition making it into a commutative monoid and a multiplication making it into a monoid, with multiplication distributing over addition. In other words, it is a monoid in the monoidal category of commutative monoids. A **topological rig** is a rig with a topology in which addition and multiplication are continuous. To make $[0, \infty]$ into a rig, we define addition as before, define multiplication in the usual way for numbers in $[0, \infty)$, and set

$$0a = a0 = 0$$

for all $a \in [0, \infty]$. One can verify that multiplication is continuous: but again, the key point is that we need to use the upper topology, since $\infty \cdot a$ suddenly jumps from ∞ to 0 as a reaches zero. Thus:

A.3. Lemma. *With its upper topology, $[0, \infty]$ is a topological rig.*

More important now is that $[0, \infty]$ is a module over the rig $[0, \infty)$, where addition and multiplication in the latter are defined as usual and we define the action of $[0, \infty)$ on $[0, \infty]$ using multiplication, with the proviso that $0 \cdot a = 0$ even when $a = \infty$. And here we see:

A.4. Lemma. *The topological monoid $[0, \infty]$ with its upper topology becomes a topological module over the rig $[0, \infty)$ with its usual topology.*

B. Convex algebras

We define the **monad for convex sets** to be the monad on Set sending any set X to the set of finitely-supported probability distributions on X. For example, this monad sends

$\{1, \ldots, n\}$ to the set

$$P_n = \{p \in [0,1]^n : \sum_{i=1}^{n} p_i = 1\}$$

which can be identified with the $(n-1)$-simplex. This monad is finitary, so can be thought about in a few different ways.

First, a finitary monad can thought of as a finitary algebraic theory. The monad for convex sets can be presented by a family $(*_\lambda)_{\lambda \in [0,1]}$ of binary operations, subject to the equations

$$x *_0 y = x,$$
$$x *_\lambda x = x,$$
$$x *_\lambda y = y *_{1-\lambda} x,$$
$$(x *_\mu y) *_\lambda z = x *_{\lambda\mu} (y *_{\frac{\lambda(1-\mu)}{1-\lambda\mu}} z)$$

For $\lambda = \mu = 1$, the fraction $\frac{\lambda(1-\mu)}{1-\lambda\mu}$ in the last equation may be taken to be an arbitrary number in $[0,1]$. See [2] for more detail on how to derive this presentation from the monad.

A finitary algebraic theory can also be thought of as an operad with extra structure. In a symmetric operad O, one has for each bijection $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ an induced map $\sigma_* : O_n \to O_n$. In a finitary algebraic theory, one has the same thing for *arbitrary* functions between finite sets, not just bijections. In other words, a finitary algebraic theory amounts to a non-symmetric operad O together with, for each function $\theta : \{1, \ldots, m\} \to \{1, \ldots, n\}$ between finite sets, an induced map $\theta_* : O_m \to O_n$, satisfying suitable axioms.

B.1. DEFINITION. *The underlying symmetric operad for the monad for convex sets is called the* **operad for convex algebras** *and denoted* P. *An algebra of* P *is called a* **convex algebra**.

The space of n-ary operations for this operad is P_n, the space of probability distributions on $\{1, \ldots, n\}$. The composition of operations works as follows. Given probability distributions $p \in P_n$ and $r_i \in P_{k_i}$ for each $i \in \{1, \ldots, n\}$, we obtain a probability distribution $p \circ (r_1, \ldots, r_n) \in P_{k_1 + \cdots + k_n}$, namely

$$p \circ (r_1, \ldots, r_n) = (p_1 r_{11} \ldots, p_1 r_{1k_1}, \ldots p_n r_{n1}, \ldots, p_n r_{nk_n}).$$

The maps $\theta_* : P_m \to P_n$ can be defined by pushforward of measures. An algebra for the algebraic theory of convex algebras is an algebra X for the operad with the further

property that the square

$$
\begin{array}{ccc}
P_m \times X^n & \xrightarrow{\;1\times\theta^*\;} & P_m \times X^m \\
\Big\downarrow{\scriptstyle \theta_*\times 1} & & \Big\downarrow \\
P_n \times X^n & \xrightarrow{\hspace{3cm}} & X
\end{array}
$$

commutes for all $\theta\colon \{1,\dots,m\} \to \{1,\dots,n\}$, where the unlabelled arrows are given by the convex algebra structure of X.

Note that P is naturally a topological operad, where the topology on P_n is the usual topology on the $(n-1)$-simplex. In this paper we have implicitly been using algebras of P in various topological categories E with finite products. We call these **convex algebras** in E. Here are some examples:

- Any convex subset of $\mathbb{R}^n$ is a convex algebra in `Top`.

- The additive monoid $[0,\infty]$ with its upper topology becomes a convex algebra in `Top` if we define convex linear combinations by treating $[0,\infty]$ as a topological module of the rig $[0,\infty)$ as in Lemma A.4. We must equip $[0,\infty]$ with its upper topology for this to work, because the convex linear combination $\lambda\cdot\infty + (1-\lambda)\cdot a$ equals ∞ when $\lambda > 0$, but suddenly jumps down to a when λ reaches zero.

- The category `Cat(Top)` of small topological categories and continuous functors is itself a large topological category. If we regard $[0,\infty]$ with its upper topology as a one-object topological category as in Appendix A, then it becomes a convex algebra in `Cat(Top)` thanks to the previous remark.

- The categories `FinProb`, `FinStat` should be 'weak convex algebras' in `Cat(Top)`, though we have not carefully checked this. By this, we mean that axioms for an algebra of the operad P hold up to coherent natural isomorphism, in the sense made precise by Leinster [4].

- Similarly, Leinster has shown that FP is a weak convex algebra in `Cat(Top)`. In fact, it is equivalent to the free convex algebra in `Cat(Top)` on an internal convex algebra [5].

References

[1] J. Baez, T. Fritz and T. Leinster, A characterization of entropy in terms of information loss, *Entropy* **13** (2011), 1945–1957. Also available as arXiv:1106.1791.

[2] T. Fritz, Convex spaces I: definition and examples, available as arXiv:0903.5522.

[3] L. Itti, P.F. Baldi, Bayesian surprise attracts human attention, in *Advances in Neural Information Processing Systems* 19 (2005), 547–554. Also available as http://ilab.usc.edu/publications/doc/Itti_Baldi06nips.pdf.

[4] T. Leinster, *Higher Operads, Higher Categories*, London Mathematical Society Lecture Note Series **298**, Cambridge U. Press, Cambridge, 2004. Also available as arxiv:math.CT/0305049.

[5] T. Leinster, An operadic introduction to entropy, *The n-Category Café*, 18 May 2011. Available at http://golem.ph.utexas.edu/category/2011/05/an_operadic_introduction_to_en.html.

[6] T. Leinster, The categorical origins of Lebesgue integration, talk at Category Theory 2014, 4 July 2014. Available at http://www.maths.ed.ac.uk/~tl/cambridge_ct14/

[7] M. Kuczma, *An Introduction to the Theory of Functional Equations and Inequalities: Cauchy's Equation and Jensen's Inequality*, Birkhäuser, Basel, 2009.

[8] D. Petz, Characterization of the relative entropy of states of matrix algebras, *Acta Math. Hungar.* **59** (1992), 449–455. Also available at http://www.renyi.hu/~petz/pdf/52.pdf.

[9] D. Petz, *Quantum entropy and its use*, Texts and Monographs in Physics, Springer (1993).

Department of Mathematics
University of California
Riverside CA 92521, USA
and
Centre for Quantum Technologies
National University of Singapore
Singapore 117543

Perimeter Institute for Theoretical Physics
31 Caroline St. N, Waterloo, ON, Canada N2L 2Y5
Email: `baez@math.ucr.edu`
          `tfritz@perimeterinstitute.ca`

This article may be accessed at `http://www.tac.mta.ca/tac/`