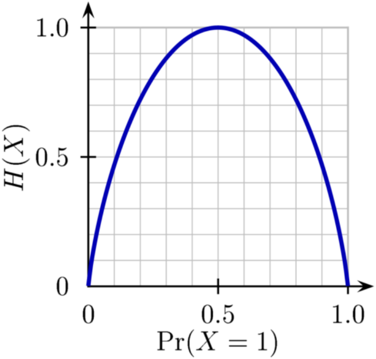
Characterizing entropy



John Baez, Tobias Fritz, Tom Leinster

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 for any x, the numbers f_{yx} form a probability distribution on Y.

We call f_{yx} **the probability of** y **given** x**.**

So, we demand:

- $f_{yx} \geq 0$ for all $x \in X$, $y \in Y$,
- $\displaystyle\sum_{y \in Y} f_{yx} = 1$ for all $x \in X$.

We can compose stochastic maps $f\colon X \to Y$ and $g\colon Y \to Z$ by matrix multiplication:

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

and get a stochastic map $g \circ f\colon X \to Z$.

We let FinStoch be the category with

- finite sets as objects,
- stochastic maps $f\colon X \rightsquigarrow Y$ as morphisms.

Every function $f\colon X \to Y$ is a stochastic map, so we get

$$\mathtt{FinSet} \hookrightarrow \mathtt{FinStoch}$$

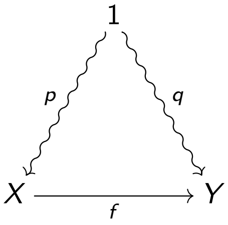
Let 1 be your favorite 1-element set. A stochastic map

$$1 \overset{p}{\rightsquigarrow} X$$

is a probability distribution on X.

We call $p\colon 1 \rightsquigarrow X$ a **finite probability measure space**.

A **measure-preserving map** between finite probability measure spaces is a commuting triangle

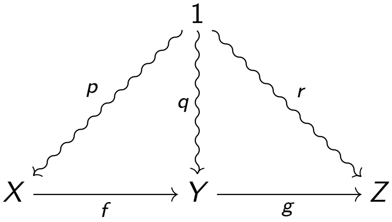


So, $f\colon X \to Y$ sends the probability distribution on X to that on Y:

$$q_y = \sum_{x\colon f(x)=y} p_x$$

It's a 'deterministic way of processing random data'.

We can compose measure-preserving maps:



So, we get a category `FinProb` with

- finite probability measure spaces as objects
- measure-preserving maps as morphisms.

Any finite probability measure space $p\colon 1 \rightsquigarrow X$ has an **entropy**:

$$S(p) = -\sum_{x \in X} p_x \ln p_x$$

This says how 'evenly spread' p is.

Or: how much information you learn, on average, when someone tells you an element $x \in X$, if all you'd known was that it was randomly distributed according to p.

Flip a coin!

If $X = \{h, t\}$ and $p_h = p_t = \frac{1}{2}$, then

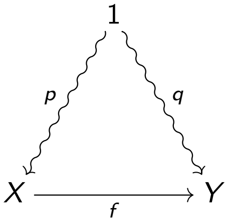
$$S(X, p) = -\left(\tfrac{1}{2}\ln\tfrac{1}{2} + \tfrac{1}{2}\ln\tfrac{1}{2}\right) = \ln 2$$

so you learn $\ln 2$ **nats** of information on average, or 1 **bit**.

But if $p_h = 1, p_t = 0$ you learn

$$S(X, p) = -(1\ln 1 + 0\ln 0) = 0$$

What's so good about entropy? Let's focus on the **information loss** of a measure-preserving map:

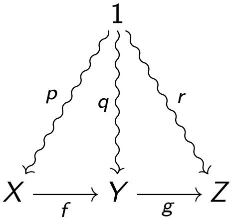


$$\mathtt{IL}(f) = S(X, p) - S(Y, q)$$

The **data processing inequality** says that

$$\mathtt{IL}(f) \geq 0$$

Deterministic processing of random data always *decreases* entropy!



Clearly we have

$$\begin{aligned} \mathtt{IL}(g \circ f) &= S(X,p) - S(Z,r) \\ &= S(X,p) - S(Y,q) + S(Y,q) - S(Z,r) \\ &= \mathtt{IL}(f) + \mathtt{IL}(g) \end{aligned}$$

So, information loss should be a *functor* from `FinProb` to a category with numbers $[0,\infty)$ as morphisms and addition as composition.

Indeed there is a category $[0, \infty)$ with:

- one object $*$
- nonnegative real numbers c as morphisms $c \colon * \to *$
- addition as composition.

We've just seen that

$$\mathtt{IL} \colon \mathtt{FinProb} \to [0, \infty)$$

is a functor. *Can we characterize this functor?*

Yes. The key is that IL is 'convex-linear' and 'continuous'.

We can define **convex linear combinations** of objects in `FinProb`. For for any $0 \leq c \leq 1$, let

$$c\,(X,p)\ +\ (1-c)\,(Y,q)$$

be the disjoint union of X and Y, with the probability distribution given by cp on X and $(1-c)q$ on Y.

We can also define convex linear combinations of morphisms.

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

give

$$cf + (1-c)g\colon c(X,p) + (1-c)(Y,q) \to c(X',p') + (1-c)(Y',q')$$

This is simply the function that equals f on X and g on Y.

Information loss is **convex linear**:

$$\mathtt{IL}\big(cf + (1-c)g\big) = c\,\mathtt{IL}(f) \; + \; (1-c)\,\mathtt{IL}(g)$$

The reason is that

$$S(c(X,p) + (1-c)(Y,q)) = c\,S(X,p) \; + \; (1-c)\,S(Y,q) \; + \; S_c$$

where

$$S_c = -\Big(c \ln c \; + \; (1-c)\ln(1-c)\Big)$$

is the entropy of a coin with probability c of landing heads-up. This extra term cancels when we compute information loss.

$\mathtt{FinProb}$ and $[0, \infty)$ are also **topological categories**: they have topological spaces of objects and morphisms, and the category operations are continuous.

$\mathtt{IL} \colon \mathtt{FinProb} \to [0, \infty)$ is a **continuous functor**: it is continuous on objects and morphisms.

Theorem (Baez, Fritz, Leinster). Any continuous convex-linear functor

$$F\colon \mathtt{FinProb} \to [0, \infty)$$

is a constant multiple of the information loss: for some $\alpha \geq 0$,

$$g\colon (X, p) \to (Y, q) \quad \Longrightarrow \quad F(g) = \alpha\, \mathtt{IL}(g).$$

Theorem (Baez, Fritz, Leinster). Any continuous convex-linear functor

$$F\colon \texttt{FinProb} \to [0, \infty)$$

is a constant multiple of the information loss: for some $\alpha \geq 0$,

$$g\colon (X, p) \to (Y, q) \quad \Longrightarrow \quad F(g) = \alpha\, \texttt{IL}(g).$$

The easy part of the proof: show that

$$F(g) = \Phi(X, p) - \Phi(Y, q)$$

for some quantity $\Phi(X, p)$. The hard part: show that

$$\Phi(X, p) = -\alpha \sum_{x \in X} p_x \ln p_x$$

This part relies on an earlier characterization due to Faddeev.

Two generalizations:

1) There is precisely a one-parameter family of convex structures on the category $[0, \infty)$. Using these we get information loss functors

$$\mathtt{IL}_\beta \colon \mathtt{FinProb} \to [0, \infty)$$

based on Tsallis entropy:

$$S_\beta(X, p) = \frac{1}{\beta - 1}\left(1 - \sum_{x \in X} p_x^\beta\right)$$

which reduces to the ordinary entropy as $\beta \to 1$.

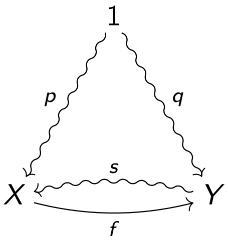
2) The entropy of one probability distribution on X **relative to** another:

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

is the expected amount of information you gain when you *thought* the right probability distribution was q and you discover it's really p. It can be infinite!

There is also category-theoretic characterization of relative entropy.

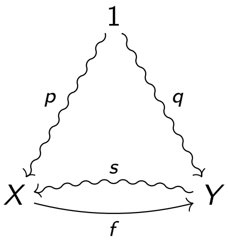
This uses a category `FinStat` where the objects are finite probability measure spaces, but the morphisms look like this:



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

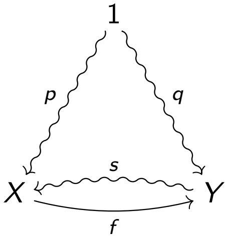
We have a measure-preserving map $f\colon X \to Y$ equipped with a stochastic right inverse $s\colon Y \rightsquigarrow X$. Think of f as a ‘measurement process’ and s as a ‘hypothesis’ about the state in X given the measurement in Y.

Any morphism in FinStat



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

gives a relative entropy $D(p \,||\, s \circ q)$. This says how much information we gain when we learn the 'true' probability distribution p on the states of the measured system, given our 'guess' $s \circ q$ based on the measurements q and our hypothesis s.



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

Our hypothesis s is **optimal** if $p = s \circ q$: our guessed probability distribution equals the true one! In this case $D(p \,||\, s \circ q) = 0$.

Morphisms with an optimal hypothesis form a subcategory

$$\mathtt{FP} \hookrightarrow \mathtt{FinStat}$$

Theorem (Baez, Fritz). Any lower semicontinuous convex-linear functor

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

vanishing on morphisms in $\texttt{FP}$ is a constant multiple of relative entropy.

The proof is hard! Can you simplify it?