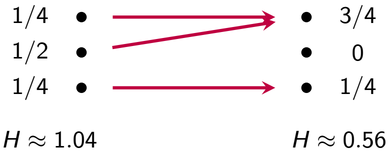
Shannon Entropy from Category Theory



John Baez

Categorical Semantics of Entropy

11 May 2022

SHANNON ENTROPY

A probability distribution p on a finite set X has a **Shannon entropy**:

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

This says how 'evenly spread' p is.

Or: how much information you learn, on average, when someone picks an element $x \in X$ according to the distribution p and tells you what it is — if all you'd known before 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

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

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

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

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

THE EXPECTED SURPRISE

To compute Shannon entropy we turn probabilities p_x into **surprisals** by taking their negative logarithm, and then compute their expected value:

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

So, Shannon entropy is the “expected surprise”.

WHAT'S SO GREAT ABOUT SHANNON ENTROPY?

There are many alternatives notions of entropy. For example, the **Tsallis entropy**:

$$\frac{1}{\alpha - 1}\left(1 - \sum_{x \in X} p_x^\alpha\right)$$

for real $\alpha \neq 1$, and the **Rényi entropy**:

$$\frac{1}{1 - \alpha} \ln\left(\sum_{x \in X} p_x^\alpha\right)$$

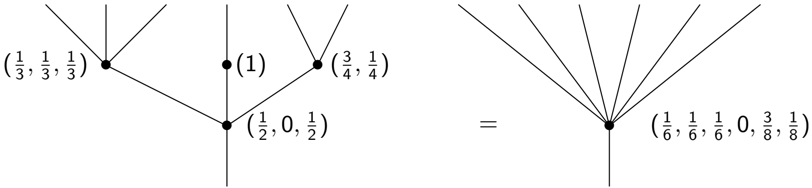
for $\alpha \geq 0$ with $\alpha \neq 1$.

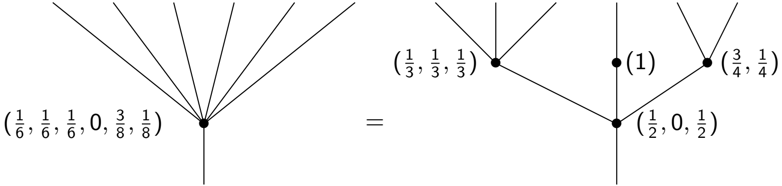
Both approach the Shannon entropy as $\alpha \to 1$. Both have good properties, discussed here:

- Tom Leinster, *Entropy and Diversity: the Axiomatic Approach*, 2020.

So, we should say which good properties single out Shannon entropy!

The most important is the 'chain rule'. To state this, note that we can *compose* probability distributions in a tree-like way:

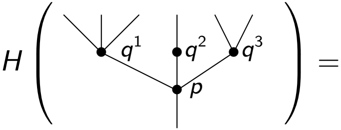


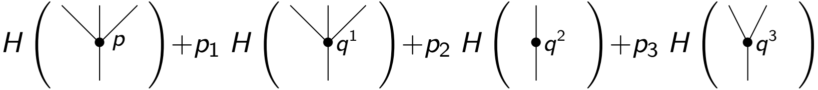


Whenever you compose probability distributions in a tree-like way, Shannon entropy obeys the 'chain rule':

$$H(\tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, 0, \tfrac{3}{8}, \tfrac{1}{8}) \;=\; H(\tfrac{1}{2}, 0, \tfrac{1}{2}) + \tfrac{1}{2}\, H(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}) + 0\, H(1) + \tfrac{1}{2}\, H(\tfrac{3}{4}, \tfrac{1}{4})$$

More generally:





In a more compressed notation, the **chain rule** says

$$H(p \circ (q^1, \ldots, q^n)) = H(p) + \sum_{i=1}^{n} p_i \, H(q^i)$$

when p is a probability distribution on $\{1, \ldots, n\}$.

Theorem (Faddeev, Leinster). Suppose I is a map sending any probability distribution on any finite set to a nonnegative real number, and:

1. I is invariant under bijections.
2. I is continuous.
3. I obeys the chain rule.

Then I is a constant nonnegative multiple of Shannon entropy.

This is a modern version of Dmitry Faddeev's 1956 theorem, due to Leinster: it's Theorem 2.5.1 in Leinster's book *Entropy and Diversity: the Axiomatic Approach*.

How does the logarithm function show up?

If we let

$$\phi(n) = I(\tfrac{1}{n}, \ldots, \tfrac{1}{n})$$

then the chain rule implies

$$\phi(mn) = \phi(m) + \phi(n)$$

This has obvious solutions

$$\phi(n) = c \ln n$$

but to rule out *nonobvious* solutions we must use the continuity condition on I. We then need more tricks to show

$$I(p_1, \ldots, p_n) = -c \sum_{i=1}^{n} p_i \ln p_i$$

It would be nice to see Shannon entropy emerge naturally from category theory! That was our goal here:

- John Baez, Tobias Fritz and Tom Leinster, A characterization of entropy in terms of information loss, 2011.

The key idea:

Category theory is really about *morphisms*, not objects. So we should talk not about the Shannon entropy of an object — a finite set with a probability measure — but the *change in entropy* due to some kind of *morphism* between these objects.

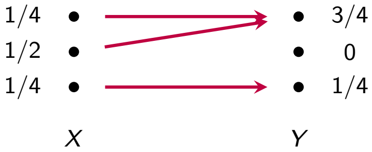
Given finite sets with probability distributions (X, p) and (Y, q), a **measure-preserving map** from the first to the second is a function

$$f : X \to Y$$

that sends p to q in this way:

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

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



The composite of measure-preserving maps is measure-preserving.
So, we get a category FinProb with

- finite sets equipped with probability distributions as objects
- measure-preserving maps as morphisms.

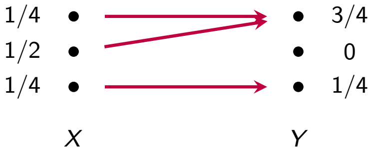
Let's define the **entropy loss** of a measure-preserving map $f\colon (X, p) \to (Y, q)$ by

$$\mathrm{Loss}(f) = H(X, p) - H(Y, q)$$

The **data processing inequality** says that

$$\mathrm{Loss}(f) \geq 0$$

Deterministic processing of random data always *decreases* entropy!



$H(X, p) \approx 1.04$ $\qquad$ $H(Y, q) \approx 0.56$

We have

$$\begin{aligned}\mathrm{Loss}(g \circ f) &= H(X,p) - H(Z,r) \\ &= H(X,p) - H(Y,q) + H(Y,q) - H(Z,r) \\ &= \mathrm{Loss}(f) + \mathrm{Loss}(g)\end{aligned}$$

So, information loss is a *functor* from FinProb to some category with numbers in $[0,\infty)$ 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

$$\mathrm{Loss}\colon \mathrm{FinProb} \to [0, \infty)$$

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

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

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

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

be the disjoint union of X and Y, with the probability distribution given by λp on X and $(1 - \lambda)q$ 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

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

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

We can show entropy loss is **convex linear**:

$$\mathrm{Loss}\big(\lambda f + (1-\lambda)g\big) = \lambda \mathrm{Loss}(f) \ + \ (1-\lambda)\mathrm{Loss}(g)$$

This follows from the chain rule:

$$H(\lambda(X,p) + (1-\lambda)(Y,q)) = H_\lambda \ + \ \lambda H(X,p) \ + \ (1-\lambda)H(Y,q)$$

where

$$H_\lambda = -\Big(\lambda \ln \lambda \ + \ (1-\lambda)\ln(1-\lambda)\Big)$$

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

FinProb and $[0, \infty)$ are also **topological categories**: they have topological spaces of objects and morphisms, and composition of morphisms is continuous.

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

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

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

is a constant multiple of the entropy loss: for some $c \geq 0$,

$$g\colon (X, p) \to (Y, q) \quad \Longrightarrow \quad F(g) = c\ \mathrm{Loss}(g)$$

The easy part of the proof: show that

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

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

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

This boils down to Faddeev's theorem.

There are many generalizations!

There is precisely a one-parameter family of convex structures on the category $[0, \infty)$. For each one, there is an entropy loss functor

$$\mathrm{Loss}_q \colon \mathrm{FinProb} \to [0, \infty)$$

that is continuous and convex-linear. It is defined using Tsallis entropy:

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

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

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

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

There is a category-theoretic characterization of relative entropy:

- John Baez and Tobias Fritz, A Bayesian characterization of relative entropy, 2014.

Later Leinster gave a simplified proof in the case where $q_x = 0 \Rightarrow p_x = 0$, and some generalizations:

- Tom Leinster, A short characterization of relative entropy, 2017.

Relative entropy generalizes nicely to *infinite* measurable spaces:

$$I(\mu, \nu) = \int_X \ln\left(\frac{d\mu}{d\nu}\right) d\mu$$

where μ, ν are probability measures, μ is absolutely continuous with respect to ν, and $d\mu/d\nu$ is the Radon–Nikodym derivative.

Gagné and Panagaden generalized the categorical characterization of relative entropy to this case:

- Nicolas Gagné and Prakash Panangaden, A categorical characterization of relative entropy on standard Borel spaces, 2017.

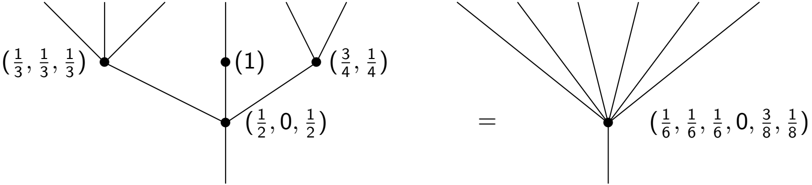
Parzygnat generalized the categorical characterization of Shannon information to the quantum case:

- Arthur Parzygnat, A functorial characterization of von Neumann entropy, 2020.

He is now working toward a categorical characterization of the quantum version of *relative* entropy:

- Arthur Parzygnat, Towards a functorial description of quantum relative entropy, 2021.

Also, this picture should remind you of 'operads', a formalism for composing operations in a tree-like way:



Leinster's thoughts on this topic led him to characterize Shannon entropy using operads:

- Tom Leinster, An operadic introduction to entropy, 2011.

Our work with Tobias Fritz was an attempt to *simplify* this beautiful but rather abstract result.

Bradley has recently given another characterization of entropy using operads:

- Tai-Danae Bradley, Entropy as a topological operad derivation, 2021.

And this is what she'll talk about next!