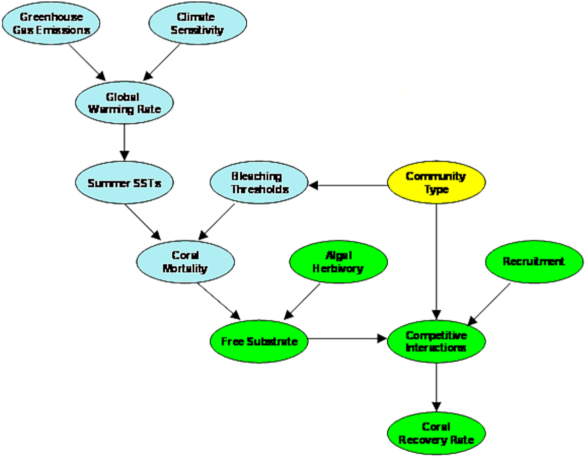
Network Theory III: Bayesian Networks, Information and Entropy

John Baez, Brendan Fong, 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 $\boldsymbol{y}$ given $\boldsymbol{x}$.**

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 $\mathtt{FinStoch}$ 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

$$\begin{tikzcd} & 1 \arrow[dl, rightsquigarrow, "p"'] \arrow[dr, rightsquigarrow, "q"] & \\ X \arrow[rr, "f"'] & & Y \end{tikzcd}$$

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

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

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

We can compose measure-preserving maps:

$$\begin{tikzcd}
 & 1 \arrow[dl, squiggly, "p"'] \arrow[d, squiggly, "q"'] \arrow[dr, squiggly, "r"] & \\
X \arrow[r, "f"'] & Y \arrow[r, "g"'] & Z
\end{tikzcd}$$

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:

$$\begin{tikzcd}
 & 1 \arrow[dl, squiggly, "p"'] \arrow[dr, squiggly, "q"] & \\
X \arrow[rr, "f"'] & & Y
\end{tikzcd}$$

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

The **data processing inequality** says that

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

Deterministic processing of random data always *decreases* entropy!

$$
\begin{tikzcd}
& 1 \arrow[dl, squiggly, "p"'] \arrow[d, squiggly, "q"'] \arrow[dr, squiggly, "r"] & \\
X \arrow[r, "f"'] & Y \arrow[r, "g"'] & Z
\end{tikzcd}
$$

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 $\mathtt{IL}$ is 'convex-linear' and 'continuous'.

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

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

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.

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

$\texttt{IL}\colon \texttt{FinProb} \to [0,\infty)$ 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 $\alpha \geq 0$,

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

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

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

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:

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

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

$$\begin{tikzcd}
 & 1 \arrow[dl, squiggly, "p"'] \arrow[dr, squiggly, "q"] & \\
X \arrow[rr, bend right, "f"'] & & Y \arrow[ll, squiggly, bend right, "s"']
\end{tikzcd}$$

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

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`

$$\begin{tikzcd}
& 1 \arrow[dl, rightsquigarrow, "p"'] \arrow[dr, rightsquigarrow, "q"] & \\
X \arrow[rr, bend right, "f"'] & & Y \arrow[ll, rightsquigarrow, bend right, "s"']
\end{tikzcd}$$

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

gives a relative entropy $S(p, s \circ q)$. 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.

$$\begin{tikzcd} & 1 \arrow[dl, squiggly, "p"'] \arrow[dr, squiggly, "q"] & \\ X \arrow[rr, bend right, "f"'] & & Y \arrow[ll, squiggly, "s"'] \end{tikzcd}$$

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

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

Morphisms with an optimal hypothesis form a subcategory

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

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

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

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

The proof is hard! Can you simplify it?

The category `FinStoch` and its big brother `Stoch` also appear in the work of Brendan Fong:

- *Causal Theories: a Categorical Perspective on Bayesian Networks*.

$$
\begin{tikzcd}
a \arrow[r] \arrow[dr] \arrow[dd, bend right] & b \arrow[d] & & c \arrow[ddl, bend left] \\
& d \arrow[dl] \arrow[dr] & & \\
e & & f &
\end{tikzcd}
$$

As usual in Bayesian network theory, he starts with a directed acyclic graph G where, intuitively speaking:

- each vertex is a ‘variable’
- each directed edge $a \to b$ is a ‘causal relationship’: the value of a may affect that of b.

Roughly speaking, starting from a directed acyclic graph G, he forms the category with finite products $\mathcal{C}_G$ freely generated by:

- one object for each vertex of G,
- one morphism $f_b\colon a_1 \times \cdots \times a_n \to b$ whenever a_i are all the parents of b:

$$\begin{tikzcd}
a_1 \arrow[drr] & a_2 \arrow[dr] & \cdots & a_{n-1} \arrow[dl] & a_n \arrow[dll] \\
& & b & &
\end{tikzcd}$$

(and thus $f_b\colon 1 \to b$ if b has no parents).

This category $\mathcal{C}_G$ is the **causal theory** described by the graph G. A **model** of this theory in $\mathtt{FinStoch}$ is a symmetric monoidal functor

$$F \colon \mathcal{C}_G \to \mathtt{FinStoch}$$

This gives

- a finite set $F(b)$ for each vertex b of the graph
- a probability measure $F(f_b) \colon 1 \rightsquigarrow F(b)$ for each vertex with no parents
- a stochastic map $F(f_b) \colon F(a_1) \times \cdots \times F(a_n) \rightsquigarrow F(b)$ whenever a_i are all the parents of b
- and thus a random variable for each vertex
- *automatically obeying the 'independence' assumptions we want in Bayesian network theory!* If two vertices have no common ancestors, their random variables are stochastically independent.

So: we're starting to see how category theory connects

- signal flow diagrams
- electrical circuit diagrams
- stochastic Petri nets
- chemical reaction networks
- Bayesian networks, entropy and information

These connections can help us develop a unified toolkit for modelling complex systems made of interacting parts... like living systems, and our planet.

But there's a lot of work to do! Please help. Check this out:

The Azimuth Project

www.azimuthproject.org