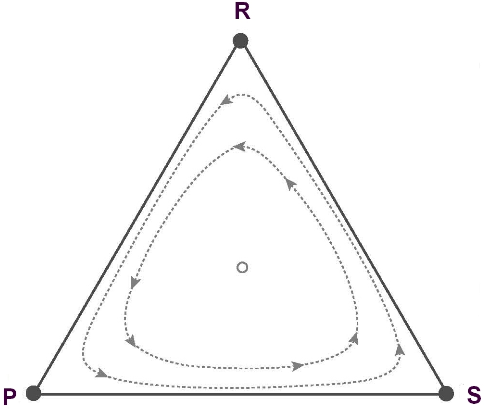
INFORMATION THEORY
IN POPULATION DYNAMICS



John Baez

Queen Mary University, 2022 December 6

IT'S ALL RELATIVE — EVEN INFORMATION!

When you learn something, how much information do you gain?

It depends on what you believed before!

We can model hypotheses as probability distributions. When you update your prior hypothesis p to a new one q, how much information do you gain?

This much:

$$I(q,p) = \sum_{i=1}^{n} q_i \ln\left(\frac{q_i}{p_i}\right)$$

This is the **information of q relative to p**, also called the 'information gain' or 'Kullback–Leibler divergence'.

$$I(q,p) \geq 0 \qquad \text{and} \qquad I(q,p) = 0 \iff q = p$$

For example, suppose we flip a coin you think is fair. Your prior hypothesis is this:

$$p_H = \frac{1}{2} \qquad p_T = \frac{1}{2}$$

Then you learn it landed heads up:

$$q_H = 1 \qquad q_T = 0$$

The relative information is 1 bit:

$$I(q,p) = 1 \ln\left(\frac{1}{1/2}\right) + 0 \ln\left(\frac{1}{1/2}\right) = \ln 2$$

where we define $0 \ln 0 = 0$. You have gained 1 bit of information.

But suppose you think there's only a 25% chance of heads:

$$p_H = \frac{1}{4} \qquad p_T = \frac{3}{4}$$

Then you learn the coin landed heads up:

$$q_H = 1 \qquad q_T = 0$$

Now the relative information is higher:

$$I(q,p) = 1 \ln\left(\frac{1}{1/4}\right) + 0 \ln\left(\frac{1}{3/4}\right) = \ln 4 = 2 \ln 2$$

You have gained 2 bits of information!

MAXIMIZING ENTROPY

Say we have a set of states $i = 1, 2, 3, \ldots$. Let E_i be any real-valued function of states — let's call it the **energy**.

If the probability of the system being in its ith state is q_i, then its **Shannon entropy** is

$$S(q) = -\sum_i q_i \ln(q_i)$$

Any probability distribution p that maximizes $S(p)$ subject to a constraint on

$$\langle E \rangle_p = \sum_i p_i E_i$$

is a **Boltzmann distribution**

$$p_i \propto \exp(-E_i/T)$$

for some value of the real parameter T, called **temperature**.

FREE ENERGY

The **free energy** of a probability distribution q at temperature T is defined to be

$$\begin{aligned} F(q) &= \langle E \rangle_q - TS(q) \\ &= \sum_i q_i E_i \; + \; T \sum_i q_i \ln(q_i) \end{aligned}$$

The basic reason free energy is interesting:

maximizing $S(q)$ for some fixed value of $\langle E \rangle_q$

is the same as

minimizing $F(q)$ for some fixed value of T.

FREE ENERGY AND RELATIVE ENTROPY

Supose p is the probability distribution that maximizes Shannon entropy for some fixed value of $\langle E \rangle_p$. Suppose q is any other probability distribution on the set of states.

A fun calculation shows that

$$\boldsymbol{F(q) - F(p) = T\, I(q,p)}$$

where $F(q)$ is the free energy of q and $F(p)$ is the free energy of the Boltzmann distribution — thus, the least possible free energy at temperature T.

NATURAL SELECTION AND RELATIVE INFORMATION

Suppose we have self-replicating entities of different kinds:

- organisms belonging to different species
- genes of different alleles
- restaurants belonging to different chains
- people with different beliefs
- game-players with different strategies
- instances of a genetic algorithm
- etc.

I'll call them **replicators** of different **species**.

Let P_i be the population of the ith species, as a function of time.

Suppose the **Lotka-Volterra equation** holds:

$$\frac{dP_i}{dt} = f_i \, P_i$$

where $f_i = f_i(P_1, \ldots, P_n)$, the **fitness** of the ith species, is any function of the populations of all the species.

The probability that a randomly chosen replicator belongs to the *i*th species is

$$p_i = \frac{P_i}{\sum_j P_j}$$

If we think of each species as a *strategy*, the probability distribution p_i can be seen as a *hypothesis about the best strategy*. The quality of a strategy is measured by its fitness f_i.

We can show p_i evolves according to the **replicator equation**:

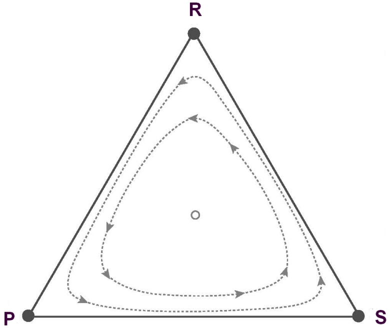
$$\frac{dp_i}{dt} = (f_i - \langle f \rangle) p_i$$

where $\langle f \rangle = \sum_j f_j \, p_j$ is the **mean fitness**.

The replicator equation is a continuous-time version of *Bayesian hypothesis updating*:

- Marc Harper, The replicator equation as an inference dynamic, arXiv:0911.1763.

But the probability distribution $p_i(t)$ may not converge to an equilibrium:



INFORMATION GEOMETRY

How can we quantify the *rate of learning?*

Here we face a ‘paradox’:

For any probability distribution $p(t)$ that changes smoothly with time, we have

$$\frac{d}{dt} I(p(t), p(t_0)) \Big|_{t=t_0} = 0$$

for all times t_0.

“To first order, you’re never learning anything.”

However, as long as the velocity $\dot{p}(t_0)$ is nonzero, we have

$$\frac{d^2}{dt^2} I(p(t), p(t_0)) \Big|_{t=t_0} > 0$$

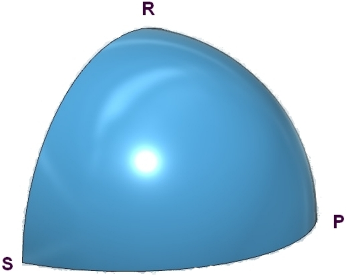
"To second order, you're always learning something... unless your opinions are fixed."

This lets us define a 'rate of learning' — that is, the 'speed' of the changing probability distribution $p(t)$.

Namely, define the length of the vector $\dot{p}(t_0)$ by

$$\|\dot{p}(t_0)\|^2 = \frac{d^2}{dt^2} I(p(t), p(t_0))\Big|_{t=t_0}$$

This notion of length defines a 'Riemannian metric' on the space of probability distributions: the **Fisher information metric**. This makes the space of probability distributions round:



Now suppose we have populations obeying the replicator equation:

$$\frac{dP_i}{dt} = f_i \, P_i$$

so the probability distribution of species evolves via

$$\frac{dp_i}{dt} = (f_i - \langle f \rangle) p_i$$

Then a fun calculation shows

$$\boldsymbol{\left\| \frac{dp}{dt} \right\|^2 = \sum_i (f_i - \langle f \rangle)^2 p_i}$$

The square of the rate of learning is the variance of the fitness!

In 1972, George R. Price wrote:

> *It has long been a mystery how Fisher derived his famous 'fundamental theorem of Natural Selection' and exactly what he meant by it.*
>
> *He compared this result to the second law of thermodynamics, and described it as holding 'the supreme position among the biological sciences'. Also, he spoke of the 'rigour' of his derivation of the theorem and of 'the ease of its interpretation'. But others have variously described his derivation as 'recondite' (Crow & Kimura), 'very difficult' (Turner), or 'entirely obscure' (Kempthorne). And no one has ever found any other way to derive the result that Fisher seems to state.*

I claim this equation

$$\left\| \frac{dp}{dt} \right\|^2 = \sum_i (f_i - \langle f \rangle)^2 p_i$$

is the theorem Fisher *should* have stated — but did not.

For more, see:

- John Baez and Blake Pollard, Relative entropy in biological systems, *Entropy* **18** (2016), 46.
- John Baez, The fundamental theorem of natural selection, *Entropy* **23** (2021), 1436.