Information and Entropy in Biological Systems

John Baez
http://www.nimbios.org/wordpress-training/entropy/

April 8, 2015
NIMBioS

Goal: to unify various ways that information and entropy are used in biology.

For example:

- biological communication systems
- the 'action-perception loop'
- the thermodynamic foundations of biology
- the structure of ecosystems
- measures of biodiversity
- evolution

The **Shannon entropy** of a probability distribution $p\colon S \to [0,1]$ on a set S is

$$H(p) = -\sum_{i \in S} p(i)\, \log(p(i))$$

It says how much information we learn upon discovering the value of an element of S that was randomly chosen according to this probability distribution.

We use base 2 for our logarithm if we want to measure information in **bits**. But base e is also natural: then we're measuring information in **nats**.

All this generalizes from sums to integrals, but let's not now.

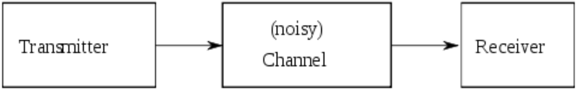
Shannon was concerned with communication.

His **source coding theorem** puts a bound on how much you can compress a signal — a string of symbols — in which each symbol is independently chosen at random from a set S, with probability distribution p.

In the limit of long signals, you can find a way to encode each symbol using a string of $H(p) + \epsilon$ bits, with ϵ probability of error, where $\epsilon > 0$ is as small as you want.

You cannot use $< H(p)$ bits to encode each symbol while still achieving an arbitrarily small probability of error.

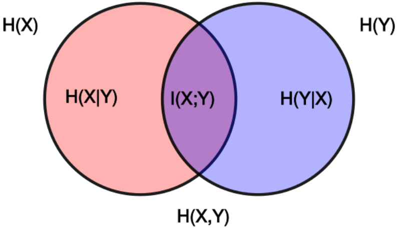
Shannon's **noisy-channel coding theorem** generalizes this to communication channels with noise.



It says how to compute the **channel capacity**: the maximum number of bits per code word that can be transmitted with arbitrarily small error probability.

I won't explain how. But the key idea is **mutual information**: how much information two random variables have in common.

A ‘pair of random variables’ X and Y is a probability distribution on a set $S \times T$. This gives probability distributions on S and on T. We may thus define three entropies: the **joint entropy** $H(X, Y)$ and the **individual entropies** $H(X)$ and $H(Y)$.



The **mutual information** is

$$I(X; Y) = H(X, Y) - H(X) - H(Y)$$

The **entropy of *X* conditioned on *Y*** is

$$H(X|Y) = H(X, Y) - H(Y)$$

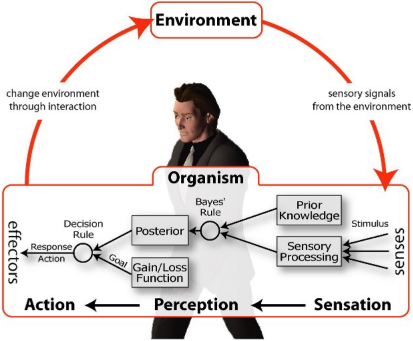
Shannon's other main achievement was founding **rate distortion theory**.

Here we put a *distance function* on our set S of symbols, and seek to encode them in way that lets the signal be reconstructed *to within some distance* $d > 0$, called the **distortion**. We seek to do this using the minimum number of bits per symbol.

All these ideas of Shannon may be important in understanding:

- communication between organisms
- the nervous system, which communicates signals via nerve impulses and neurotransmitters
- other forms of intercellular communication, for example via hormones and cytokines
- intracellular communication, for example via gene expression and gene regulation

If biological communication is near-optimized by evolution, we may use Shannon's ideas on optimal communication — *appropriately generalized* — to help generate testable hypotheses. But beware: in biology, communication is always just a means to an end.



University of Bielefeld

Natural selection maximizes fitness, not 'bits per second'.

Communication typically deals with a few bits — or terabytes — of *relevant* information. The complete description of a physical object uses vastly more information, most of which is *irrelevant* for understanding its macroscopic properties. Physicists call this irrelevant information **entropy**.

- your genome: 10^{10} bits.
- all words ever spoken by human beings: $\sim 4 \times 10^{19}$ bits.
- genomes of all living humans: 6×10^{19} bits.
- one gram of water at room temperature: 4×10^{24} bits.

The tendency for information to shift from more relevant to less relevant forms — the **Second Law of Thermodynamics** — underlies chemistry and thus biology.

Maximizing entropy is a powerful way to choose hypotheses.

The **maximum entropy method** for choosing a probability distribution $p\colon S \to [0,1]$ says we should maximize

$$H(p) = -\sum_{i\in S} p(i)\, \log(p(i))$$

subject to whatever constraints we want p to obey.

For example, suppose we have a function $f\colon S \to \mathbb{R}$ and we want to choose p that maximizes $H(p)$ subject to the constraint that the expected value of f is some number c:

$$\sum_{i \in S} p(i) f(i) = c$$

Then we should choose a **Boltzmann distribution**:

$$p(i) = \frac{e^{-\beta f(i)}}{\displaystyle\sum_{i \in S} e^{-\beta f(i)}}$$

Which β should we choose? It depends on which c we want.

All this generalizes painlessly when we have a collection of functions $f_1, \ldots, f_n\colon S \to \mathbb{R}$.

Physicists have developed the maximum entropy method to a high art when S is the set of states of a physical system in thermal equilibrium.

Jaynes emphasized that we can use all this machinery more generally. For example: we can let S be a set of species, and $p(i)$ be the probability that an organism belongs to the ith species. Ideas of this type underlie John Harte's work on ecology.

Intriguingly, in this case the entropy $H(p)$, and generalizations like the Rényi entropy $H_q(p)$, are widely used as measures of biodiversity!

Is there a sense in which nature maximizes biodiversity subject to constraints?

The truth seems to be more complicated....

Let

$$P = (P_1, \ldots, P_n)$$

be the vector of populations of n different self-replicating entities: for example, species of organisms.

The probability that an organism belongs to the ith species is

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

We can think of this probability distibution as a 'hypothesis' and its change with time as a 'learning process'. Natural selection is analogous to Bayesian updating.

Let p and q be a two probability distributions. The **information of q relative to p**, or **Kullback–Leibler divergence**, is

$$I(q||p) = \sum_i q_i \ln\left(\frac{q_i}{p_i}\right)$$

This is the amount of information *left to learn* if p is our current hypothesis and q is the 'true' probability distribution describing a situation.

In Bayesian language, p is our **prior**.

Suppose the population $P(t)$ evolves according to the **replicator equation**:

$$\frac{d}{dt}P_i(t) = F_i(P_1(t), \dots, P_n(t))\ P_i(t)$$

where F_i, the **fitness** of the ith species, depends smoothly on all the populations.

Suppose q is a 'dominant distribution' — a distribution of species whose mean fitness is at least as great as that of any other distribution it could find itself amidst. Then Shahshahani proved

$$\frac{d}{dt}I(q||p(t))\ \leq\ 0$$

As time passes, the information the population has 'left to learn' never increases.

Reality is more complicated; Marc Harper will say more.