Characterizations of Shannon and Rényi entropy

Tobias Fritz

NIMBioS, April 2015

Entropy and Information in Biological Systems

Shannon entropy

Let $p : S \to [0, 1]$ is a probability distribution on a finite set S.

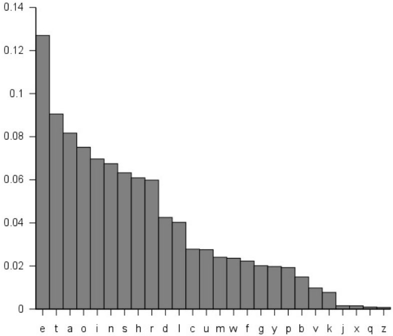
- **Shannon entropy** is defined to be

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

Possible interpretations:

- $H(p)$ measures the **amount of randomness** in p.
- $H(p)$ measures the **amount of information** that we gain when learning a particular $i \in S$.
- The exponentiated entropy $e^{H(p)}$ measures the **effective size** of p.
- $H(p)$ measures the **compressibility** of a sequence of elements sampled from p.

Example: relative frequencies of letters in English



English has an entropy of about

$$2.9 \text{ nats} \approx 4.1 \text{ bits}$$

per letter. BUT: this incorrectly ignores correlations between subsequent letters. Taking these into account results in a much lower value. English text can be compressed to a fraction of its length!

Why Shannon entropy?

Why do we use Shannon entropy as a measure of information?

In information theory applications, the answer is given by the **asymptotic equipartition property:**

- There is $T \subseteq S^n$ with

$$|T| \leq e^{n(H(p)+\varepsilon)}$$

  such that sampling n times from p yields an element of T with probability $> 1 - \varepsilon$, and $\varepsilon \to 0$ as $n \to \infty$.

- T is the **typical set** whose size is governed by the entropy. Basic idea: in n coin flips, we expect roughly $\frac{n}{2}$ heads and $\frac{n}{2}$ tails. Sequences which strongly deviate from this are "untypical".

- $H(p)$ is the smallest number with the above property.

- This is Shannon's source coding theorem on compressibility.

Rényi entropies I

This explains why Shannon entropy is so ubiquitous in information theory. But could other measures of information be useful in other contexts?

- For $\beta \in [0, \infty]$, the **Rényi entropy of order** β is given by

$$H_\beta(p) = \frac{1}{1-\beta} \log \left(\sum_{i \in S} p_i^\beta \right).$$

- The scaling factor is conventional: it makes H_β nonnegative for all β and ensures $H_\beta(u_n) = \log n$, where u_n is the uniform distribution on an n-element set.

- The main property which the Rényi entropies have in common with Shannon entropy is **additivity**:

$$H_\beta(p \times r) = H_\beta(p) + H_\beta(r).$$

Rényi entropies II

Interesting special cases:

- For $\beta = 0$, we obtain the **max entropy**, which is the cardinality of the support of p:

$$H_0(p) = \log |\{ i \in S \mid p(i) > 0 \}|.$$

- For $\beta = 1$, we recover Shannon entropy:

$$\begin{aligned} H_1(p) &= \lim_{\beta \to 1} H_\beta(p) \\ &= \frac{d}{d\beta} \left(\frac{1}{1-\beta} \log \left(\sum_i p(i)^\beta \right) \right)_{\beta=1} = -\sum_i p(i) \log p(i). \end{aligned}$$

- For $\beta = \infty$, we obtain the **min entropy**:

$$H_\infty(p) = -\log \max_i p(i) = \log \min_i \frac{1}{p(i)}$$

Rényi entropies III

The **partition function**

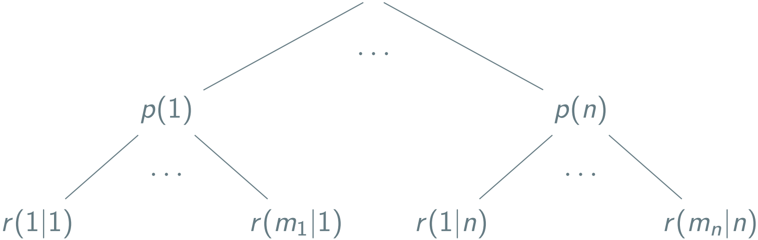
$$Z_p(\beta) = \sum_{i \in S} p(i)^\beta = e^{(1-\beta)H_\beta(p)}$$

provides an alternative point of view on the Rényi entropies.

- Knowing the partition function lets us recover p up to permutations of the outcomes $i \in S$.
- So if we know all the Rényi entropies, we also know p up to permutations.
- This is one way to explain why the Rényi entropies are useful: every other invariant quantity can be expressed in terms of the H_β's.

The chain rule I

Consider an ecosystem inhabited by taxonomic families $i \in F$ with relative abundances $p(i)$. In each family, there are species $j \in S_i$ with relative abundances $r(j|i)$. We get a taxonomic tree labelled by relative frequencies like this:



So each $q(-|i)$ is a probability distribution itself. But the *overall* relative abundance of species j of family i is given by

$$(r \circ p)(i, j) = p(i) \cdot r(j|i).$$

The chain rule II

The Shannon entropy as a diversity measure has the following appealing property:

$$H(r \circ p) = H(p) + \sum_i p_i H(r(-|i)).$$

- In Shannon's own words: if a choice is broken down into two successive choices, the original H should be the weighted sum of the individual values of H [weighted by the relative frequency with which each choice occurs].
- Known under various names such as **chain rule**, **glomming formula**, etc.
- This is a surprising property of Shannon entropy not satisfied by the other Rényi entropies.

Faddeev's characterization

Theorem (Faddeev 1956)

The chain rule, together with permutation invariance and continuity, characterize Shannon entropy up to a constant multiple.

- So if you want your measure of information to satisfy the chain rule, you are essentially forced to use Shannon entropy!
- In 2011, Baez, Fritz and Leinster reformulated this characterization in terms of three very natural axioms on the change of information under deterministic processing.
- There are intriguing connections to group cohomology.

The minimal requirements I

What are minimal requirements that we could impose on a measure of information?

- It should be additive under product measures, i.e. the amount of information in $p \times q$ should be the sum of the amount of information in p and q individually,

$$H(p \times r) = H(p) + H(r).$$

- If $p(i) > p(j)$, then moving a bit of weight from $p(i)$ to $p(j)$ makes the distribution unambiguously more random. $\Longrightarrow$ The measure of information should not decrease under this operation.

- This is equivalent to postulating that if p majorizes r, then $H(p) \leq H(r)$.

The minimal requirements II

- All the Rényi entropies H_β satisfy both of these properties.
- So do all positive linear combinations of Rényi entropies.
- More generally, so do all integrals of Rényi entropies, i.e. information measures of the form

$$p \longmapsto \int_0^\infty H_\beta(p)\, f(\beta)\, \mathrm{d}\beta$$

  for some nonnegative weight function (measure) f.

Conjecture (Fritz 2015)

Every measure of information satisfying both properties is of this form.