Self-information

Self-information attempts to describe the amount of information that we gain from a specific outcome of an experiment. Intuitively, if an experimental outcome gives us a "surprising" result, then we gained more information from the experiment than if the result had been something we were expecting. In essence, a surprising result changes our understanding of the world whereas an expected result supports our current understanding. We see that intuitively "surprise" is a good measure of the information content of a result. For this reason, the definition of self-information is built upon the probability of a result (Definition 1).

The value of the self-information of an event depends on the base of the logarithm used in I. Different bases thus define which units we are using to measure the quantity of the information. Some common units of information are:

- Base 2: bit
- Base 3: trit
- Base e: nat

As can be shown in Shannon's Coding Theorem, the base of the logarithm can be interpreted as the size of an alphabet that is used to communicate the outcome of the event. The base 2 logarithm asserts that the occurrence of the event is communicated using bits (i.e. ones and zeros) whereas a base 3 logarithm can be interpreted as communicating the outcome using trits (i.e. zeros, ones, and twos).

Definition 1 ***Self-information*** *is a function* I

$$I : [0, 1] \to [0, \infty)$$

defined as

$$\begin{aligned} I(p) &:= \log \frac{1}{p} \\ &= -\log p \end{aligned}$$

where $p \in [0, 1]$.

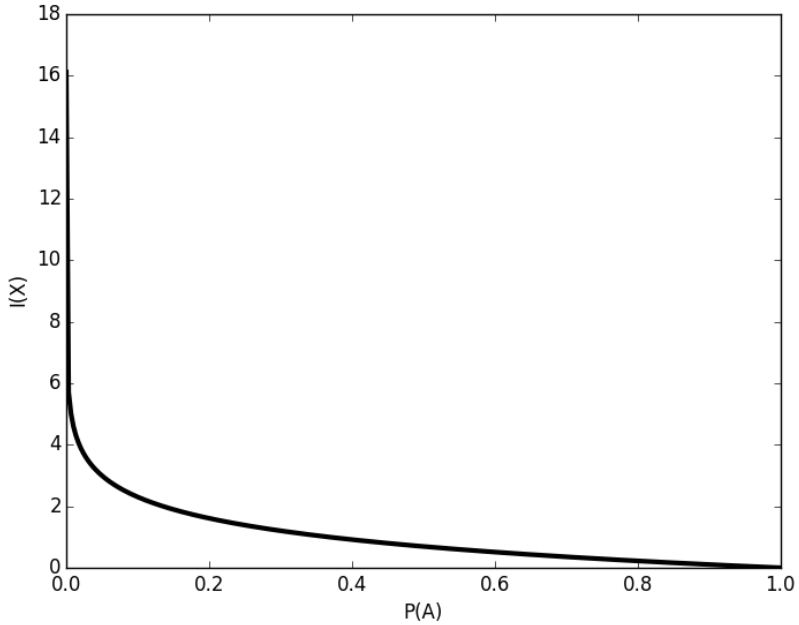

Figure 1: The self-information of an event A versus its probability. Above, we see that the self-information of a zero-probability event is infinite. This makes intuitive sense; the occurrence of an event with zero probability would be infinitely surprising. Likewise, an event that is sure to occur would not be surprising at all; hence it's self-information is zero.

Intuition

We see that the amount of "surprise" of an outcome is a function of the probability of the outcome and not on the value of the outcome itself. Therefore, it makes sense to define I to be function on the probability of the outcome p rather than the value of the outcome. Furthermore, the definition of self-information has several characteristics that make it a good choice for encoding this notion of "surprise."

- Self-information is a continuous function thereby encoding our intuition that the amount of "surprise" exists on a continuum.
- Self-information is a non-negative quantity

  $$I(p) \geq 0$$

  That is, we cannot be negatively surprised.
- Self-information is a monotonic function

  $$p_1 < p_2 \implies I(p_1) > I(p_2)$$

  That is, the smaller the probability of an event is to occur, the greater is our surprise when it occurs.
- If an event has probability 1, then the self-information content is zero.

  $$I(1) = 0$$

  That is, if an even is certain to occur, we gain no information from it.
- The self-information from two independent events is the sum of the self-informations from each individual event:

  $$I(p_1 * p_2) = I(p_1) + I(p_2)$$

  Generalizing to more than two events:

  $$I(p^n) = nI(p)$$

  This encodes the intuition that if we perform two independent experiments, then the amount of self-information we've gained is a sum of the self-informations from each individual experiment.