2.3. Entropy

The last section shows that a lower-bound on the "rate" of an encoding (i.e., the compressibility of the encoded string) is governed by:

$$H(X) = -\sum_{s \in \Sigma} \log_2 p(X = s) \cdot p(X = s)$$

It turns out that $H(X)$ is a more general measure that applies to all random variables called *entropy*. If X is a discrete random variable with range $\mathcal{X}$:

$$H(X) = -\sum_{x \in \mathcal{X}} \log_2 p(X = x) \cdot p(X = x)$$

and likewise if X is a continuous random variable:

$$H(X) = -\int_{x=-\infty}^{\infty} \log_2 p(x) p(x) dx$$

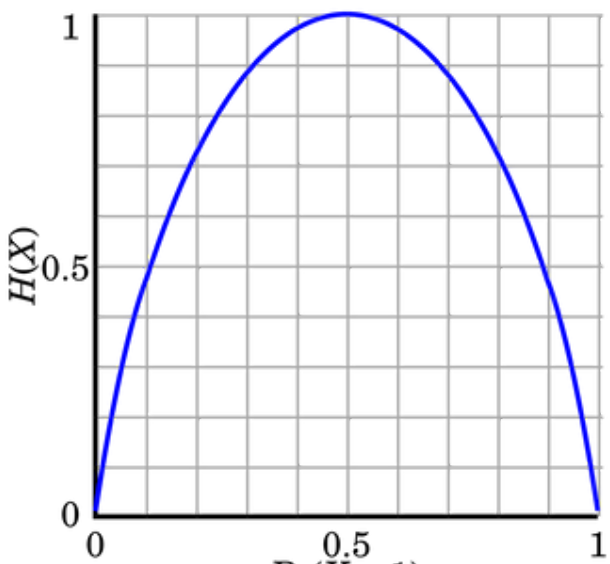
The Shannon source coding theorem describes entropy in terms of "compressibility" but we can more generally interpret it as the average level of "information", "surprise", or "uncertainty" inherent in a random variable's possible outcomes.

2.3.1. Examples

Let's consider a Bernouilli Random Variable X with parameter p (which models an unfair coin flip). The entropy is:

$$H(X) = -p \log_2 p - (1 - p) \log_2 (1 - p)$$

Let's plot this function out as a function of the parameter p.



The entropy is highest when the coin is perfectly fair (there is no bias towards heads or tails). The entropy quickly drops as we make the coin less fair. Intuitively, the more biased the coin is, the less "surprising" the outcomes of the random variable are.

Let's look at a case of differential entropy. The entropy of a normal distribution parametrized by μ and σ is:

$$\frac{1}{2} \log(2\pi e \sigma^2)$$

As the variance σ increases, the random variable has higher entropy.

2.3.2. Properties

Beyond being a lower bound on data compression. Entropy is, in a sense, the right way to measure the information content in a random variable. To understand the meaning of $-\sum_{x \in \mathcal{X}} \log_2 p(X = x) \cdot p(X = x)$, let's first define an information function I in terms of an event e. The information in event i is:

$$I(e) = \log(\frac{1}{\mathbf{Pr}[e]}) = -\log(\mathbf{Pr}[e])$$

$I(e)$ captures the rarity of certain events. Based on this definition of information, we can see that entropy is an expected "rarity" or "suprise" of a random variable. In the discrete case:

$$H(X) = \mathbf{E}[I(X = x)]$$

The amount of information acquired due to the observation of event e has certain fundementa properties:

- $I(e)$ is monotonically decreasing in p_e: an increase in the probability of an event decreases the information from an observed event, and vice versa.
- I(e) ≥ 0: information is a non-negative quantity.
- I(1) = 0: events that always occur do not communicate information.
- I(e1, e2) = I(e1) + I(e2): the information learned from independent events is the sum of the information learned from each event.

By Sanjay Krishnan