Shannon entropy

Shannon entropy of a chance event with probabilities $\mathbf{p} = (p_1, \ldots, p_n)$ is defined by

$$H(\mathbf{p}) = H(p_1, \ldots, p_n) = \sum_{i=1}^{n} -p_i \log p_i.$$

This is a measure of how much "choice" is involved in the selection of the outcome.

Notation: If X denotes the random outcome, we also write $H(X)$ to denote the entropy.

If log is to base 2, entropy is in "bits."

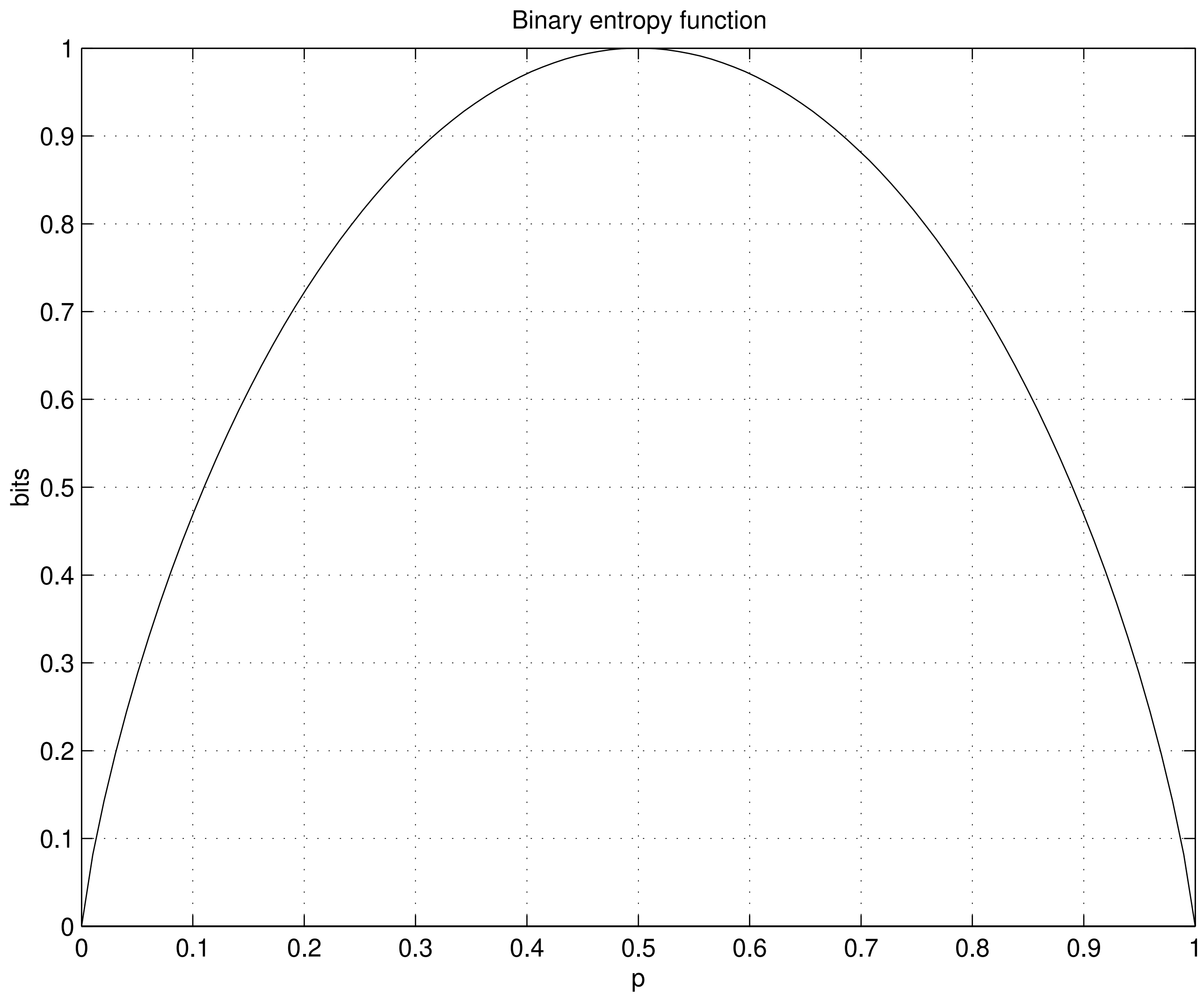
Take $0 \cdot \log 0 = 0$ in entropy calculations.

Binary entropy

The entropy for the case of two outcomes with probabilities p and $1-p$,

$$H(p) = -p \log p - (1-p)\log(1-p),$$

is plotted below.



Some properties of entropy

1. $H(p_1, \ldots, p_n) \geq 0$ with equality iff $p_i = 1$ for some i.

2. $H(p_1, \ldots, p_n) \leq \log n$ with equality iff $p_i = 1/n$ for all i.

3. Concavity. For any

$$\mathbf{p} = (p_1, \ldots, p_n); \;\; \mathbf{q} = (q_1, \ldots, q_n)$$

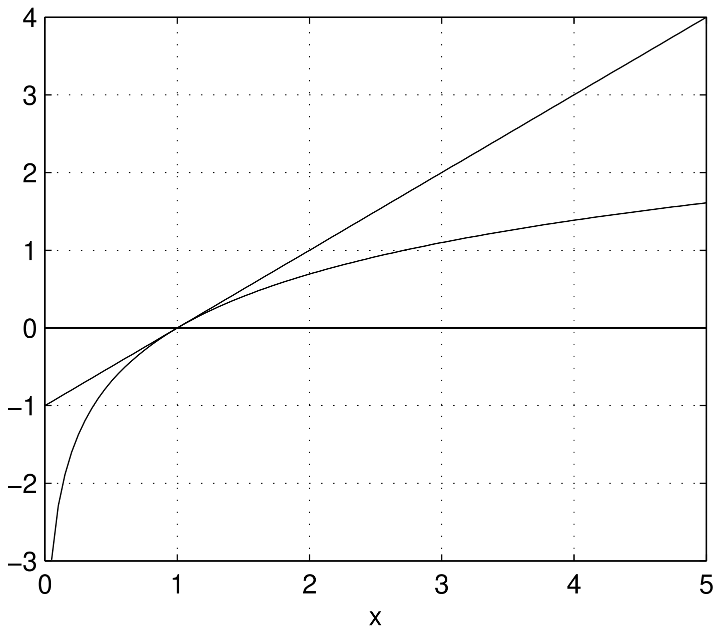
$$\mathbf{r} = \lambda\mathbf{p} + (1 - \lambda)\mathbf{q}, \;\; 0 \leq \lambda \leq 1,$$

$$H(\mathbf{r}) \geq \lambda H(\mathbf{p}) + (1 - \lambda)H(\mathbf{q}).$$

A useful inequality

$$\ln x \leq x - 1, \quad \text{for all } x > 0$$

with equality iff $x = 1$.



Proof of property 2

$$
\begin{aligned}
H(\mathbf{p}) - \log(n) &= \log(e)\left[\sum_i p_i \ln(1/p_i) - \ln(n)\right] \\
&= \log(e)\left[\sum_i p_i \ln(\frac{1}{np_i})\right] \\
&\leq \left[\sum_i p_i(\frac{1}{np_i} - 1)\right] \\
&= \sum_i 1/n - \sum_i p_i \\
&= 0.
\end{aligned}
$$

Entropy of an information source

Consider a discrete memoryless source (DMS) that emits the ith letter with probability p_i.

A typical message of length N by such a souce will contain about $p_i N$ occurrences of the ith symbol.

Hence, the probability p of such a typical message will be

$$p \approx p_1^{p_1 N} \cdots p_N^{p_N N} = 2^{N \sum_i p_i \log p_i} = 2^{-NH},$$

where $H = H(p_1, \ldots, p_n)$ is the entropy of the DMS.

Asymptotic equipartition property (AEP)

Consider a DMS with entropy H. Given any $\epsilon > 0$ and $\delta > 0$, we can find an $N_0(\epsilon, \delta)$ such that sequences of any length $N \geq N_0$ by the DMS fall into two classes:

1. A set whose total probability is less than ϵ.

2. The remainder, called the typical set $T = T(\epsilon, \delta)$, all of whose members have probabilities satisfying the inequality
$$2^{-N[H+\delta]} \leq p \leq 2^{-N[H-\delta]}.$$

The number of typical elements satisfies

$$(1-\epsilon)2^{N[H-\delta]} \leq |T| \leq 2^{N[H+\delta]}.$$

Source coding

To compress the output of a DMS, take messages of length $N \geq N_0(\epsilon, \delta)$ for a given ϵ, δ.

Enumerate the elements of the typical set $T = T(\epsilon, \delta)$ by integers 1 through $|T| \leq 2^{N[H+\delta]}$.

If the output $\mathbf{x} = (x_1, \ldots, x_N)$ of the DMS is in T, send its index using roughly NH bits.

If $\mathbf{x}$ is not typical, send it without compression using roughly $N \log K$ bits where K is the number of letters in the source alphabet.

On average, we send less than roughly

$$\epsilon \log K + (1 - \epsilon)H = H + O(\epsilon)$$

bits per source letter. This is Shannon's "noiseless source coding" theorem.