measuring information

info-20002: foundations of informatics

information theory

- Shannon's seminal paper
- Provide a framework to deal with the efficiency of information transmission in noisy channels
- Measure information quantity objectively
- Data as facts
- Devoid of the subjective aspects of information, no semantics and pragmatics.
- Information represents **the degree of freedom in choosing one particular symbol from all possible ones.**
- Bit as the fundamental unit of information.

Shannon, C. (1948). **A Mathematical Theory of Communication**, Bell System Technical Journal (27)
Klir, G. J., Wierman, M. J. (1999). **"Uncertainty-Based Information. Elements of Generalized Information Theory"**

Binary Code

- Two-symbol language: two alphabets "0" and "1", binary digits
- Bit is equivalent to the selection between two equally likely choices
- Information quantity depends on the number of alternative message choices encoded in the binary system
- Most economical information encoding (von Baeyer, 2004)

Consider encoding a number between 0 and 127 inclusive using a set of numbered flags. Number 126 can be represented using various encodings as follows:

Method	Required flags	Encoded number
One number each flag	128	126
Hexadecimal code	23 = $7 + (1 \times 16)$	7E
Decimal code	21 = $1 + (2 \times 10)$	126
Octal code	17 = $1 + (2 \times 8)$	176
Binary code	14 = 2×7	11111110

Measure of Information

Hartley function (Hartley uncertainty measure)

- Selecting up a particular symbol on a finite set X of uniformly discrete symbols.
- The amount of information needed to remove the uncertainty.
- The amount of uncertainty associated with a set of n alternatives.
- Quantifies how many b-choice questions need to be asked to uncover the selected symbol

$$H_0(X) = \log_b |X| = \log_b n$$

$b = 2$, the unit of measurement is called **shannon** (Sh).
$b = e$, the unit of measurement is called **nat** (for natural logarithmic).
$b = 10$, the unit of measurement is called **hartley** (Hart).

[1] Hartley, R. V. L. (1928). **"Transmission of information,"**, Bell System Technical Journal 7(3): 535-563

Think a card, and select the row in which that card appears.

Row 1

Row 2

Next

Row 3

Entropy of Information

- Shannon's Entropy probabilistic-based measurement
- Information as events (random variable), e.g. tossing coins, receiving a piece of text
- Hartley works for uniformly distributed symbol, Shannon's for non-uniform distribution
- Measure of the amount of uncertainty in selecting a symbol from symbol space
- The more unpredictable a symbol is, the higher its information value

Shannon, C. (1948). **A Mathematical Theory of Communication**, Bell System Technical Journal (27)

Entropy of Information

Let:

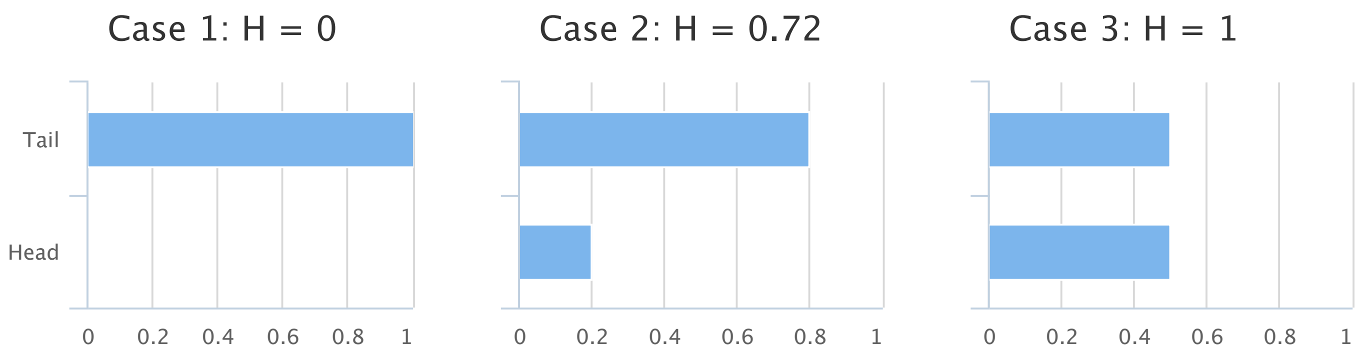
- a message space of n symbols in a possible space X. Probablistically speaking, X is a discrete random variable with n outcomes $\{x_1, \ldots, x_n\}$.
- $p(x_i)$ is the probability of selecting x_i
- measure of information content (self-information or surprisal) of symbol x_i is $-\log p(x_i)$
- Entropy of selecting a symbol from X is:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

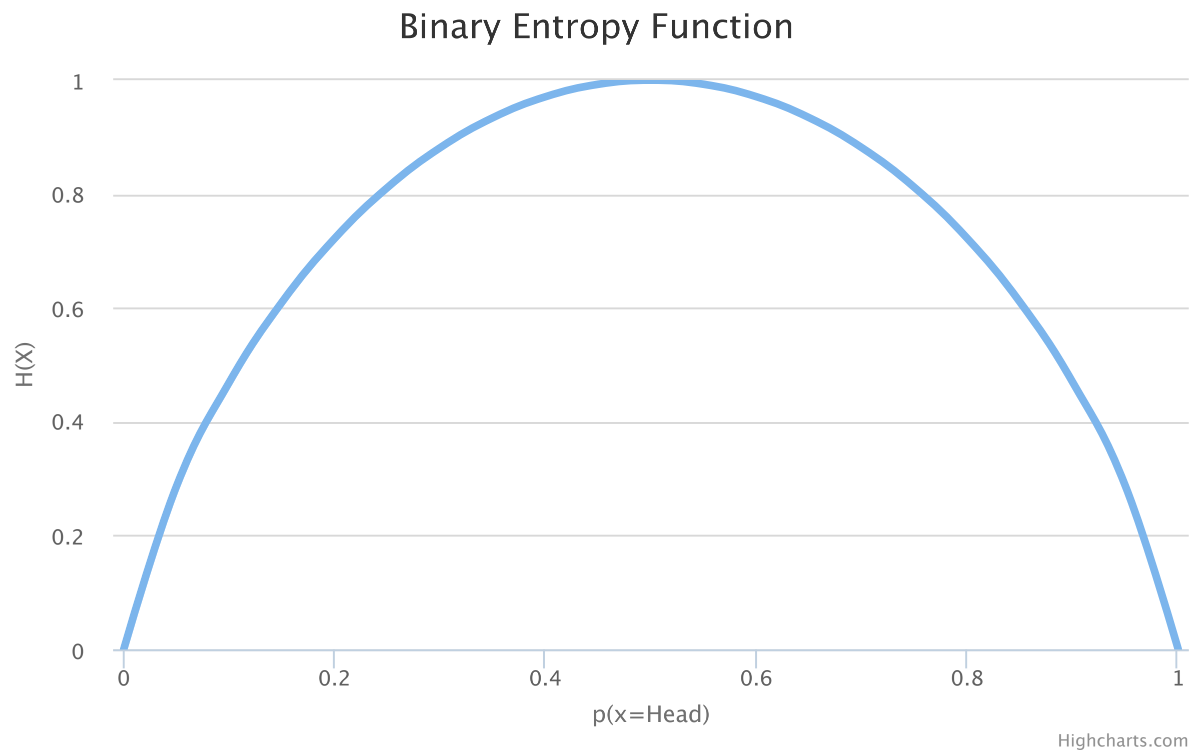
Entropy of a Binary Message

Consider a message, Head or Tail, constructed from the result of tossing a coin

- Case 1: Only Heads
  $H = -(1\log_2 1 + 0\log_2 0) = 0$ bit
- Case 2: Biased coin, 80% Heads and 20% Tails
  $H = -(0.8\log_2 0.8 + 0.2\log_2 0.2) = 0.72$ bit
- Case 3: Fair Coin
  $H = -(2 \times 0.5\log_2 0.5) = 1$ bit



Properties of Entropy

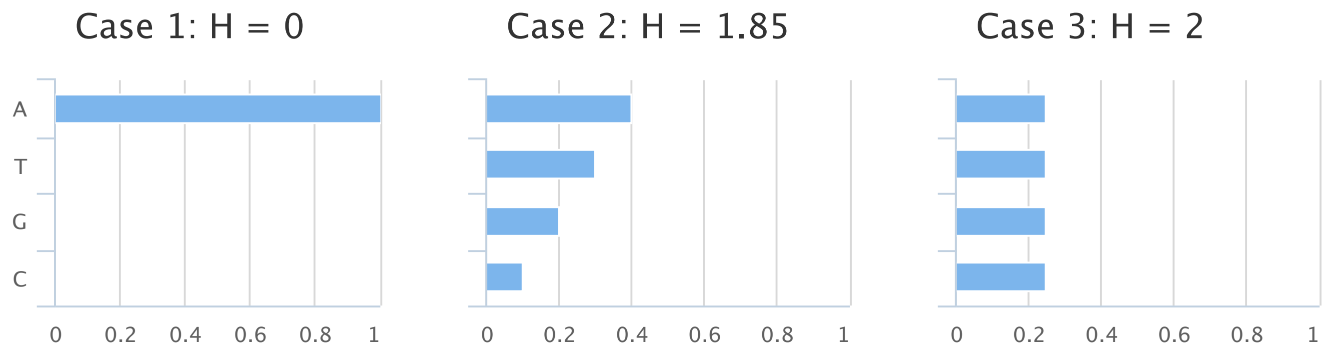


- H is maximum when all $p(x_i)$ are equal.
- H is minimum when any of $p(x_i)$ is 1.

Example with 4 symbols

Consider a message space of A, T, G, C characters with various probability distributions.

- Case 1: $H = -(1 \log_2 1 + 3 \times 0 \log_2 0) = 0$
- Case 2: $H = -(0.4 \log_2 0.4 + 0.3 \log_2 0.3 + 0.2 \log_2 0.2 + 0.1 \log_2 0.1) = 1.85$
- Case 3: $H = -(4 \times \frac{1}{4} \log_2 \frac{1}{4}) = 2$



H indicates the theoretical bounds for the average bits needed to represent/compress the symbols. Consider a message constructed from `A`, `T`, `G`, `C` with distibution of $[0.4, 0.3, 0.2, 0.1]$

- Binary encode `A` as `00`, `T` as `01`, `G` as `10`, and `C` as `11` (fixed-length)
- Average message length = $(0.4 \times 2 + 0.3 \times 2 + 0.2 \times 2 + 0.1 \times 2) = 2$

Alternatively:

- Binary encode `A` as `0`, `T` as `10`, `G` as `110`, and `C` as `111` (variable-length)
- Average message length = $(0.4 \times 1 + 0.3 \times 2 + 0.2 \times 3 + 0.1 \times 3) = 1.9 > H = 1.85$
- Using variable-length encoding (in this case **Huffman coding**), the size is 5% smaller