Information Theory and Coding

Quantitative measure of information

Cédric RICHARD
Université Côte d'Azur

Self-information

Information content

Let A be an event with non-zero probability $P(A)$.

The greater the uncertainty of A, the larger the information $h(A)$ provided by the realization of A. This can be expressed as follows:

$$h(A) = f\left(\frac{1}{P(A)}\right).$$

Function $f(\cdot)$ must satisfy the following properties:

- ▷ $f(\cdot)$ is an increasing function over $\mathbb{R}_+$
- ▷ information provided by 1 sure event is zero: $\lim_{p\to 1} f(p) = 0$
- ▷ information provided by 2 independent events: $f(p_1 \cdot p_2) = f(p_1) + f(p_2)$

This leads us to use the logarithmic function for $f(\cdot)$

Self-information

Information content

Lemme 1. *Function* $f(p) = -\log_b p$ *is the only one that is both positive, continue over* $(0, 1]$*, and that satisfies* $f(p1 \cdot p2) = f(p1) + f(p2)$*.*

Proof. The proof consists of the following steps:

1. $f(p^n) = n\, f(p)$
2. $f\left(p^{1/n}\right) = \frac{1}{n}\, f(p)$ after replacing p with $p^{1/n}$
3. $f\left(p^{m/n}\right) = \frac{m}{n}\, f(p)$ by combining the two previous equalities
4. $f(p^q) = q\, f(p)$ where q is any positive rational number
5. $f(p^r) = \lim_{n\to+\infty} f(p^{q_n}) = \lim_{n\to+\infty} q_n\, f(p) = r\, f(p)$ because rationals are dense in the reals

Let p and q in $(0, 1[$. One can write: $p = q^{\log_q p}$, which yields:

$$f(p) = f\left(q^{\log_q p}\right) = f(q)\, \log_q p.$$

We finally arrive at: $f(p) = -\log_b p$

Self-information

Information content

Definition 1. *Let $(\Omega, \mathcal{A}, P)$ be a probability space, and A an event of $\mathcal{A}$ with non-zero probability $P(A)$. The information content of A is defined as:*

$$h(A) = -\log P(A).$$

Unit. The unit of $h(A)$ depends on the base chosen for the logarithm.

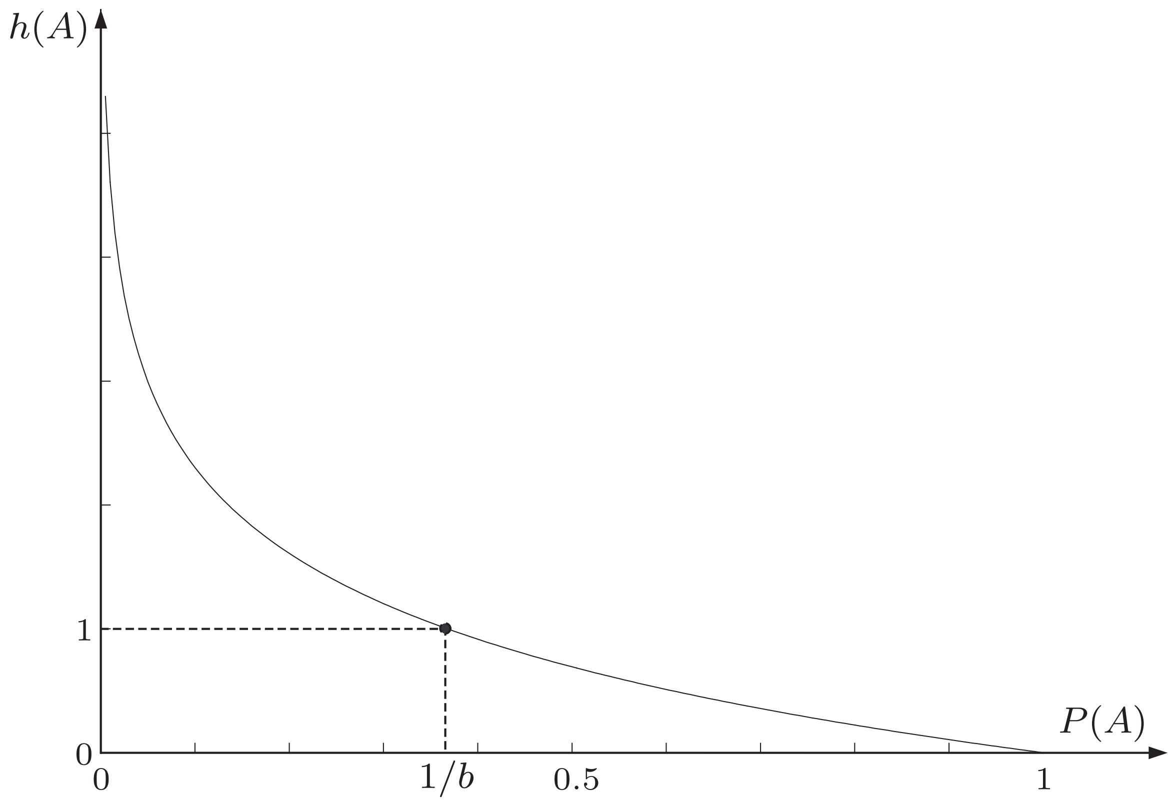
- $\log_2$: Shannon, bit (binary unit)
- $\log_e$: logon, nat (natural unit)
- $\log_{10}$: Hartley, decit (decimal unit)

Vocabulary. $h(\cdot)$ represents the *uncertainty* of A, or its *information content*.

Self-information

Information content

Information content or uncertainty: $h(A) = -\log_b P(A)$



Self-information

Information content

Example 1. Consider a binary source $S \in \{0, 1\}$ with $P(0) = P(1) = 0.5$. Information content conveyed by each binary symbol is equal to: $h\left(\frac{1}{2}\right) = \log 2$, namely, 1 bit or Shannon.

Example 2. Consider a source S that randomly selects symbols s_i among 16 equally likely symbols $\{s_0, \ldots, s_{15}\}$. Information content conveyed by each symbol is $\log 16$ Shannon, that is, 4 Shannon.

Remark. The bit in Computer Science (*binary digit*) and the bit in Information Theory (*binary unit*) do not refer to the same concept.

Self-information

Conditional information content

Self-information applies to 2 events A and B. Note that $P(A, B) = P(A)\,P(B|A)$. We get:

$$h(A, B) = -\log P(A, B) = -\log P(A) - \log P(B|A)$$

Note that $-\log P(B|A)$ is the information content of B that is not provided by A.

Definition 2. *Conditional information content of B given A is defined as:*

$$h(B|A) = -\log P(B|A),$$

that is: $h(B|A) = h(A, B) - h(A)$.

Exercise. Analyze and interpret the following cases: $A \subset B$, $A = B$, $A \cap B = \emptyset$.

Self-information

Mutual information content

The definition of conditional information leads directly to another definition, that of mutual information, which measures information shared by two events.

Definition 3. *We call mutual information of A and B the following quantity:*

$$i(A, B) = h(A) - h(A|B) = h(B) - h(B|A).$$

Exercise. Analyze and interpret the following cases: $A \subset B$, $A = B$, $A \cap B = \emptyset$.

Entropy of a random variable

Definition

Consider a memoryless stochastic source S with alphabet $\{s_1, \ldots, s_n\}$. Let p_i be the probability $P(S = s_i)$.

The entropy of S is the average amount of information produced by S:

$$H(S) = E\{h(S)\} = -\sum_{i=1}^{n} p_i \log p_i.$$

Definition 4. *Let X be a random variable that takes its values in $\{x_1, \ldots, x_n\}$. Entropy of X is defined as follows:*

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

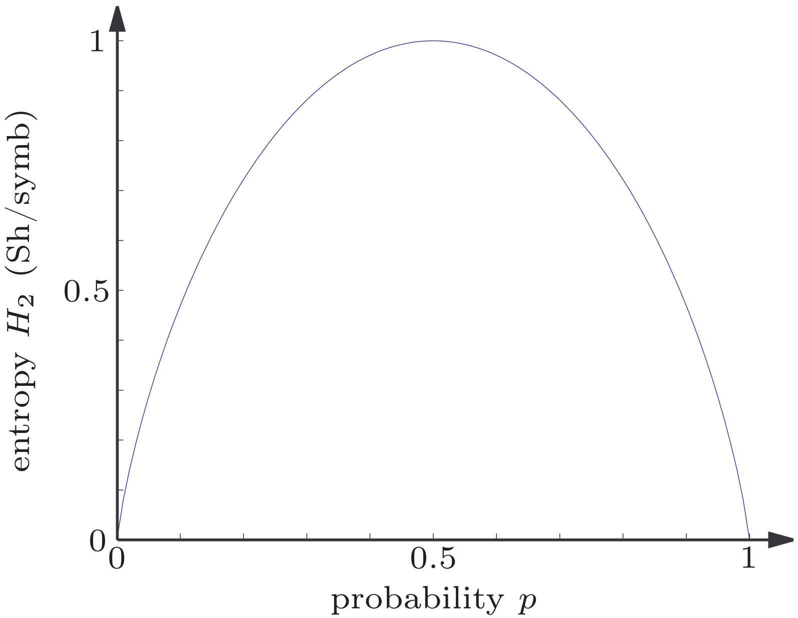
Entropy of a random variable

Example of a binary random variable

The entropy of a binary random variable is given by:

$$H(X) = -p\log p - (1-p)\log(1-p) \triangleq H_2(p).$$

$H_2(p)$ is called the binary entropy function.



Lemme 2 (Gibbs' inequality). *Consider 2 discrete probability distributions with mass functions $(p_1, \ldots, p_n)$ and $(q_1, \ldots, q_n)$. We have:*

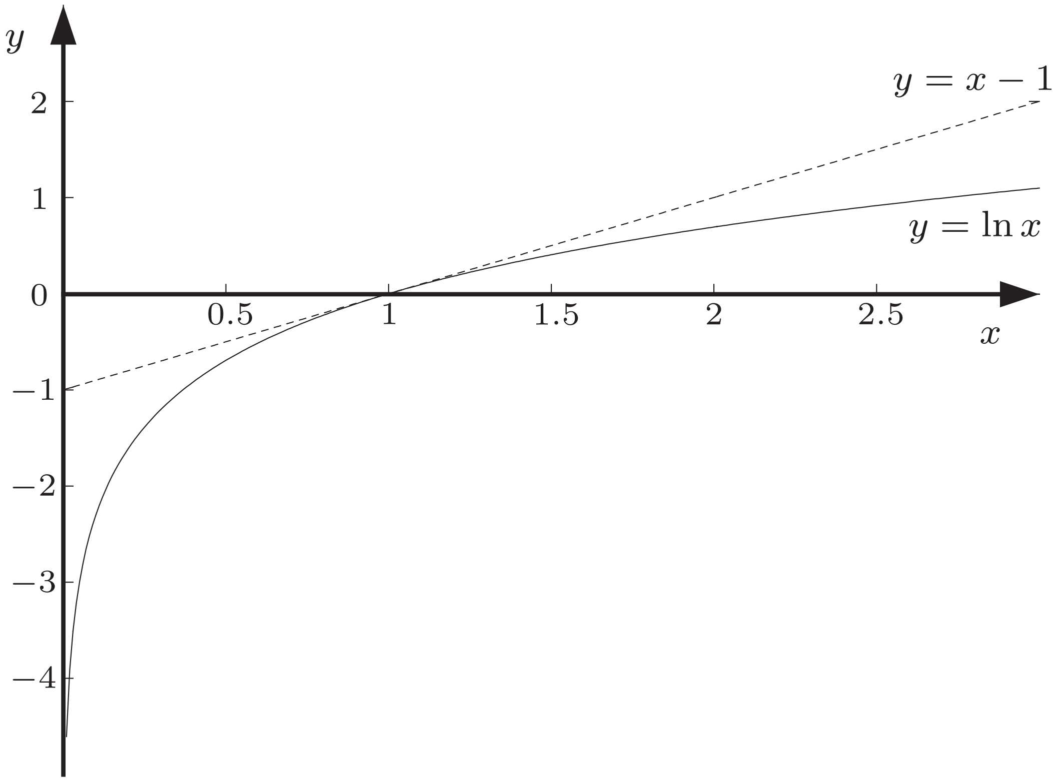
$$\sum_{i=1}^{n} p_i \log \frac{q_i}{p_i} \leq 0$$

Equality is achieved when $p_i = q_i$ for all i

Proof. The proof is carried out in the case of the neperian logarithm. Observe that $\ln x \leq x - 1$, with equality for $x = 1$. Let $x = \frac{q_i}{p_i}$. We have:

$$\sum_{i=1}^{n} p_i \ln \frac{q_i}{p_i} \leq \sum_{i=1}^{n} p_i \left(\frac{q_i}{p_i} - 1 \right) = 1 - 1 = 0.$$

Graphical checking of inequality $\ln x \leq x - 1$



Entropy of a random variable

Properties

Property 1. *The entropy satisfies the following inequality:*

$$H_n(p_1, \ldots, p_n) \leq \log n,$$

Equality is achieved by the uniform distribution, that is, $p_i = \frac{1}{n}$ *for all* i*.*

Proof. Based on Gibbs' inequality, we set $q_i = \frac{1}{n}$.

Uncertainty about the outcome of an experiment is maximum when all possible outcomes are equiprobable.

Entropy of a random variable

Properties

Property 2. *The entropy increases as the number of possible outcomes increases.*

Proof. Let X be a discrete random variable with values in $\{x_1, \ldots, x_n\}$ and probabilities $(p_1, \ldots, p_n)$, respectively. Consider that state x_k is split into two substates x_{k_1} et x_{k_2}, with non-zero probabilities p_{k_1} et p_{k_2} such that $p_k = p_{k_1} + p_{k_2}$.

Entropy of the resulting random variable X' is given by:

$$
\begin{aligned}
H(X') &= H(X) + p_k \log p_k - p_{k_1} \log p_{k_1} - p_{k_2} \log p_{k_2} \\
&= H(X) + p_{k_1}(\log p_k - \log p_{k_1}) + p_{k_2}(\log p_k - \log p_{k_2}).
\end{aligned}
$$

The logarithmic function being strictly increasing, we have: $\log p_k > \log p_{k_i}$. This implies: $H(X') > H(X)$.

Interpretation. Second law of thermodynamics

Property 3. *The entropy H_n is a concave function of $p_1, \ldots, p_n$.*

Proof. Consider 2 discrete probability distributions $(p_1, \ldots, p_n)$ and $(q_1, \ldots, q_n)$. We need to prove that, for every λ in $[0, 1]$, we have:

$H_n(\lambda p_1 + (1-\lambda)q_1, \ldots, \lambda p_n + (1-\lambda)q_n) \geq \lambda H_n(p_1, \ldots, p_n) + (1-\lambda)H_n(q_1, \ldots, q_n).$

By setting $f(x) = -x \log x$, we can write:

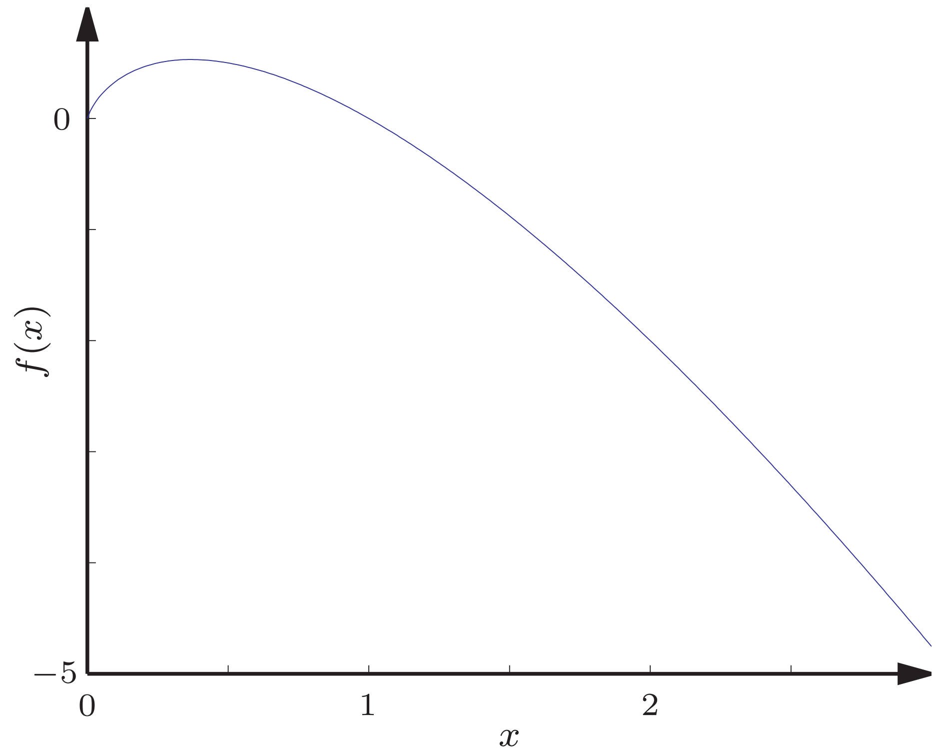
$$H_n(\lambda p_1 + (1-\lambda)q_1, \ldots, \lambda p_n + (1-\lambda)q_n) = \sum_{i=1}^{n} f(\lambda p_i + (1-\lambda)q_i).$$

The result is a direct consequence of the concavity of $f(\cdot)$ and Jensen's inequality.

Entropy of a random variable

Properties

Graphical checking of the concavity of $f(x) = -x \log x$



Entropy of a random variable

Properties

Concavity of H_n can be generalized to any number m of distributions.

Property 4. *Given* $\{(q_{1j},\ldots,q_{nj})\}_{j=1}^{m}$ *a finite set of discrete probability distributions, the following inequality is satisfied:*

$$H_n(\sum_{j=1}^{m}\lambda_j\, q_{1j},\ldots,\sum_{j=1}^{m}\lambda_j\, q_{mj}) \geq \sum_{j=1}^{m}\lambda_j\, H_n(q_{1j},\ldots,q_{mj}),$$

where $\{\lambda_j\}_{j=1}^{m}$ *is any set of constants in* $[0,1]$ *such that* $\sum_{j=1}^{m}\lambda_j = 1$.

Proof. As in the previous case, the demonstration of this inequality is based on the concavity of $f(x) = -x\log x$ and Jensen's inequality.

Pair of random variables

Joint entropy

Definition 5. *Let X and Y be two random variables with values in $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_m\}$, respectively. The joint entropy of X and Y is defined as:*

$$H(X, Y) \triangleq -\sum_{i=1}^{n} \sum_{j=1}^{m} P(X = x_i, Y = y_j) \log P(X = x_i, Y = y_j).$$

$\triangleright$ The joint entropy is symmetric: $H(X, Y) = H(Y, X)$

Example. Case of two independent random variables

Pair of random variables

Conditional entropy

Definition 6. *Let X and Y be two random variables with values in $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_m\}$, respectively. The conditional entropy of X given $Y = y_j$ is:*

$$H(X|Y = y_j) \triangleq -\sum_{i=1}^{n} P(X = x_i|Y = y_j) \log P(X = x_i|Y = y_j).$$

$H(X|Y = y_j)$ is the amount of information needed to describe the outcome of X given that we know that $Y = y_j$.

Definition 7. *The conditional entropy of X given Y is defined as:*

$$H(X|Y) \triangleq \sum_{j=1}^{m} P(Y = y_j)\, H(X|Y = y_j),$$

Example. Case of two independent random variables

Pair of random variables

Relations between entropies

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y).$$

These equalities can be obtained by first writing:

$$\log P(X = x, Y = y) = \log P(X = x|Y = y) + \log P(Y = y),$$

and then taking the expectation of each member.

Property 5 (chain rule)**.** *The joint entropy of n random variables can be evaluated using the following chain rule:*

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_1 \ldots X_{i-1}).$$

Pair of random variables

Relations between entropies

Each term of $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$ is positive. We can conclude that:

$$H(X) \leq H(X,Y)$$
$$H(Y) \leq H(X,Y)$$

Pair of random variables

Relations between entropies

From the *generalized concavity* of the entropy, setting $q_{ij} = P(X = x_i|Y = y_j)$ and $\lambda_j = P(Y = y_j)$, we get the following inequality:

$$H(X|Y) \leq H(X)$$

Conditioning a random variable reduces its entropy. Without proof, this can be generalized as follows:

Property 6 (entropy decrease with conditioning)**.** *The entropy of a random variable decreases with successive conditionings, namely,*

$$H(X_1|X_2, \ldots, X_n) \leq \ldots \leq H(X_1|X_2, X_3) \leq H(X_1|X_2) \leq H(X_1),$$

where $X_1, \ldots, X_n$ denote n discrete random variables.

Pair of random variables

Relations between entropies

Consider X and Y two random variables, respectively with values in $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_m\}$. We have:

$$0 \leq H(X|Y) \leq H(X) \leq H(X, Y) \leq H(X) + H(Y) \leq 2H(X, Y).$$

Pair of random variables

Mutual information

Definition 8. *The mutual information of two random variables X and Y is defined as follows:*

$$I(X,Y) \triangleq H(X) - H(X|Y)$$

or, equivalently,

$$I(X,Y) \triangleq \sum_{i=1}^{n} \sum_{j=1}^{m} P(X = x_i, Y = y_j) \log \frac{P(X = x_i, Y = y_j)}{P(X = x_i)\, P(Y = y_j)}.$$

The mutual information quantifies the amount of information obtained about one random variable through observing the other random variable.

Exercise. Case of two independent random variables

Pair of random variables

Mutual information

In order to give a different interpretation of mutual information, the following definition is recalled beforehand.

Definition 9. *We call the Kullback-Leibler distance between two distributions P_1 and P_2, here supposed to be discrete, the following quantity:*

$$d(P_1, P_2) = \sum_{x \in X(\Omega)} P_1(X = x) \log \frac{P_1(X = x)}{P_2(X = x)}.$$

The mutual information corresponds to the Kullback-Leibler distance between the marginal distributions and the joint distribution of X and Y.

Pair of random variables

Venn diagram

A Venn diagram can be used to illustrate relationships among measures of information: entropy, joint entropy, conditional entropy and mutual information.