Information Theory and Coding

Shannon's communication model

Cédric RICHARD
Université Côte d'Azur

Information Theory

Models of communication

Models of communication are conceptual models used to explain the human communication process.

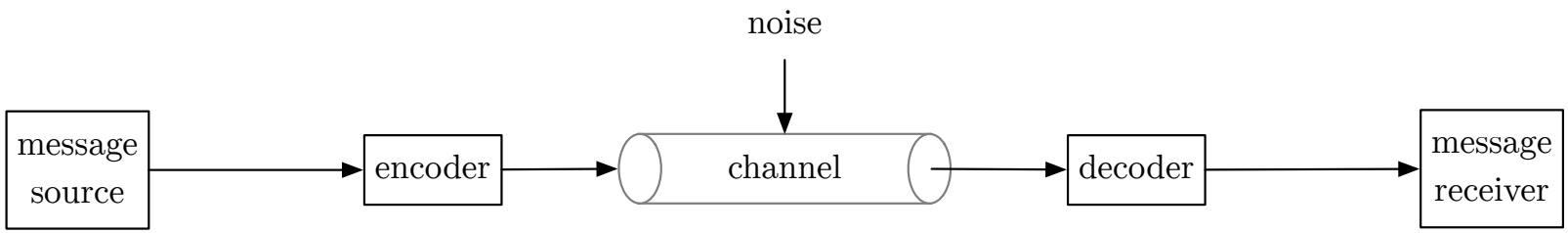
Following the basic concept, communication is the process of sending and receiving messages or transferring information from one part (sender) to another (receiver).

The Shannon-Weaver model was designed in 1949 to mirror the functioning of radio and telephone technology. It is referred to as the mother of all models.

This model has been expanded later by other scholars: Berlo (1960), ...

Information Theory

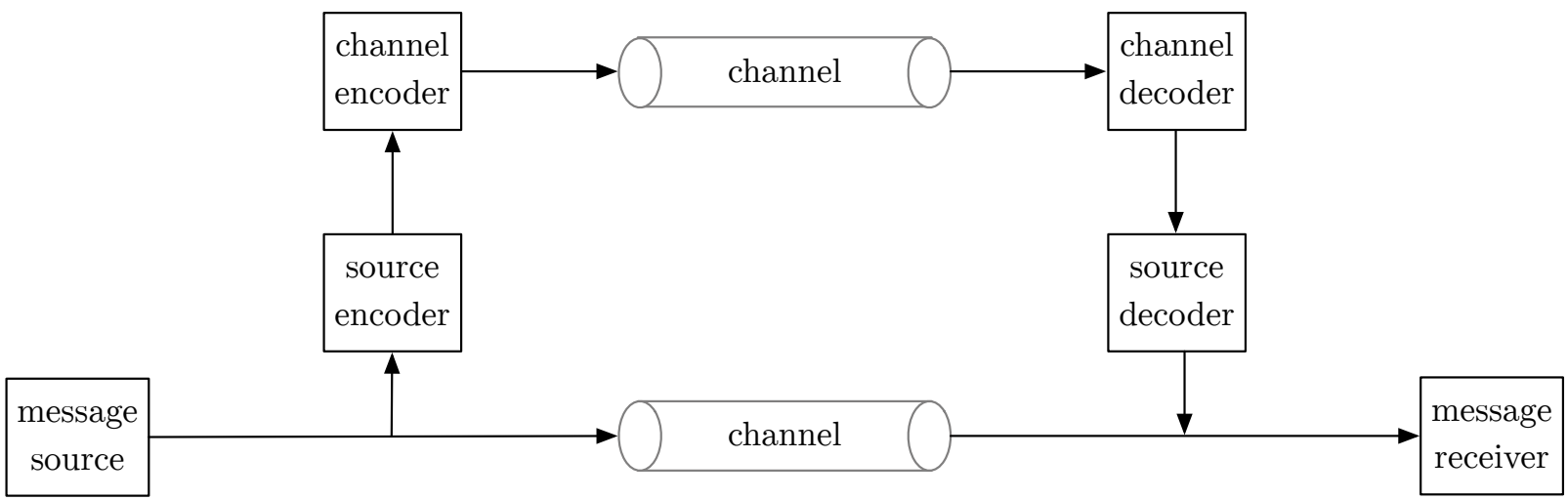
Shannon's communication model



An information source, which produces a message
An encoder, which encodes the message into signals
A channel, for which signals are adapted for transmission
A decoder, which reconstructs the encoded message
An information destination, where the message arrives

Information Theory

Shannon's communication model



Information Theory

Objectives

Information theory studies the quantification, storage, and communication of information.

It was originally proposed by Claude Shannon in 1948 to find fundamental limits on signal processing and communication operations such as data compression.

Applications of fundamental topics of information theory include lossless data compression, lossy data compression, and channel coding.

Information theory is used in information retrieval, intelligence gathering, gambling, statistics, and even in musical composition.

A key measure is entropy. It quantifies the amount of uncertainty involved in the value of a random variable or the outcome of a random process.

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 $f(\cdot)$

A and B two events which are independent.

By definition : $P(A,B) = P(A)P(B)$

$P(A,B)$ → $P(A \cap B)$ ("and")

$$h(A,B) \triangleq f\left(\frac{1}{P(A,B)}\right) \overset{\text{indep.}}{=} f\left(\frac{1}{P(A)} \cdot \frac{1}{P(B)}\right)$$

we want → $= f\left(\frac{1}{P(A)}\right) + f\left(\frac{1}{P(B)}\right)$

$$= h(A) + h(B)$$

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. $f\left(p^n\right) = n\, f(p)$
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. $f\left(p^r\right) = \lim_{n\to+\infty} f\left(p^{q_n}\right) = \lim_{n\to+\infty} q_n\, f(p) = r\, f(p)$ 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: $f(p) = -\log_b p$

1. $f(p^n) = f(p^{n-1} \cdot p) = f(p^{n-1}) + f(p)$

$= \dots$

$= \underbrace{f(p) + \dots + f(p)}_{n \text{ times}}$

$= n\, f(p)$

$$\boxed{f(p^n) = n\, f(p)}$$

2. $f(p) = f\left((p^{1/n})^n\right) \overset{1.}{=} n\, f(p^{1/n})$

$\Rightarrow f(p^{1/n}) = \frac{1}{n} f(p)$

$$\boxed{f(p^{1/n}) = \frac{1}{n} f(p)}$$

3. $f(p^{m/n}) = f\left((p^{1/n})^m\right) \overset{1.}{=} m\, f(p^{1/n})$

$\overset{2.}{=} \frac{m}{n} f(p)$

$$\boxed{f(p^{m/n}) = \frac{m}{n} f(p)}$$

4. Because of 3.

$$f(p^q) = q\, f(p) \quad \text{for all } q \in \underbrace{\mathbb{Q}^+}_{\text{rationals}}.$$

5. Because $\mathbb{Q}$ is dense in $\mathbb{R}$, which means that

$\forall x \in \mathbb{R}$, $\exists (q_1, \dots, q_n)$, such that

$q_i \in \mathbb{Q}$

$x \dots$ $\qquad x = \lim_{n \to +\infty} q_n$

$$f(p^x) = x\, f(p) \quad , \quad \forall x \in \mathbb{R}$$

by continuity of f

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

$$q^{\log_q p} = e^{\ln q \times \log_q p} = e^{\ln q \times \frac{\ln p}{\ln q}}$$

$$= p$$

$$\log_b p = \frac{\ln p}{\ln b}$$

$$f(p) = f\left(q^{\log_q p}\right) \overset{5.}{=} \log_q p \; f(q)$$

$$\Rightarrow \frac{f(p)}{f(q)} = \log_q p = \frac{\ln p}{\ln q}$$

$$\Rightarrow \boxed{f(p) = C \ln p}$$

where $C \in \mathbb{R}$

I want $f(p) > 0$, $\forall p \in]0,1]$ ← probability.

$\Rightarrow$ C must be negative

Let say that $C = \frac{-1}{\ln b}$ with $b > 1$

$$\Rightarrow \boxed{f(p) = -\log_b p}$$

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)

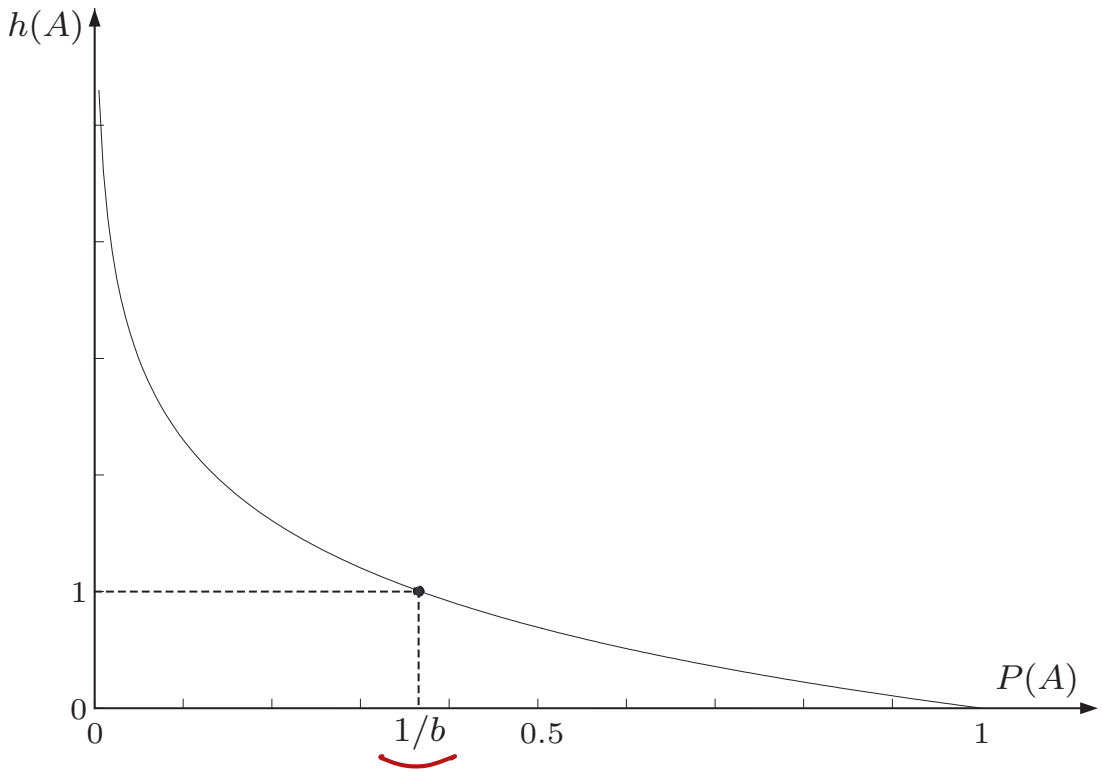
Vocabulary. $h(\cdot)$ represents the *uncertainty* of A, or its *information content*.

$P(A) \searrow$, uncertainty $\nearrow$ and $R(A) \nearrow$

Self-information

Information content

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



→ if $b = 2$, $h(1/2) = 1$ Shannon

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.

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.

$$h(S=0) = -\log_2 P(S=0)$$
$$= -\log_2\left(\frac{1}{2}\right)$$
$$= \log_2 2$$
$$= 1 \quad \text{Sh}$$

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.

$$h(S=s_i) = -\log_2 P(S=s_i)$$
$$= -\log_2\left(\frac{1}{16}\right)$$
$$= \log_2(2^4)$$
$$= 4 \quad \text{Sh}$$

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:

$$h(A,B) = -\log P(A,B) = \underbrace{-\log P(A)}_{h(A)} \underbrace{-\log P(B|A)}_{h(B|A)}$$

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: $A \subset B$, $A = B$, $A \cap B = \emptyset$.

$$h(A,B) = -\log_2 P(A,B)$$

$$= -\log_2 P(A) - \log_2 P(B|A)$$

$$= h(A) + h(B|A)$$

$h(B|A) = ?$ when A and B independent.

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:*

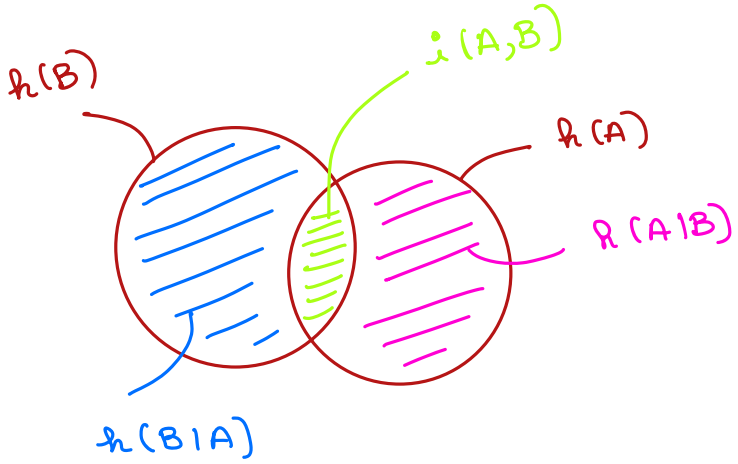
$$i(A, B) = h(A) - h(A|B) = h(B) - h(B|A).$$

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

$$h(A,B) = h(A) + h(B|A)$$
$$= h(B) + h(A|B)$$

$$i(A,B) \triangleq h(A) - h(A|B)$$
$$\triangleq h(B) - h(B|A)$$

mutual information shared by A and B



Venn diagram

$$h(A,B) = h(B|A) + h(A)$$
$$= h(A|B) + h(B)$$
$$= h(B|A) + h(A|B) + i(A,B)$$

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:

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

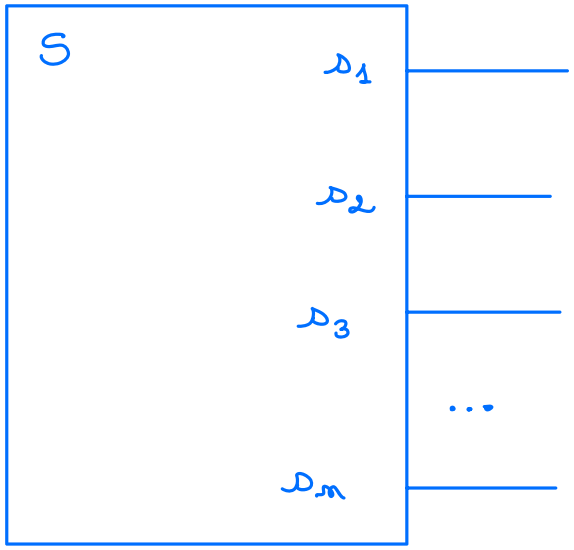
Sh / state

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:*

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

$S \in \{s_1, \dots, s_m\}$

$P(S = s_i) = p_i \quad , \; i \in \{1, \dots, m\}$



$$h(S = s_1) = -\log_2 P(S = s_1) = -\log_2 p_1$$

$$h(S = s_m) = -\log_2 p_m$$

$$H(S) = \text{Mean}\{h(S)\}$$

$$= \sum_{i=1}^{m} p_i \, h(S = s_i)$$

$$= \sum_{i=1}^{m} p_i \left(-\log_2 p_i\right)$$

$$\boxed{H(S) = -\sum_{i=1}^{m} p_i \log_2 p_i}$$

unit: Sh / state of S

Binary Source

Case of a binary source : $S = \{0, 1\}$

$$P(S=0) = p \quad \Rightarrow \quad P(S=1) = 1 - P(S=0) = 1 - p$$

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

$$= f(p) \quad \text{with} \quad p \in [0, 1]$$

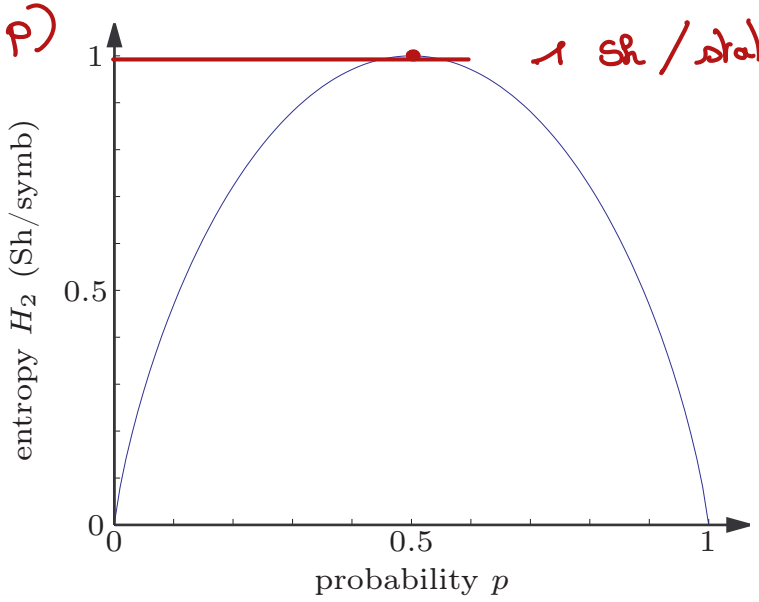
Entropy of a random variable

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.



$$\lim_{p \to 1} \left(-p \underbrace{\log_2 p}_{0} - (1-p) \log_2 (1-p) \right) = \lim_{p \to 0} p \log_2 p = 0$$

Entropy of a random variable

Notation and preliminary properties

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:*

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

pmf (p1,...,pn) and (q1,...,qn)

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:

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

We have : $\ln x \leq x - 1$, $\forall x \in]0, +\infty[$

We write $x = \frac{q_i}{p_i} > 0$

$$\sum_{i=1}^{m} p_i \ln\left(\frac{q_i}{p_i}\right) \leq \underbrace{\sum_{i=1}^{m} p_i \left(\frac{q_i}{p_i}\right)}_{= \sum_{i=1}^{m} q_i = 1 \text{ because } \{q_1, \dots, q_m\} \text{ distrib.}} - \underbrace{\sum_{i=1}^{m} 1 \times p_i}_{= 1 \text{ because } \{p_1, \dots, p_m\} \text{ distrib.}}$$

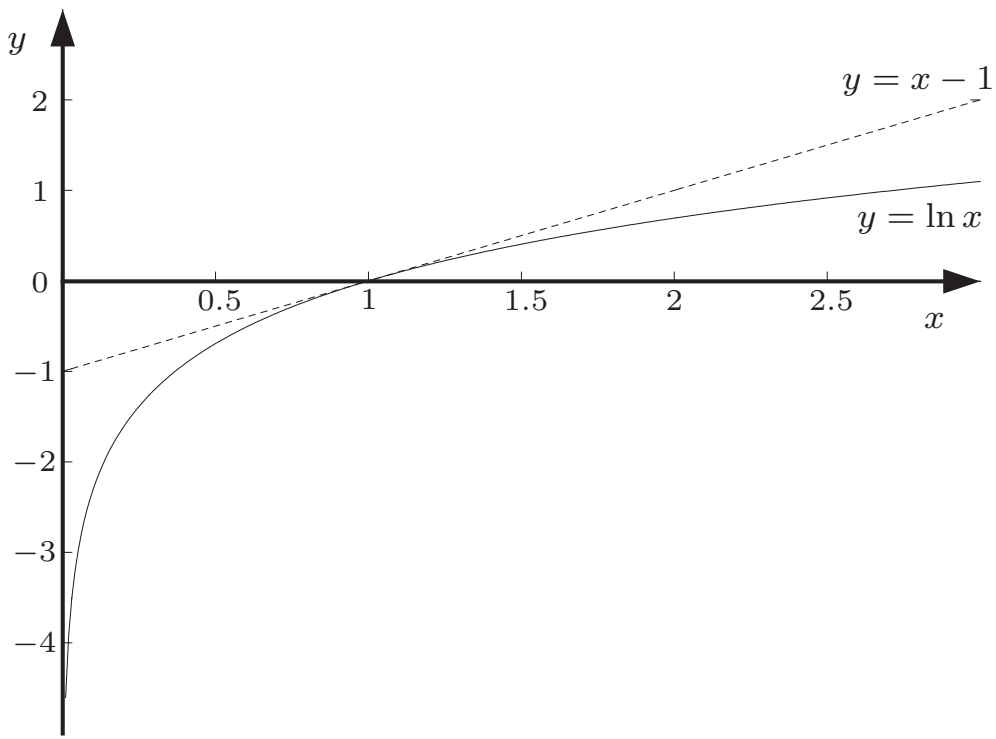
$$1 - 1 = 0$$

$$\Rightarrow \sum_{i=1}^{m} p_i \ln\left(\frac{q_i}{p_i}\right) \leq 0 \quad \text{(Gibbs)}$$

Entropy of a random variable

Notation and preliminary properties

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.

We use : $\sum_{i=1}^{n} p_i \log_2\left(\frac{q_i}{p_i}\right) \leq 0$

we set : $q_i = \frac{1}{n}$, $\forall i \in \{1, \dots, n\}$

$$\sum_{i=1}^{n} p_i \log_2\left(\frac{1}{n p_i}\right) \leq 0 \Leftrightarrow -\sum_{i=1}^{n} p_i \log p_i \leq \sum_{i=1}^{n} p_i \log_2 n$$

$$\Leftrightarrow H(S) \leq \log_2 n \times \underbrace{\sum_{i=1}^{n} p_i}_{1}$$

$$\boxed{H(S) \leq \log_2 n}$$

Equality : Gibbs inequality becomes an equality for $x = 1$

during the proof, we set $x = \frac{q_i}{p_i}$

and next $q_i = \frac{1}{n}$

Combining everything, $H(S) = \log_2 n$

if $p_i = \frac{1}{n}$, for all $i \in \{1, \dots, n\}$

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

Entropy of a random variable

Properties

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.

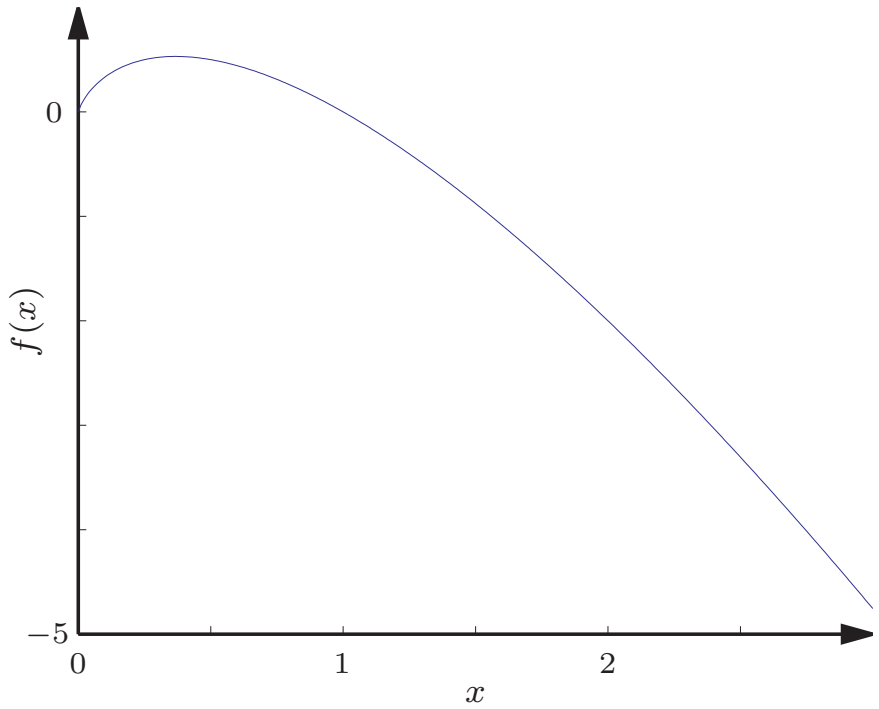
Rappel : $H(S) = \sum_{i=1}^{M} (-p_i \log p_i)$

$= \sum_{i=1}^{M} f(p_i)$

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:*

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

unit: Sh / pair of (x,y)

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

Example. Case of two independent random variables

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

Z : source

$$H(Z) = -\sum_{k=1}^{K} P(Z = z_k) \log_2 P(Z = z_k)$$

X and Y : 2 sources

$$Z = (X, Y)$$

$$H(X,Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} P(X = x_i, Y = y_j) \log_2 P(X = x_i, Y = y_j)$$

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

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

if X and Y indep., then $P(X=x_i|Y=y_j) = P(X=x_i)$

$\Rightarrow H(X|Y=y_j) = H(X)$

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

$$= \sum_j P(Y=y_j)\, H(X) \quad \text{because } X, Y \text{ indep.}$$

$$= H(X) \underbrace{\sum_j P(Y=y_j)}_{1}$$

$$= H(X)$$

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:*

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

$$H(X,Y) = -\sum_i \sum_j P(X=x_i, Y=y_j) \log P(X=x_i, Y=y_j)$$

$$= -\sum_i \sum_j P(X=x_i \mid Y=y_j) P(Y=y_j) \left[\log P(X=x_i \mid Y=y_j) + \log P(Y=y_j)\right]$$

$$= \sum_j \left[-\sum_i P(X=x_i \mid Y=y_j) \log P(X=x_i \mid Y=y_j)\right] P(Y=y_j)$$

$$-\sum_j \underbrace{\left[\sum_i P(X=x_i \mid Y=y_j)\right]}_{1} P(Y=y_j) \log P(Y=y_j)$$

$$= \sum_j H(X \mid Y=y_j)\, P(Y=y_j) - \sum_j P(Y=y_j) \log P(Y=y_j)$$

$$= H(X \mid Y) + H(Y)$$

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

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

Illustration :

$$H(X,Y,Z) = H(X) + H(Y,Z \mid X)$$
$$= H(X) + H(Y \mid X) + H(Z \mid X,Y)$$

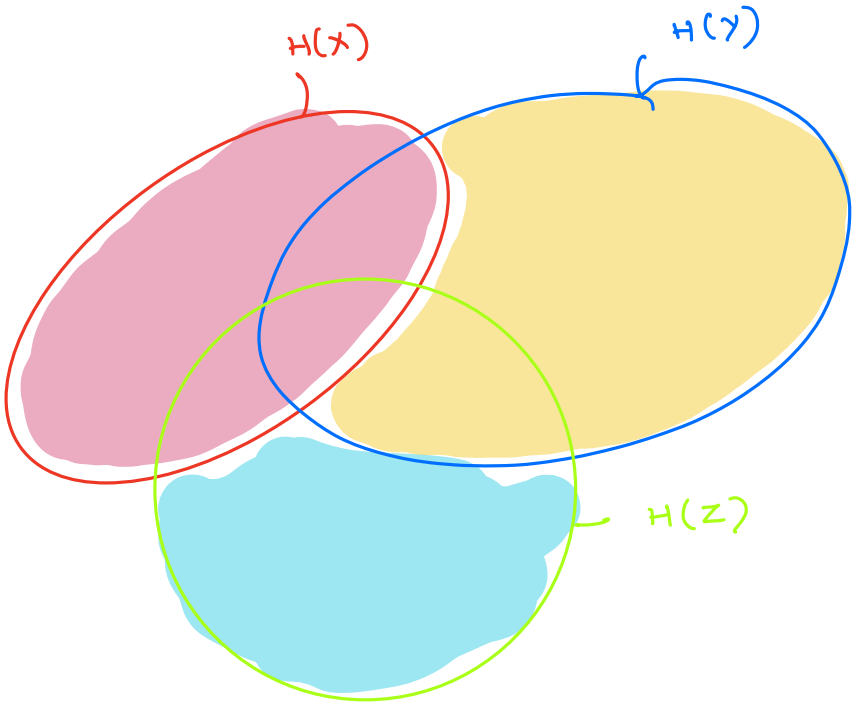
$$P(X,Y,Z) = P(X)\ P(Y,Z \mid X)$$

and : $$P(Y,Z \mid X) = P(Z \mid X,Y)\, P(Y \mid X)$$

$$\Rightarrow P(X,Y,Z) = P(X)\, P(Y \mid X)\, P(Z \mid X,Y)$$

log Σ

$$H(X,Y,Z) = H(X) + H(Y \mid X) + H(Z \mid X,Y)$$

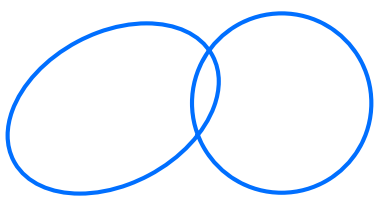

H(X)
H(Y)
H(Z)

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:

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



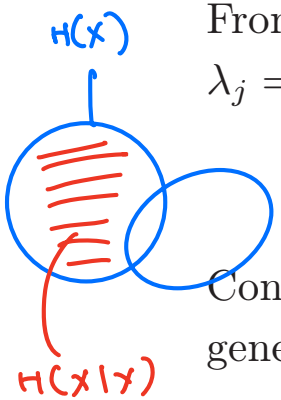
$H(X,Y) = H(X) + H(Y|X)$

We know that these three terms are positive

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

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 $\{x_1, \dots, x_n\}$ and $\{y_1, \dots, y_m\}$. We have:

$$0 \overset{①}{\leq} H(X|Y) \overset{②}{\leq} H(X) \overset{③}{\leq} H(X,Y) \overset{④}{\leq} H(X) + H(Y) \overset{⑤}{\leq} 2H(X,Y).$$

① : Entropy is positive because $\underbrace{-\sum_i p_i \log p_i}_{\geq 0}$ with $p_i \in [0,1]$

② $H(X|Y) \leq H(X)$ because conditionning decreases entropy (concavity)

③ $H(X,Y) = H(X) + H(Y|X)$ with $H(Y|X) \geq 0$

④ $H(X,Y) = H(X) + H(Y|X)$ and $H(Y|X) \leq H(Y)$

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

⑤ By ③ : $H(X) \leq H(X,Y)$

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

$\Rightarrow H(X) + H(Y) \leq 2\, H(X,Y)$

Inequalities can become equalities:

$\underline{x, y \text{ indep.}}$

or $X = f(Y)$ with f a one-to-one application

depending on the inequality.

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,

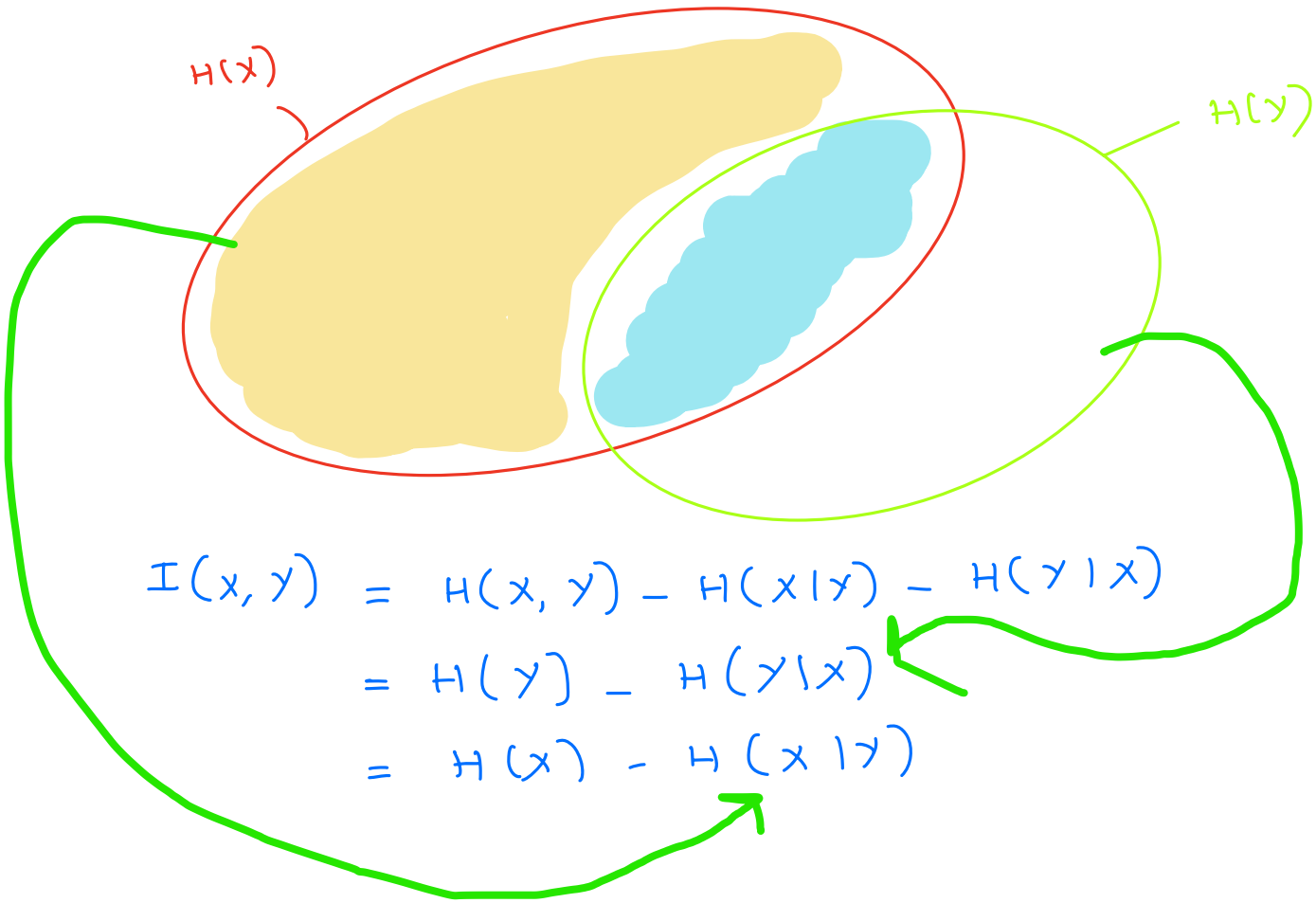
$$I(X,Y) \triangleq \sum_{i=1}^{n} \sum_{j=1}^{m} P(X = x_i, Y = y_j) \log \frac{P(X = x_i, Y = y_j)}{P(X = x_i)\, P(Y = y_j)}.$$

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

Mutual information

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



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:*

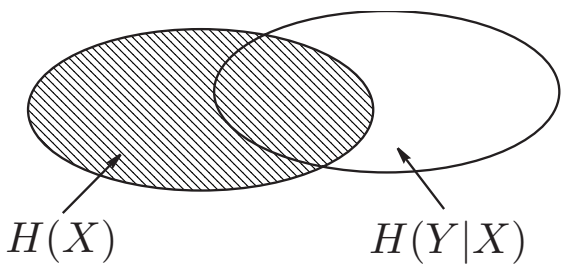
$$d(P_1, P_2) = \sum_{x \in X(\Omega)} P_1(X = x) \log \frac{P_1(X = x)}{P_2(X = x)}.$$

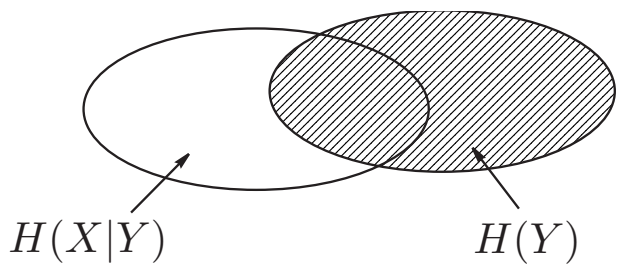
The mutual information corresponds to the Kullback-Leibler distance between the marginal distributions and the joint distribution of X and Y.

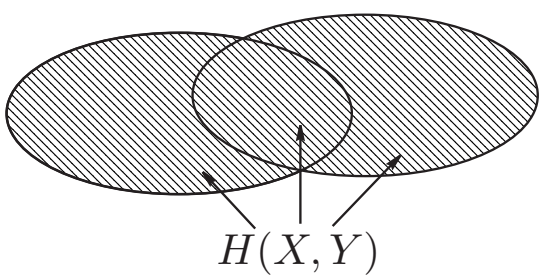
Pair of random variables

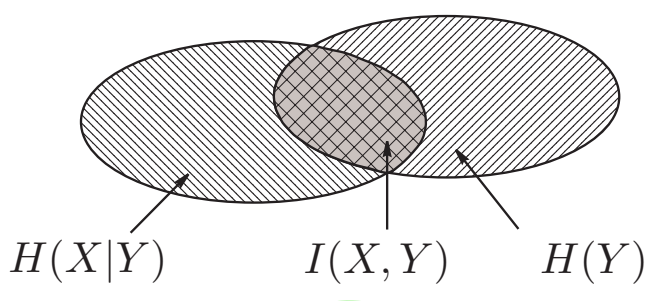
Venn diagram

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









Consider a single event A of probability $P(A)$

$$h(A) = -\log_2 P(A) \quad \text{Sh}$$

Considering an addition event B of probability $P(B)$

$$h(A,B) = -\log_2 P(A,B) \quad \text{(joint info)}$$

$$h(A|B) = -\log_2 P(A|B) \quad \text{(cond.t info)}$$

$$i(A,B) = h(A) - h(A|B) \quad \text{(shared)}$$

$S = \{s_1, \dots, s_m\}$

Richness of the source S: entropy $H(S)$

$$H(S) = \sum_{i=1}^{m} P(S=s_i)\, h(S=s_i)$$

$$= -\sum_{i=1}^{m} P(S=s_i) \log_2 P(S=s_i) \quad \text{Sh / state of S}$$

$$\max_{P(S)} H(S) = \log_2 m \quad \text{Sh / state of S}$$

reached when $P(S=s_i) = \frac{1}{m} \quad \forall i = 1, \dots, m$

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(X=x_i, Y=y_j) \log_2 P(X=x_i, Y=y_j)$$

Sh / pair of states of x and y

$$H(X|Y=y_{j_0}) = -\sum_{i=1}^{n} P(X=x_i|Y=y_{j_0}) \log_2 P(X=x_i|Y=y_{j_0})$$

Sh / state of x

$$H(X|Y) = \sum_{j=1}^{n} P(Y=y_j)\, H(X|Y=y_j) \quad \text{Sh / state of X}$$

$$I(X,Y) = H(X) - H(X|Y)$$
$$= H(Y) - H(Y|X)$$
$$= H(X,Y) - H(X|Y) - H(Y|X)$$

Properties: $H(X,Y) = H(X|Y) + H(Y)$
$= H(Y|X) + H(X)$

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

if X and Y indep., then:

$H(X|Y) = H(X)$

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

UNIVERSITÉ CÔTE D'AZUR

Cédric Richard

Quantitative Measure of Information
Part I

Exercise 1

One person says: "Today is my birthday". Calculate the amount of self-information conveyed by this statement. Calculate the average amount of information conveyed by this source over one year.

Exercise 2

The 64 squares of a chessboard are assumed to be equiprobable. Determine the average amount of information contained in a communication indicating the position of a given chess piece. Propose a dichotomous strategy, based on questions of the form "Is the chess piece on that part of the chessboard?", that would allow to guess the position of this chess piece in a minimum average number of questions. Compare this average number of questions to the entropy calculated at the beginning of the exercise.

Exercise 3

A perfectly balanced coin is tossed until the first head appears. Calculate the entropy $H(X)$ in Shannon, where the random variable X denotes the number of flips required to get the first head. Propose a dichotomous strategy, based on questions with binary response of the form "Is X is smaller or greater than (...)", making it possible to guess the value of X in a minimum average number of questions. Compare this number of questions to $H(X)$.
In order to resolve this exercise, the following equality can be used $\sum_{n=1}^{\infty} n\, a^n = \frac{a}{(1-a)^2}$.

Exercise 5

Consider a tank that consists of two compartments of identical volumes. Compartment I is filled with two inert gases with respective proportions $(\frac{2}{5}, \frac{3}{5})$. The same gases fill compartment II with respective proportions $(\frac{1}{3}, \frac{2}{3})$. Assuming the pressure and temperature in both compartments are the same, calculate the tank entropy before and after the two compartments communicate. Interpret the result.

Exercise 6

A source emits symbols 0 and 1 with probabilities $P(0) = \frac{1}{4}$ and $P(1) = \frac{3}{4}$. These symbols are transmitted to a receiver through an imperfect symmetric channel illustrated by Figure 1, with $p_0 = 10^{-1}$. Denoting by X and Y the transmitted and received symbols, calculate the following quantities: $H(X)$, $H(Y)$, $H(X,Y)$, $H(Y|X)$, $H(X|Y)$ and $I(X,Y)$.

Problem 1

Let $\{\mathcal{E}_k\}_{k=1}^n$ be a partition of $\mathcal{E}$. We denote by N and N_k the numbers of elements in sets $\mathcal{E}$ and $\mathcal{E}_k$, respectively. Assume that the elements of $\mathcal{E}$ are equiprobable. We set $p_k = N_k/N$.

1. Determine the self-information of any element of $\mathcal{E}_k$. Calculate the average amount of information needed to determine any element in $\mathcal{E}_k$.
2. Calculate the average amount of information needed to characterize any element of $\mathcal{E}$. By noticing that we can split the identification procedure of an element of $\mathcal{E}$ in 2 steps, (a) identification of the set $\mathcal{E}_k$, and then (b) identification of the element in $\mathcal{E}_k$, estimate the average amount of information needed to identify $\mathcal{E}_k$.

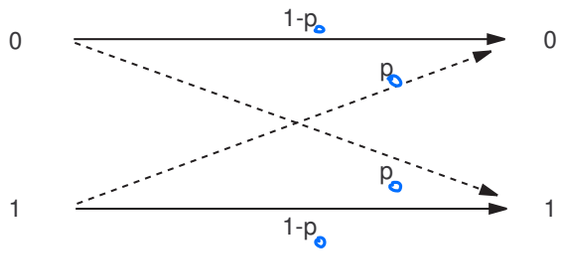


Figure 1: Imperfect channel.

Problem 2

Consider a twin-pan balance and 9 coins. We know that one of these coins is fake. The problem is to find the fake coin given that it only differs from the other 8 coins by its weight.

1. Determine the number of possible cases, considering that the fake coin may be heavier or lighter than the others. Calculate the average amount of information necessary to identify the fake coin.

2. To identify the fake coin, the weights of two sets of n coins each are compared using the twin-pan balance. Enumerate the possible outcomes of each weighting operation. Assuming these outcomes are equiprobable, determine in that case the amount of information provided by every weighing operation. Determine the average number of weighting operations to plan.

3. One wants to determine n in order to maximize the amount of information provided by each weighting operation. Let P_ℓ, resp. P_r, be the probability that the set of coins in the left pan, resp. right pan, is heavier. Let P_e be the probability that an equilibrium is achieved. Calculate P_ℓ, P_r and P_e.

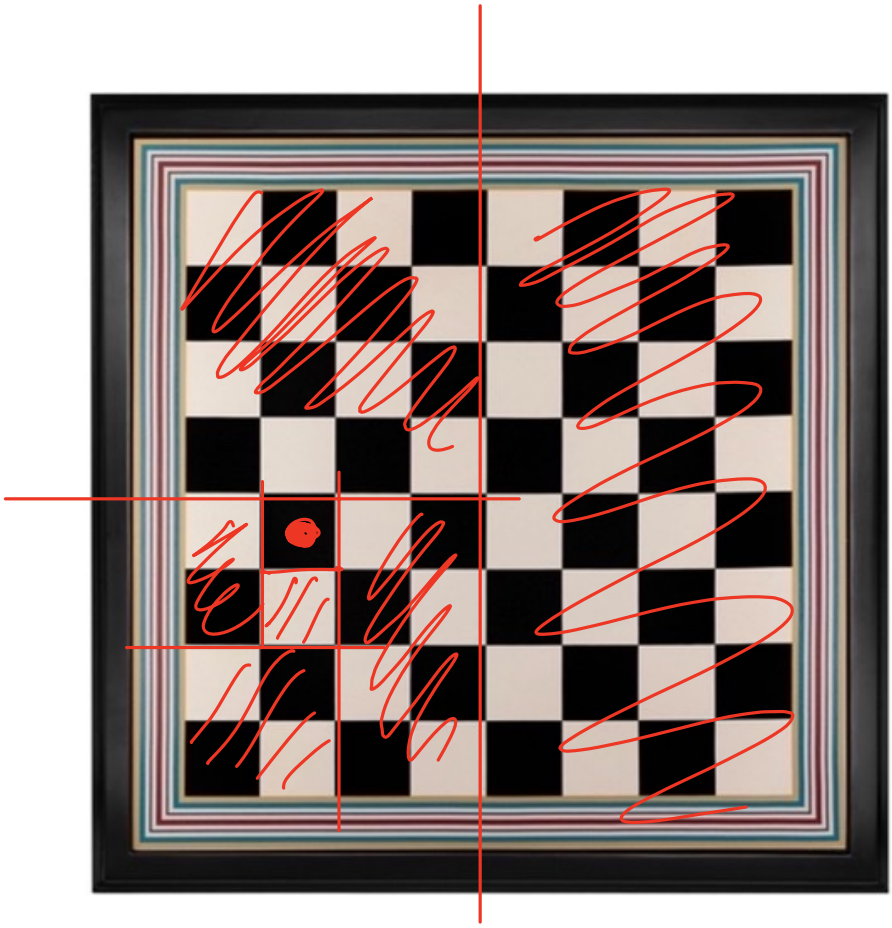
4. Calculate n to maximize the entropy of each weighting operation.

5. Calculate the minimum average number of weighting operations required to identify the fake coin.

6. Propose a strategy to identify the fake coin.

exercice 2 :

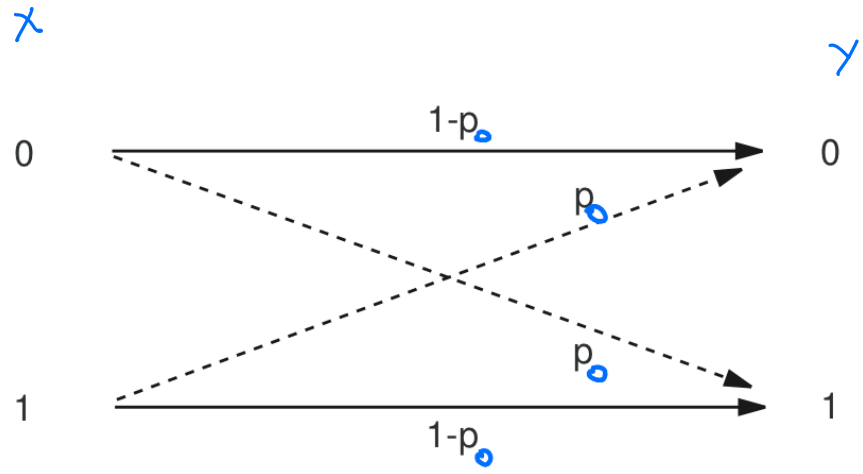
6 questions.



$H(S) = \log_2 64 = \log_2 2^6 = 6$ Sh /state

= number of questions.

Exercise 6:



$P(X=0) = \frac{1}{4}$ $\quad P(X=1) = \frac{3}{4}$ $\quad p_0 = 10^{-1}$

$p_0 = P(Y=0 \mid X=1) = P(Y=1 \mid X=0)$

$1-p_0 = P(Y=1 \mid X=1) = P(Y=0 \mid X=0)$

$H(X)$, $H(Y)$, $H(X,Y)$, $H(Y|X)$, $H(X|Y)$ and $I(X,Y)$.

UNIVERSITÉ CÔTE D'AZUR

Cédric Richard

Quantitative Measure of Information
Part II

Exercise

Let X be a discrete random variable that can take n possible values, and Y a discrete random variable uniformly distributed that can take n possible values. Throughout the exercise, no assumption will be made on the distribution of X.

1. Calcule the maximum entropy that can be reached by X, and specify the case in which this value would be obtained.

2. Calcule the entropy of Y.

3. In the case of $n = m$, rank in ascending order the following quantities: $H(Y)$, 0, $H(X)$, $H(X;Y)$, $H(X) + H(Y)$, $H(X|Y)$. Justify each step of your ranking.

4. Explain cases of equality in the ranking proposed in 3., i.e., give for each inequality the cases in which it becomes an equality.

5. We now consider the case where the joint distribution $P(X = x_i; Y = y_j)$ is given by the table below:

$P(X;Y)$	y_1	y_2	y_3	y_4
x_1	1/24	1/12	1/6	1/24
x_2	1/6	1/8	1/24	1/6
x_3	1/24	1/24	1/24	1/24

Check that Y is uniformaly distributed. Calculate $I(X;Y)$.

Problem

For a given region, the forecasts of a meteorologist are divided according to their relative frequencies given by the table below. The columns correspond to the actual weather, which is represented by the random variable T, which takes values 0 or 1 depending on whether the weather is rainy or sunny, respectively. The rows correspond to the meteorologist's forecast, identified by the random variable M, also with values in $\{0.1\}$ depending on whether he had planned a rainy weather (0) or a sunny weather (1).

$P(M = i, T = j)$	sunny weather $(T = 1)$	rainy weather $(T = 0)$
sunny weather $(M = 1)$	5/8	1/16
rainy weather $(M = 0)$	3/16	1/8

1. Calculate the probabilities $P(M = i)$ and $P(T = j)$, with $i, j \in \{0, 1\}$.

2. Show that the meteorologist is wrong once in 4 times.

3. One student says that by always forecasting sunny weather, he makes fewer mistakes than the meteorologist does. Check this assertion.

4. Let E be the random variable representing the student's prediction. As for T and M, random variable E takes values in $\{0.1\}$. Calculate $I(E;T)$.

5. Calculate $I(M;T)$.

6. Comparing $I(M;T)$ to $I(E;T)$, what Information Theory shows on the meteorologist's forecast and that of the student?

7. The student claims to have found a revolutionary method of predicting the weather. Its revised performance are provided in the table above. As before, the rows correspond to the forecast, and the columns to the actual weather.

$P(E=i,T=j)$	sunny weather $(T=1)$	rainy weather $(T=0)$
sunny weather $(E=1)$	403/512	93/512
rainy weather $(E=0)$	13/512	3/512

Calculate the probabilities $P(E=0)$ and $P(E=1)$.

8. Compare $P(E=i,T=j)$ and $P(E=i)\,P(T=j)$, for all i, $j \in \{0,1\}$. Conclude.

9. We wish to store T by using a binary coding. Using Shannon's first theorem, give the minimum average memory space required to store T, in bits per realization of T.

11. Redo the previous calculation in the case of M. Calculate the minimum memory space required to store M and T separately, in bits per realization of (M,T)?

12. Calculate the minimum memory space required to store M and T jointly, in bits per realization of (M,T)?

13. Interpret the difference between results of the 2 previous questions.

14. Propose Huffman coding to jointly encode M and T.

15. Calculate the average length of words $\overline{n}$ of the binary code found in the previous question. What double inequality is satisfied by $\overline{n}$?

$P(X;Y)$	y_1	y_2	y_3	y_4	$P(x=x_e)$
x_1	$1/24$	$1/12$	$1/6$	$1/24$	$1/3$
x_2	$1/6$	$1/8$	$1/24$	$1/6$	$1/2$
x_3	$1/24$	$1/24$	$1/24$	$1/24$	$1/6$
$P(y=y_k)$	$1/4$	$1/4$	$1/4$	$1/4$	

$$I(X,Y) = H(Y) - H(Y|X)$$
$$= 2 - H(Y|X)$$

$$H(Y|X) = \sum_{i=1}^{3} H(Y|X=x_i)\, P(X=x_i)$$

To calculate $H(Y|X=x_i)$

we need : $P(Y|X=x_i)$

$P(Y\|X)$	y_1	y_2	y_3	y_4	
x_1	$1/8$	$1/4$	$1/2$	$1/8$	$\cdot/\cdot\ \frac{1}{3}$
x_2	$1/3$	$1/4$	$1/12$	$1/3$	$\cdot/\cdot\ \frac{1}{2}$
x_3	$1/4$	$1/4$	$1/4$	$1/4$	$\cdot/\cdot\ \frac{1}{6}$

$$P(Y=y_j|X=x_i) = \frac{P(X=x_i, Y=y_j)}{P(X=x_i)}$$

$$H(Y|X=x_1) = H\left(\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8}\right)$$
$$H(Y|X=x_2) = H\left(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{4}, \tfrac{1}{12}\right)$$
$$H(Y|X=x_3) = H\left(\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}\right) = 2 \text{ Sh (b)}$$

$$I(X,Y) = 2 - \left[\tfrac{1}{3} H\left(\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8}\right) + \tfrac{1}{2} H\left(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{4}, \tfrac{1}{12}\right)\right.$$

$$+ \frac{1}{6} H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) \Big]$$

$$= 0,15 \text{ Sh} \;/\; \text{pair}(x, y)$$