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.

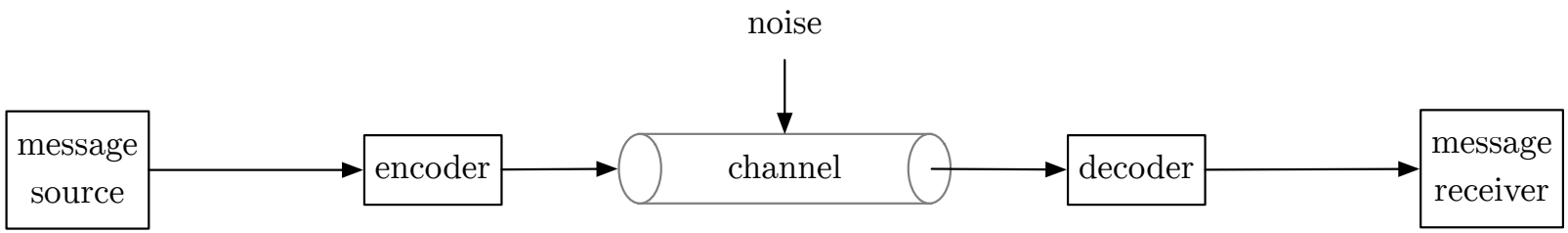
Following the basic concept, communication is the process of sending and receiving messages or transferring information from one part (sender) to another (receiver).

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

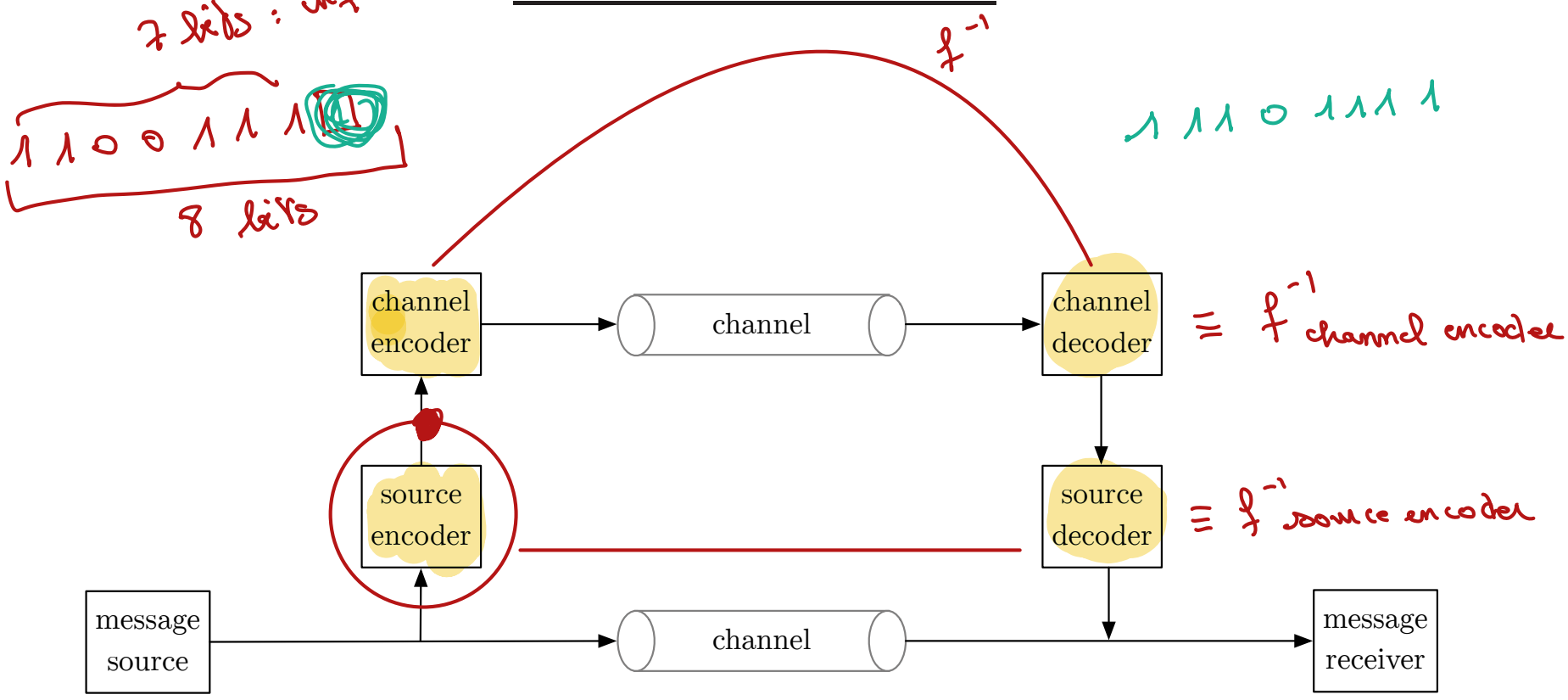
Shannon's communication model



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_1 and A_2 two events that are independent

ex: A_1 = " getting head with coin 1 "

A_2 = " getting head with coin 2 "

Then: $P(A_1 \text{ and } A_2) = P(A_1) \cdot P(A_2)$

↑ def. of independance

" and " , " given that "

$P(A_1, A_2) = P(A_2 | A_1) P(A_1)$

$= P(A_1 | A_2) P(A_2)$

with independence of A_1 and A_2 :

$P(A_1, A_2) = P(A_1) P(A_2)$

$\Leftrightarrow \quad P(A_2 | A_1) = P(A_2)$

$P(A_1 | A_2) = P(A_1)$

▷ information provided by 2 independent events: $\boxed{f(p_1 \cdot p_2) = f(p_1) + f(p_2)}$

$h(A_1, A_2) \triangleq f\left(\frac{1}{P(A_1, A_2)}\right) = f\left(\frac{1}{P(A_1)} \cdot \frac{1}{P(A_2)}\right)$

(A_1 and A_2 are indep)

$h(A_1, A_2)$: joint quantity of information

$P(A_1, A_2)$: joint proba

I want this for 2 indep. events (Shannon's axiom)

$= f\left(\frac{1}{P(A_1)}\right) + f\left(\frac{1}{P(A_2)}\right)$

$= h(A_1) + h(A_2)$

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

$$f(p) = -\log_b p$$

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

base of log

$b > 0$

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

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

$\Rightarrow f(p^{1/n}) = \frac{1}{n} f(p)$ 2.

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

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

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

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

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

$\ln p < 0$, $\ln q < 0$, $\frac{\ln p}{\ln q} > 0$

$q^{\log_q p} \overset{a}{=} q^{\frac{\ln p}{\ln q}} \overset{(a)\&(b)}{=} e^{\frac{\ln p}{\ln q} \times \ln q} = e^{\ln p} = p$

(b) $q = e^{\ln q}$

(a) $e^{ab} = (e^a)^b$

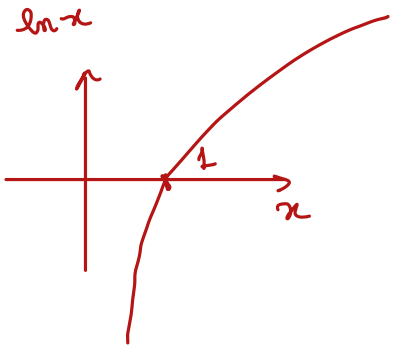
we have : $p = q^{\log_q p}$, $p, q \in]0,1]$

$$f(p) = f\left(q^{\log_q p}\right) = \log_q p \times f(q) = \frac{\ln p}{\ln q} \times f(q)$$

$$\Rightarrow \frac{f(p)}{f(q)} = \frac{\ln p}{\ln q} \Rightarrow f(p) = K \underbrace{\ln p}_{<0}$$

$\forall p, q \in]0,1]$

I want it positive



$\Rightarrow \quad K < 0$

I can write $K = -\frac{1}{\ln b}$ with $b > 1$

$$f(p) = -\frac{\ln p}{\ln b}, \quad b > 1$$

$$\triangleq -\log_b p, \quad b > 1$$

Check: $\lim_{p \to 1} -\log_b p = 0$

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:*

$$h(A) = -\log_b P(A). \qquad , \ b > 1$$

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

$b = 2$
- ▷ $\log_2$: Shannon, bit (binary unit)

$b = e^1$
- ▷ $\log_e$: logon, nat (natural unit) — ln

$b = 10$
- ▷ $\log_{10}$: Hartley, decit (decimal unit)

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

$\log_2$ x Shannon

8, 16

2048

if $x = 2^n$

$$\log_2 x = \log_2 2^n$$

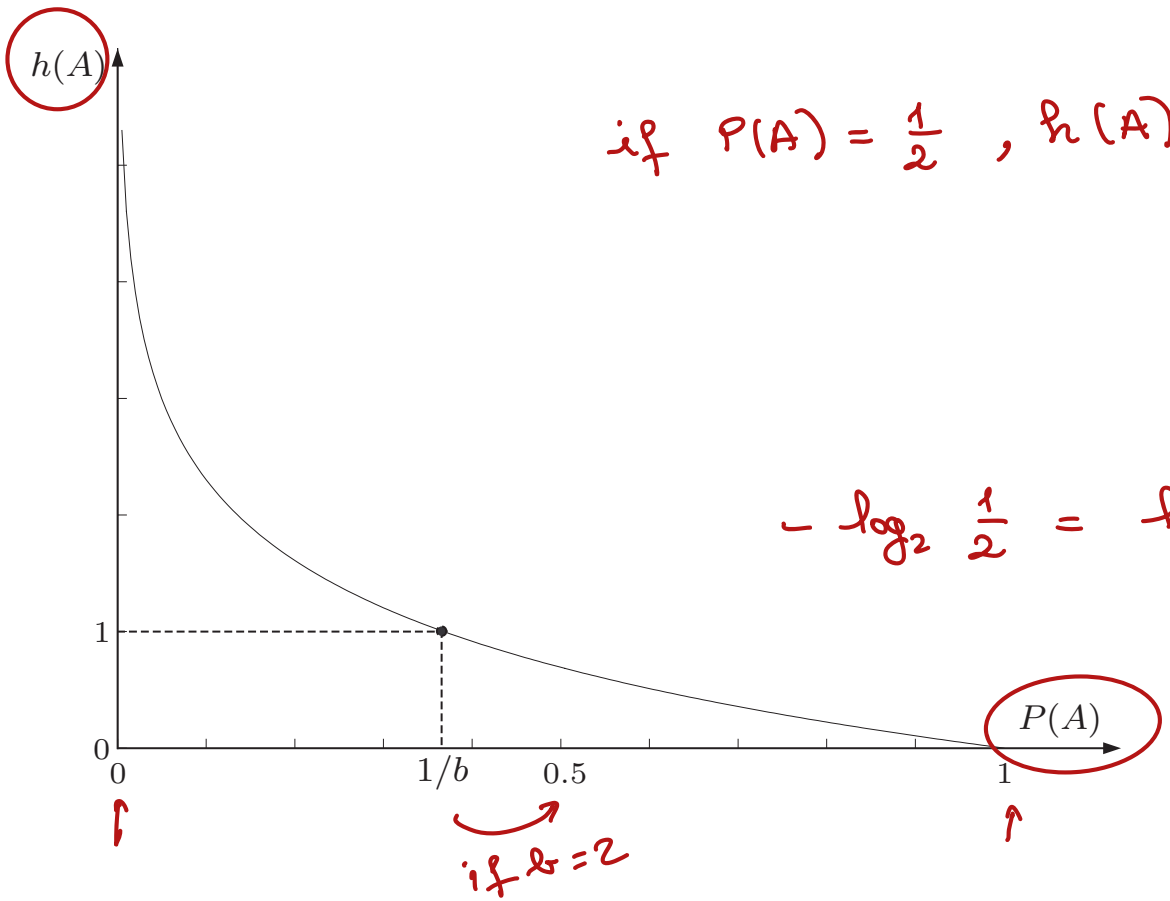
$$= \frac{\ln 2^n}{\ln 2}$$

$$= n \frac{\ln 2}{\ln 2}$$

Self-information

Information content

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



if $P(A) = \frac{1}{2}$, $h(A) = -\log_2 \frac{1}{2}$

$= 1$ Sh

$-\log_2 \frac{1}{2} = \log_2 2^1 = 1$

if $b = 2$

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 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.

$$P(S = s_i) = \frac{1}{16}$$

$$h(\underbrace{S = s_i}_{A:\text{ "}S\text{ generated }s_i\text{"}}) = -\log_2 \frac{1}{16} = \log_2 2^4 = 4 \text{ Sh (bit)}$$

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

A : with $P(A)$

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

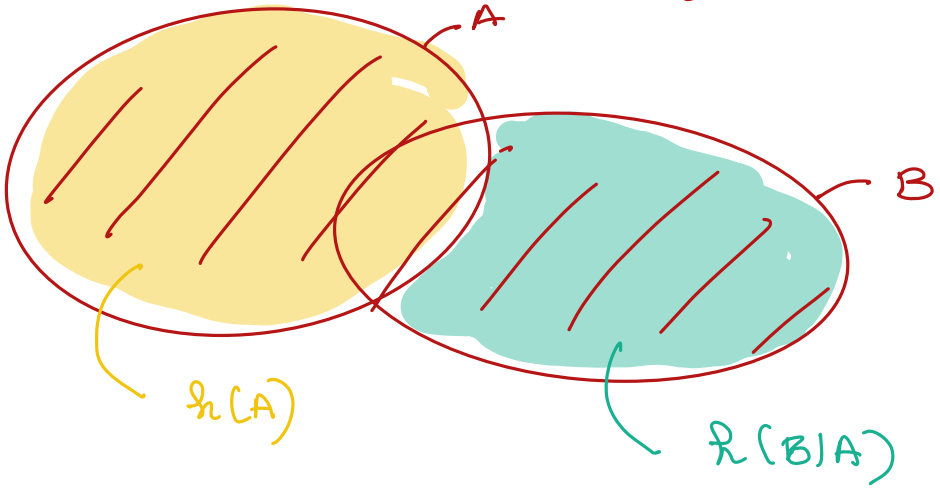
A, B : with $P(A,B)$ joint proba.

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

extension : $P(A,B) = P(A)\,P(B|A)$

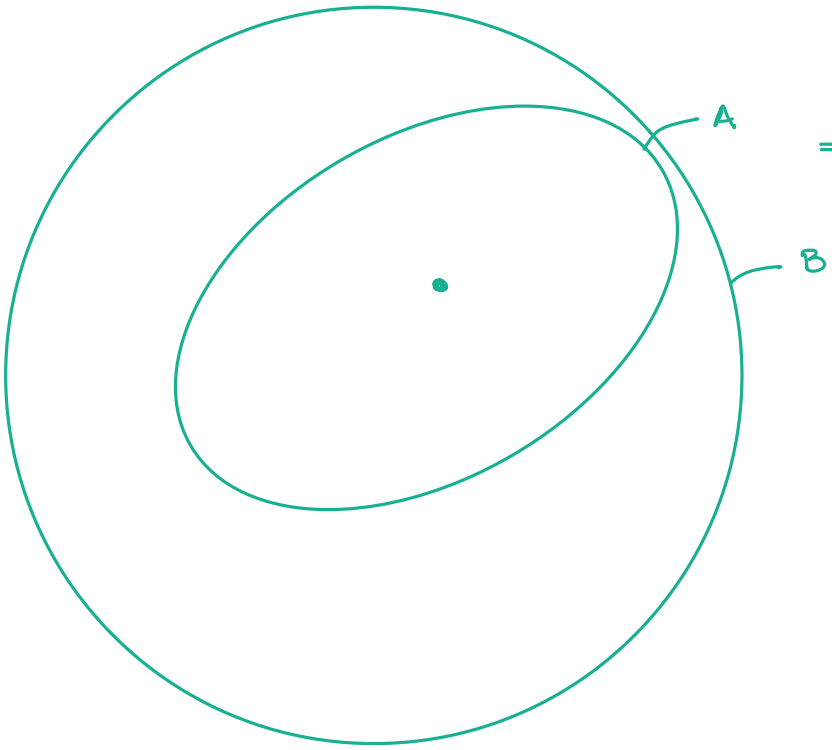
$$\begin{aligned} h(A,B) &= -\log_2 P(A,B) \\ &= -\log_2 [P(A)\,P(B|A)] \\ &= -\log_2 P(A) - \log_2 P(B|A) \\ &\overset{\Delta}{=} \underbrace{h(A)}_{\text{Sh}} + \underbrace{h(B|A)}_{\text{Sh}} \end{aligned}$$

↳ quantity of info provided by B given that you know A



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

$A \subset B$



$A \subset B$

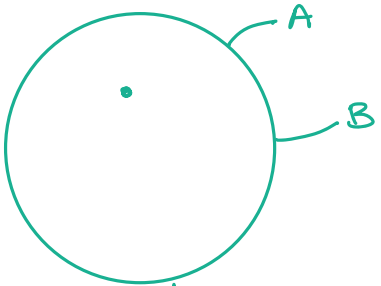
$\Rightarrow P(B|A) = 1$

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

$$h(B|A) = -\log_2 P(B|A) = 0$$

$$\Rightarrow h(A,B) = h(A)$$

$A = B$



$P(A|B) = P(B|A) = 1$

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

$A \cap B = \emptyset$ → independent

$$P(A,B) = P(A)P(B)$$

$$\Rightarrow h(A,B) = h(A) + h(B)$$

because $P(A|B) = P(A)$

$P(B|A) = P(B)$

Self-information

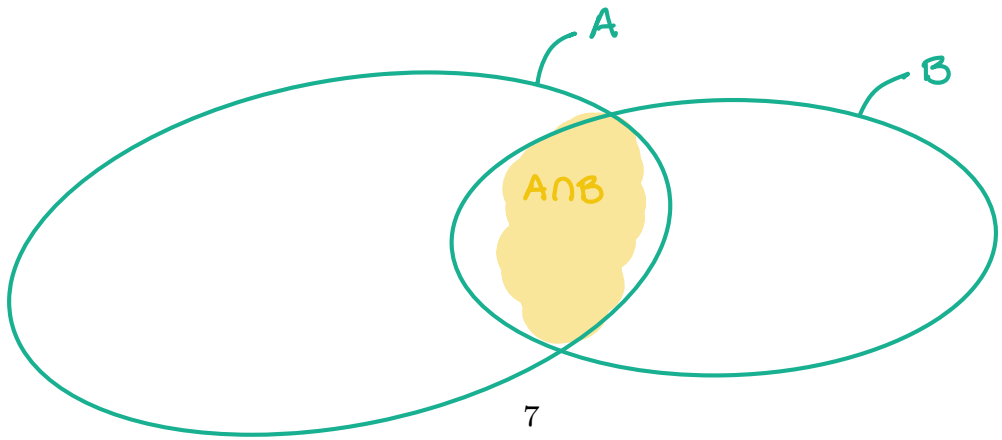
Mutual information content

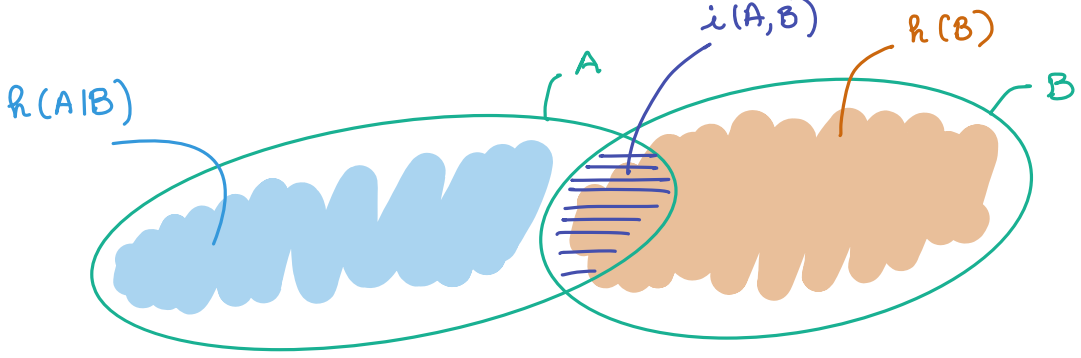
The definition of conditional information leads directly to another definition, that of mutual information, which measures information shared by two events.

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

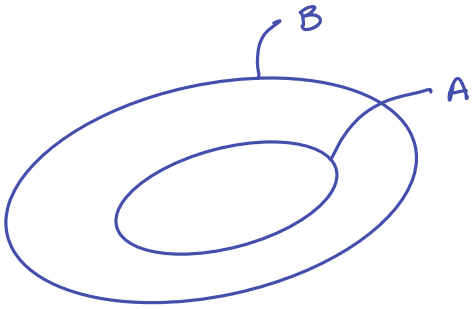




$h(A) - h(A|B) \triangleq i(A,B) \triangleq h(B) - h(B|A)$

↳ mutual information
" shared information "

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



$P(B|A) = 1$

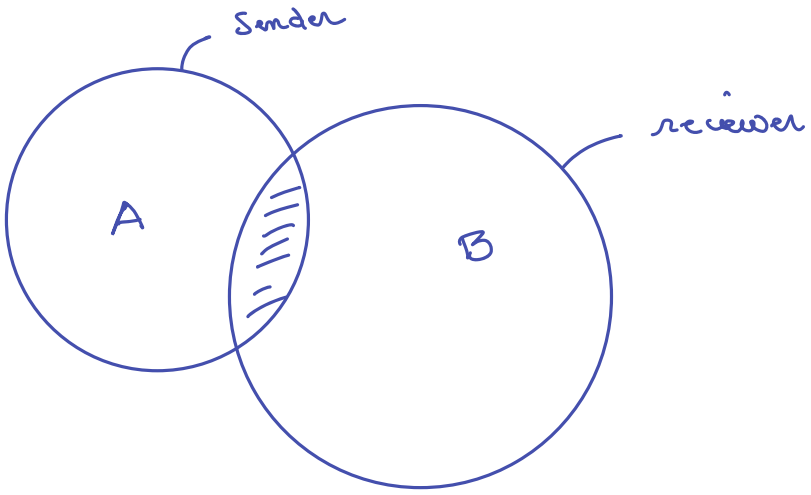
$h(B|A) = 0$

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

$\boxed{A = B}$ $\quad i(A,B) = h(A) = h(B)$

$\boxed{A \cap B = \emptyset}$ → independent.

$h(A|B) = h(A) \Rightarrow i(A,B) = 0$



good telecommunication : max i (S, R)

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:*

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

$S = \{ s_1, \dots, s_n \}$ ← event, state, character

$P(S = s_1) = p_1 \Rightarrow h(S = s_1) = -\log p_1$ Sh

$(\dots)$

$P(S = s_n) = p_n \Rightarrow h(S = s_n) = -\log p_n$ Sh

$H(S) = \sum_{i=1}^{n} p_i \, h(S = s_i)$ ← mean value $E\{\cdot\}$

↳ entropic → Sh / event or state or character.

$$H(S) = \sum_{i=1}^{n} -p_i \log_2 p_i \quad \text{Sh / event}$$

$$\boxed{H(S) = -\sum_{i=1}^{n} p_i \log_2 p_i \quad \text{Sh / event state}}$$

Alphabetical source : $a, b, c, \dots, z$

↳ $P(S = a) = \dots$

$p \log p$ with $p = 1 \longrightarrow 0$

$p \log p$ with $p \to 0 \longrightarrow 0$

if : $P(S = s_i) = 1$

and $P(S = s_j) = 0 \quad , \quad \forall j \neq i$

$\Rightarrow H(S) = 0$

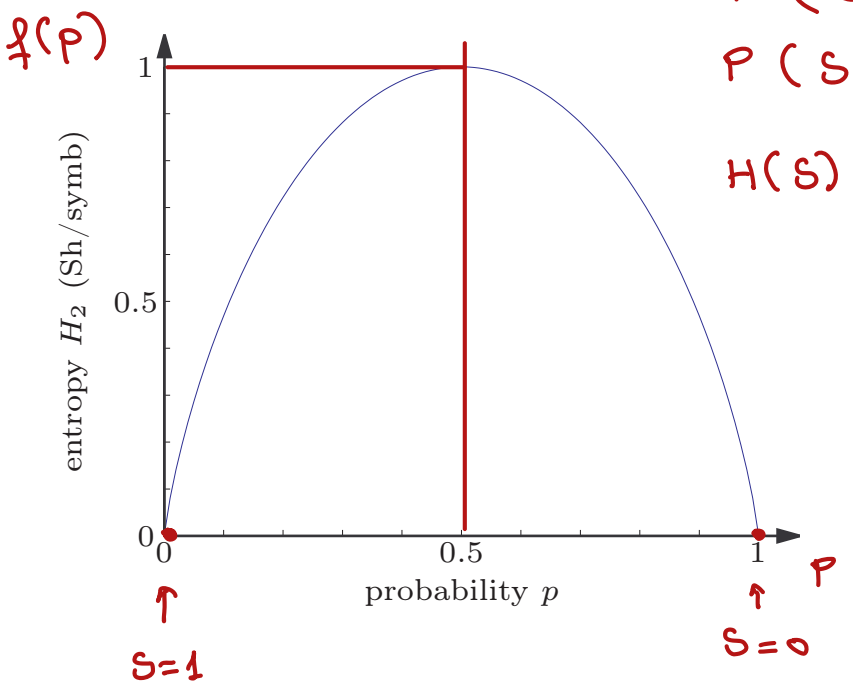
Entropy of a random variable

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.



$P(S=0) = p$

$P(S=1) = 1-p$

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

$= f(p)$

$$H(S) = 2 \times \left[-\frac{1}{2} \log \frac{1}{2} \right]$$

$$= 2 \times \frac{1}{2}$$

$$= 1 \quad \text{Sh / state.}$$

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:*

$\sum_i p_i = 1$, $p_i \geq 0$ $\forall i$

$\sum_i q_i = 1$, $q_i \geq 0$ $\forall i$

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

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:

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

We have : $\ln x \leq x - 1 \;,\; \forall x > 0$

$q_i / p_i > 0$

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

$$\sum_{i=1}^{n} p_i \log \frac{x}{\frac{q_i}{p_i}} \leq \underbrace{\sum_{i=1}^{n} p_i \times \overbrace{\left(\frac{q_i}{p_i} - 1 \right)}^{x-1}}$$

$$\sum_{i=1}^{n} \left(q_i - p_i \right)$$

$$\sum_{i=1}^{n} q_i - \sum_{i=1}^{n} p_i$$

$$1 - 1$$

$$0$$

if $p_i = q_i \;,\; \forall i$

$$\Rightarrow \sum_i p_i \log \frac{q_i}{p_i} = 0$$

$$H(S) = - \sum_i p_i \log_2 p_i$$

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

$$\Rightarrow \sum_{i=1}^{n} \left[p_i \log q_i - p_i \log p_i \right] \leq 0$$

$$\Rightarrow \underbrace{- \sum_{i=1}^{n} p_i \log p_i}_{H(S)} + \sum_{i=1}^{n} p_i \log q_i \leq 0$$

We set $q_i = \frac{1}{n}$

$$H(S) - \log_2 n \underbrace{\sum_{i=1}^{n} p_i}_{1} \leq 0$$

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

2 states : $n = 2$

$H(S) \leq 1$ Sh

(see before)

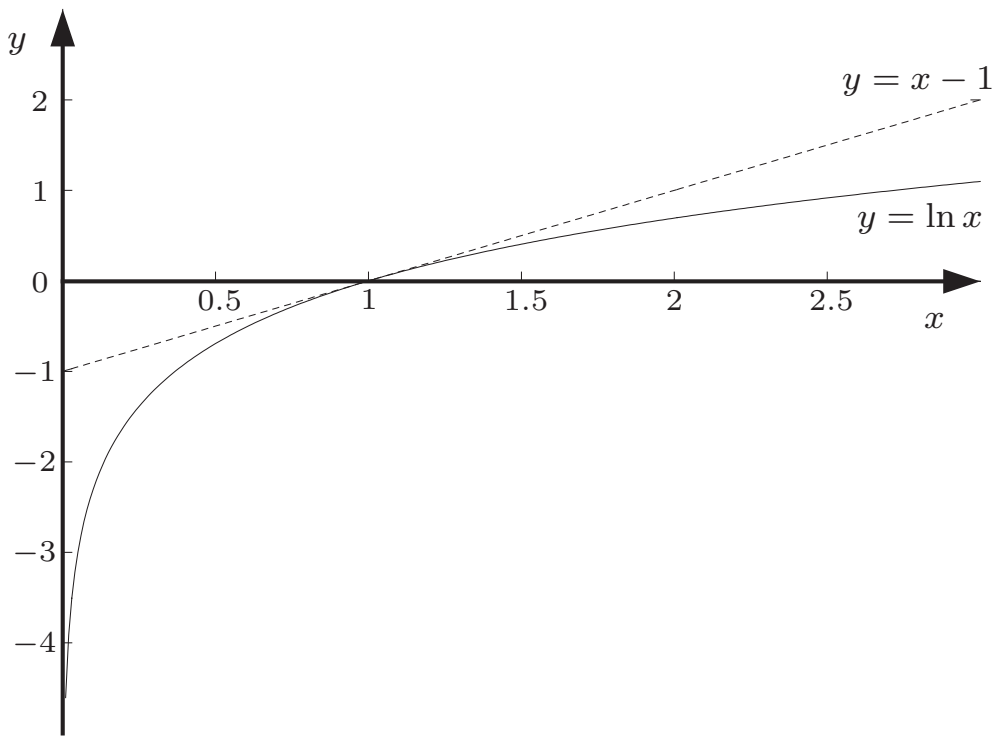
$$H(S) = \log_2 n$$

if $p_i = q_i = \frac{1}{n}$, $\forall i = 1, \dots, n$

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.

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:

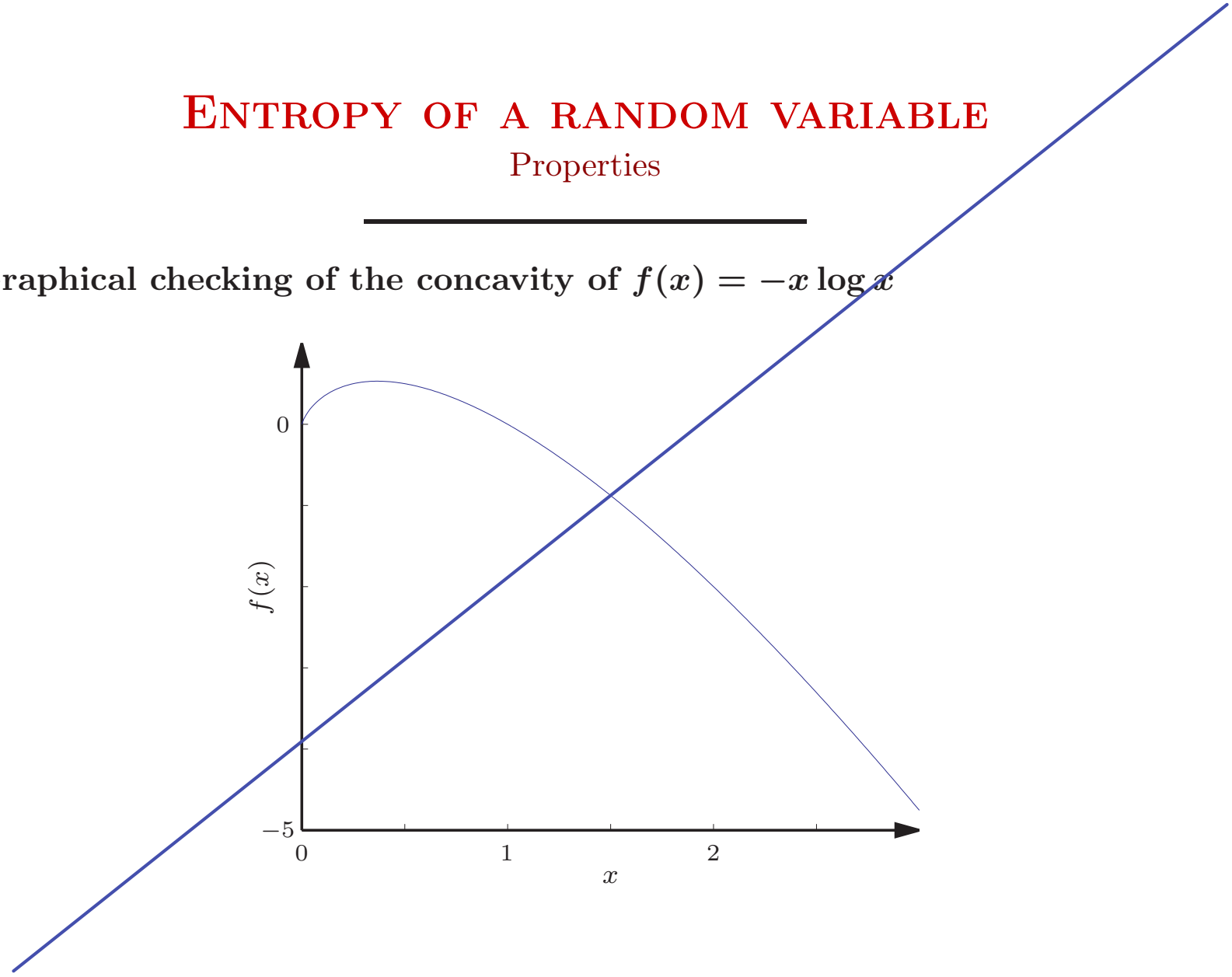
$$H_n(\lambda p_1 + (1-\lambda)q_1, \ldots, \lambda p_n + (1-\lambda)q_n) = \sum_{i=1}^{n} f(\lambda p_i + (1-\lambda)q_i).$$

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

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

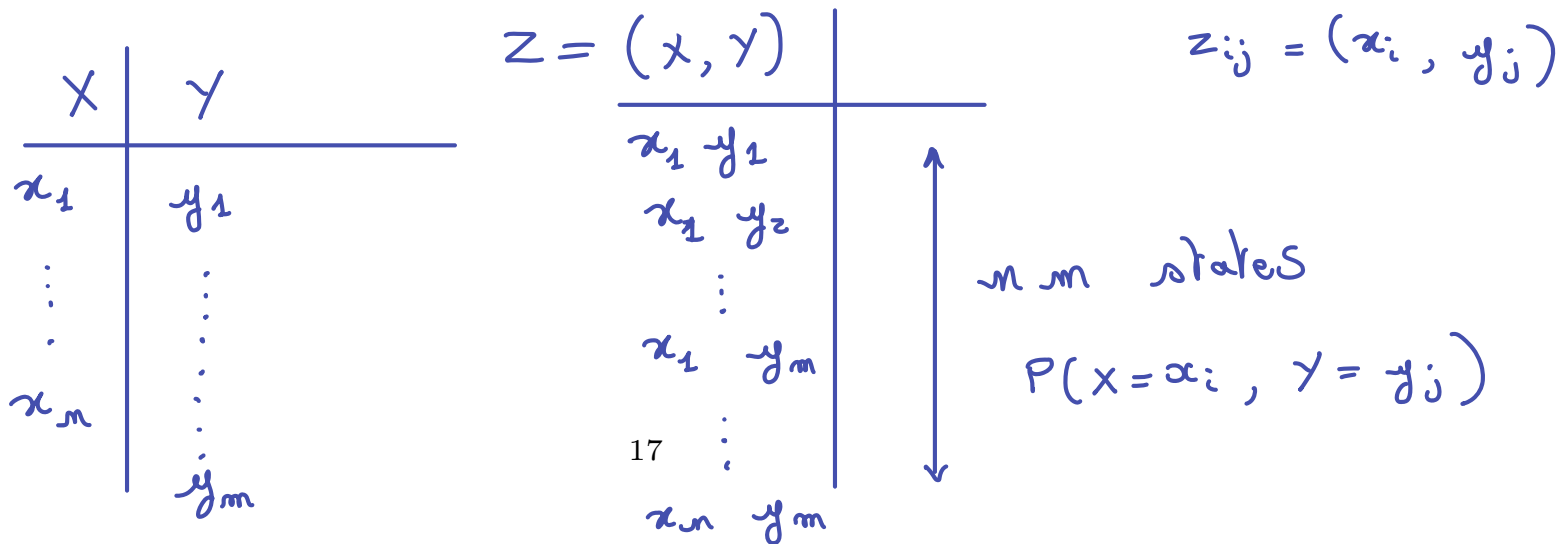
Joint entropy

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:*

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

▷ The joint entropy is symmetric: $H(X,Y) = H(Y,X)$ can $\mathbb{P}(x,y) = \mathbb{P}(y,x)$

Example. Case of two independent random variables



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

$$= -\sum_{i=1}^{i=n}\sum_{j=1}^{m} P(Z=z_{ij}) \log_2 P(Z=z_{ij})$$

$$= -\sum_{i}\sum_{j} P(X=x_i, Y=y_j) \log_2 P(X=x_i, Y=y_j)$$

Sh / pair of states of X and Y

$$H(X,Y) \leq \log_2(nm) \quad (\log_2 n + \log_2 m)$$

if X and Y indep. $H(X,Y) = H(X) + H(Y)$

$H(X) \leq \log n$ $\quad$ $H(Y) \leq \log m$

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

$H(X \mid Y = y_j)$ ⟶ X : system; Y : fixed to y_j } → single system $X \mid Y = y_j$

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

Sh / state of X ; fixed

Rm : $H(X \mid Y = y_j) \leq \log_2 n$

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

mean value over all y_j

$$\triangleq E_Y \, H(X \mid Y = y)$$

⚠ $H(X \mid Y) \neq H(Y \mid X)$

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

Example. Case of two independent random variables

X and Y independant

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

because $P(X=x_i|Y=y_j) = P(X=x_i)$

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

$$= H(X) \underbrace{\sum_{j=1}^{m} P(Y=y_j)}_{1}$$

$$= H(X)$$

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:*

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

$$P(x = x_i, y = y_j) = P(x = x_i)\, P(y = y_j \mid x = x_i)$$

$-\log$

$$\underbrace{-\log P(x = x_i, y = y_j)}_{h(x = x_i,\, y = y_j)} = \underbrace{-\log P(x = x_i)}_{h(x = x_i)} \underbrace{-\log P(y = y_j \mid x = x_i)}_{h(y = y_j \mid x = x_i)}$$

$E_{xy}\{\cdot\}$

Rm: $h(A, B) = h(A) + h(B \mid A)$

$A = (x = x_i)$

$B = (y = y_j)$

$$H(x) = E_x\{h(x = x)\}$$

$$H(x, y) = E_{x,y}\{h(x = x, y = y)\}$$

$$H(x, y) = E_{xy}\{h(x = x)\} + E_{xy}\{h(y = y \mid x = x)\}$$

$$= E_y\{E_x\{h(x = x)\}\} + E_y\{E_x\{h(y = y \mid x = x)\}\}$$

$$= E_y\{H(x)\} + E_x\{H(y \mid x = x)\}$$

$$= H(x) + H(y \mid x)$$

$$\boxed{H(x, y) = H(x) + H(y \mid x)}$$

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:

$$\boxed{\begin{aligned} H(X) &\leq H(X,Y) \\ H(Y) &\leq H(X,Y) \end{aligned}}$$

$\underbrace{H(X,Y)}_{\geqslant 0} = \underbrace{H(X)}_{\geqslant 0} + \underbrace{H(Y|X)}_{\geqslant 0} \Rightarrow$ $\boxed{\begin{aligned} H(X,Y) &\geqslant H(X) \\ H(X,Y) &\geqslant H(Y|X) \end{aligned}}$

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:

$$\boxed{H(X|Y) \leq H(X)}$$

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:

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

① : Entropie is positive (linear combination of $f(x) = -x \log x$, $0 \leq x \leq 1$)

② see previous slide (concavity) – not demonstrated.

③ $H(X,Y) = H(X) + H(Y|X) \Rightarrow H(X) \leq H(X,Y)$ because $H(Y|X) \geqslant 0$

④ $H(X,Y) \overset{③}{=} H(X) + H(Y|X)$ and $H(Y|X) \overset{②}{\leqslant} H(Y)$

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

⑤ ③ applied to X and to Y $\quad H(X) \leqslant H(X,Y)$ + $H(Y) \leqslant H(X,Y)$

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

Pair of random variables

Mutual information

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

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

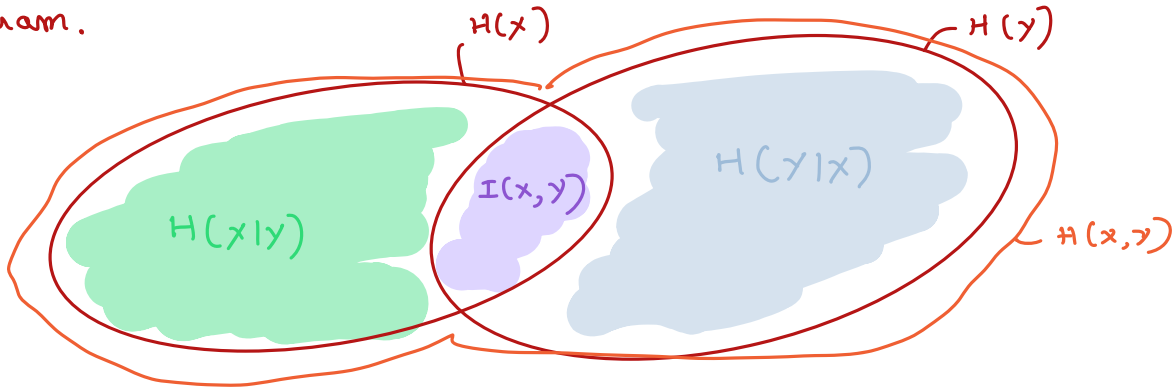
or, equivalently,

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

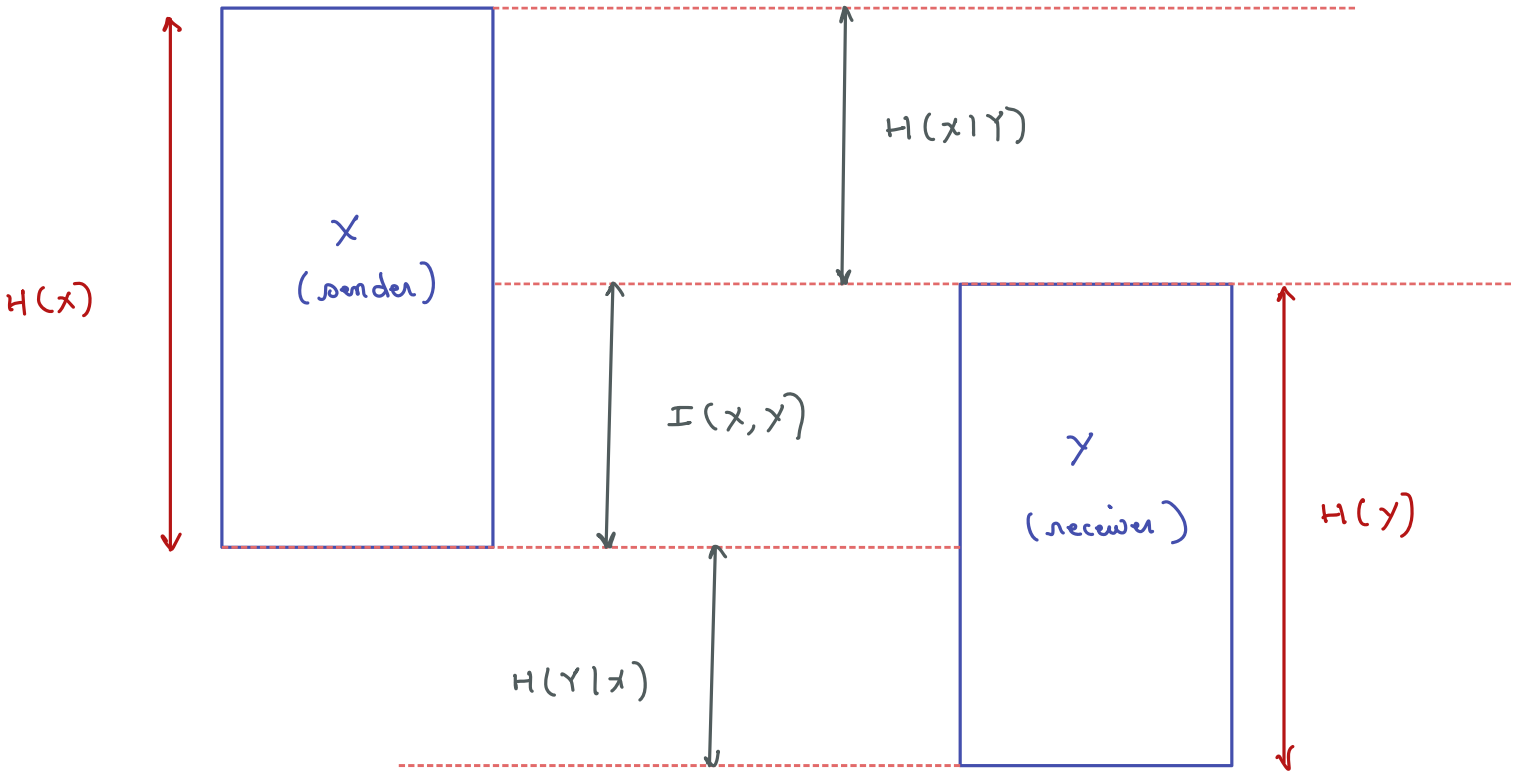
The mutual information quantifies the amount of information obtained about one random variable through observing the other random variable.

Exercise. Case of two independent random variables

Venn diagram.



$$H(x,y) = I(x,y) + H(x|y) + H(y|x)$$



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:*

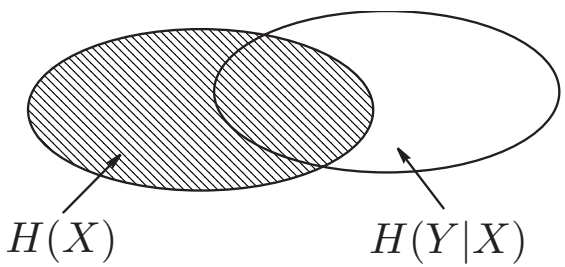
$$d(P_1, P_2) = \sum_{x \in X(\Omega)} P_1(X = x) \log \frac{P_1(X = x)}{P_2(X = x)}.$$

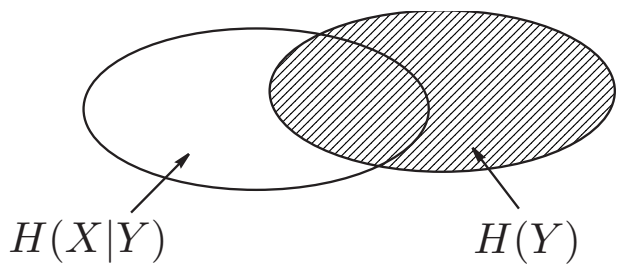
The mutual information corresponds to the Kullback-Leibler distance between the marginal distributions and the joint distribution of X and Y.

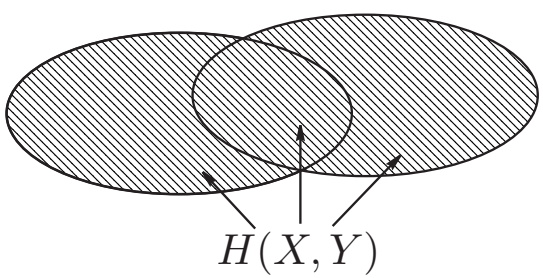
Pair of random variables

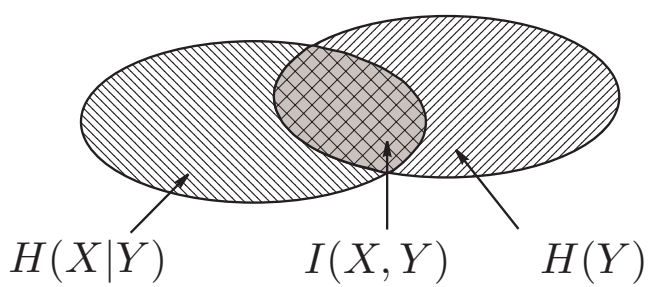
Venn diagram

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









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 $\mathcal{E}_k$.

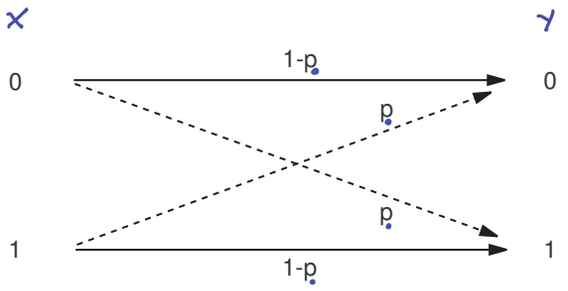


Figure 1: Imperfect channel.

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.

1. $S = \{ 1H, 1L, 2H, 2L, \dots, 9H, 9L \}$

$H(S) = \log_2 18 \simeq 4,17$ Sh

2. $P = \{ \text{left}, \text{right}, \text{equilibrium} \}$

$H_{max}(P) = \log_2 3$ Sh / weighting op.

$\overline{n}_{min} = \dfrac{\log_2 18}{\log_2 3} \approx 2,6$ weighting op.

3. $2n$ coins among 9 coins

$$P(\text{equilibrium}) = \frac{C_8^{2n}}{C_9^{2n}} = \frac{\frac{8!}{2n!\,(8-2n)!}}{\frac{9!}{2n!\,(9-2n)!}}$$

$$= \frac{9-2n}{9}$$

$$= 1 - \frac{2}{9}n$$

$$P(\text{left}) = P(\text{right}) = \frac{1}{2}\left[1 - P(\text{equilibrium})\right]$$

$$= \frac{1}{9}n$$

We want $P(\text{left}) = P(\text{right}) = P(\text{equi})$ to get max information.

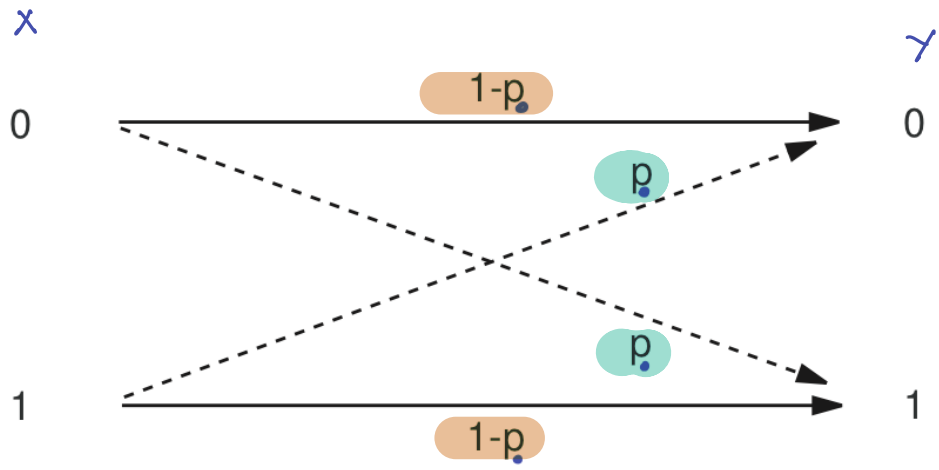
$$1 - \frac{2}{9}n = \frac{1}{9}n \Rightarrow 1 = \frac{1}{3}n$$

$$\Rightarrow \boxed{n = 3}$$

3 coins on each side / weighting operation $\longrightarrow$ max of info.

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 $I(X,Y)$.



$P(x=0) = \frac{1}{4}$

$P(x=1) = \frac{3}{4}$

$P(y=0 \mid x=0) = P(y=1 \mid x=1) = 1 - p_0$

$P(y=0 \mid x=1) = P(y=1 \mid x=0) = p_0$

$$H(X) = -\sum_{i=1}^{2} P(x=i) \log_2 P(x=i)$$

$$= -\frac{1}{4} \log_2\left(\frac{1}{4}\right) - \frac{3}{4} \log_2\left(\frac{3}{4}\right)$$

$$= 0,81 \text{ Sh /state of } x$$

$H(y) = ?$ Need to calculate

$P(y=0) = 0,3$

$P(y=1) = 0,7$

$H(y) = 0,88$ Sh /state of y

$H(X,Y) = ?$ Need to calculate $P(X=i, Y=j)$ for all $i = 1,2$, $j = 1,2$

X	Y	$(X,Y) = Z$
0	1	(0,1)
0	0	(0,0)
1	0	(1,0)
1	1	(1,0)

$H(X,Y) = 1,28$ Sh / state of (X,Y)

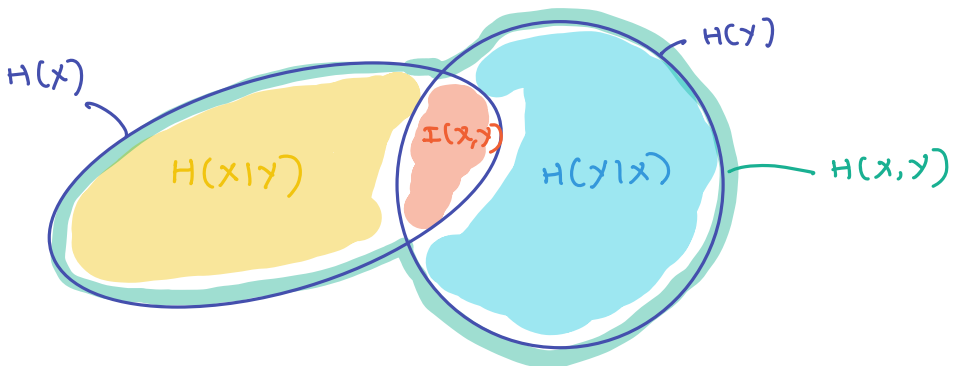
Rm: $\max_{P(Z)} H(Z) = \log_2 n$ Sh / state of Z

n: number of states of X

if $n = 2$: $H_{max}(Z) = \log_2 2 = 1$ Sh / state

if $n = 4$: $H_{max}(Z) = \log_2 4 = 2$ Sh / state

$H(X|Y)$ and $H(Y|X)$



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

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

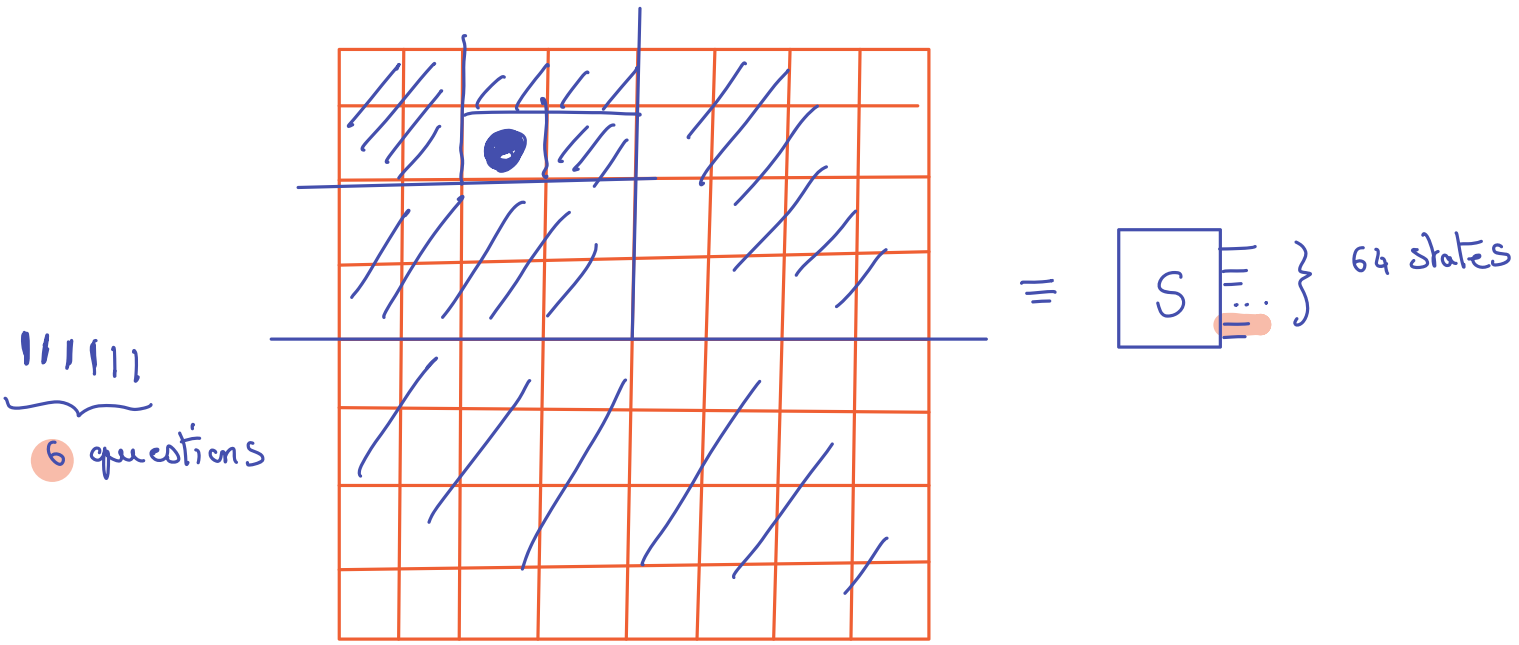
$I(X,Y) = H(X) - H(X|Y)$

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

$= 0,41$ Sh

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.



$$H(S) = \log_2 64 \text{ Sh}$$

$$= 6 \text{ Sh}$$

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 $\overline{n}$?

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 $I(X;Y)$.

Y is uniformly distributed

$$I(x,y) = H(y) - H(y|x)$$

$$= \log_2 4 - H(y|x)$$

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

$$H(y|x = x_1) = ?$$

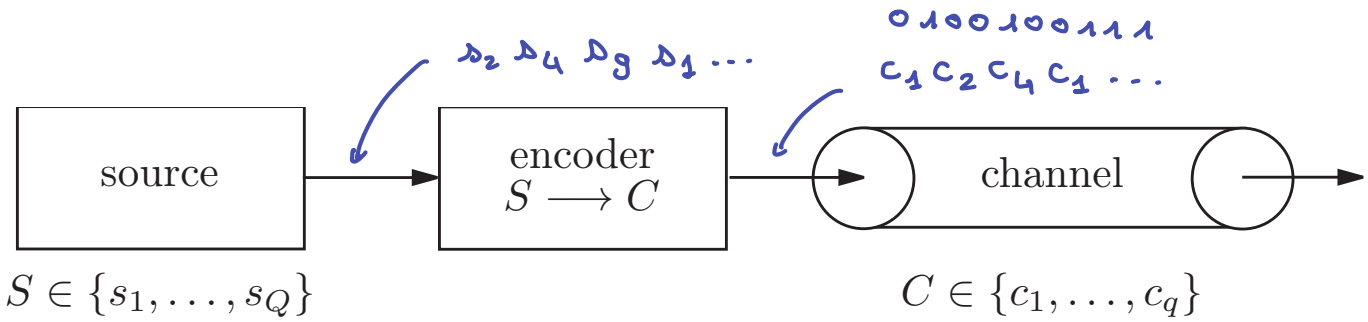
Information Theory and Coding

Discrete source coding

Cédric RICHARD
Université Côte d'Azur

Discrete source coding

Each of the Q states s_i of source S is associated with a codeword, that is, a sequence of n_i symbols of a q-ary alphabet. These constitute a source code that can be noted as follows: $C = \{c_1, \ldots, c_q\}$.



Example. The Morse code

- ▷ quaternary code (dot, dash, long space, short space)
- ▷ variable length code
- ▷ the shortest sequence is for "E"

Problem

Source coding and adaptation (ideal noiseless channel)

Let S be a source characterized by a rate D_s (Q-ary symbol per second). Consider a noiseless channel with maximum rate D_c (q-ary symbol per second). We define:

- emission rate of the source : $T \triangleq D_s\, H(S)$
- channel capacity : $C \triangleq D_c \log q$

If T > C: the channel cannot transmit the information

If T $\leq$ C : the channel can theoretically transmit the information

If we have a q-ary code where the average length $\overline{n}$ of codewords is such that $\overline{n}\, D_s \leq D_c$, then this code can be used for transmission.

Otherwise, how to encode the source words to make their transmission possible?

Source coding is used to eliminate redundant information
WITHOUT LOSS !!!

Source coding

Definitions

Source coding consists of associating to each symbol s_i generated by a source, a sequence of symbols of a q-ary alphabet, referred to as a codeword.

Example 1. ASCII (7 bits) et extended ASCII (8 bits), Morse code, etc.

Example 2.

	code A	code B	code C	code D	code E	code F	code G
s_1	1	0	00	0	0	0	0
s_2	1	10	11	10	01	10	10
s_3	0	01	10	11	011	110	110
s_4	0	11	01	110	0111	1110	111

0110

$s_3 s_2$

$s_1 s_4 s_1$

110

SOURCE CODING

Definition

Regularity. A code is said to be nonsingular if all codewords are distinct.

Decodability.

A nonsingular code is called uniquely decodable if any sequence of codewords can be decoded only in a unique way.

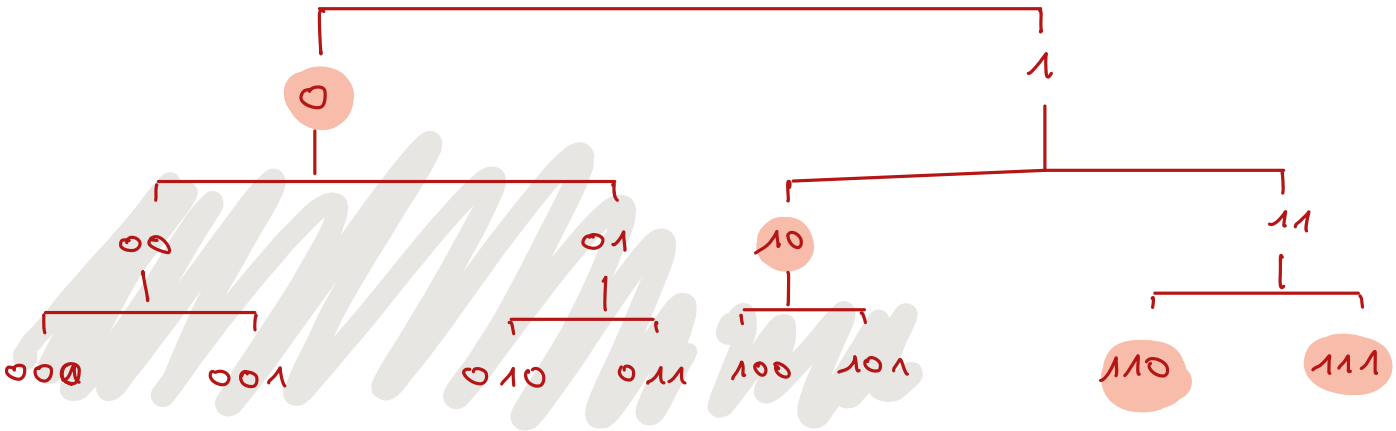
Fixed length. With fixed-length codewords, any message can be decoded without ambiguity. (C)

Separator. A symbol of the alphabet is used as a word separator. (E, F)

Without prefix. A code is called a prefix code or an instantaneous code if no codeword is a prefix of any other codeword. (G)

Exercise. Characterize codes A to G.

code G
0
10
110
111



s_3 s_2 s_1 s_4

110 | 10 | 0 | 0 | 111 → instantaneous coding method

Source coding

Expected length

Definition 1. *The expected length $\overline{n}$ of the q-ary codewords encoding the states of a source $S \in \{s_1, \ldots, s_Q\}$ with probability distribution $\{p_1, \ldots, p_Q\}$ is defined as:*

$$\overline{n} = \sum_{i=1}^{Q} p_i n_i \qquad = E\{n(S)\}$$

where n_i is the length of the codeword c_i encoding state s_i.

Example. Consider a source with 4 states $\{s_1, \ldots, s_4\}$ defined by the following distribution, with associated codewords:

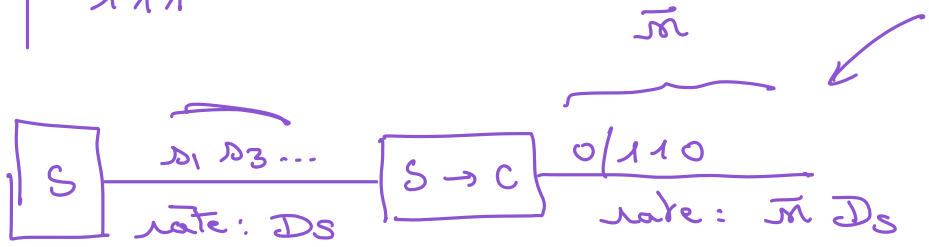
$$\begin{array}{lll} p_1 = \frac{1}{2} & c_1 = 0 & \longrightarrow \quad n_1 = 1 \\ p_2 = \frac{1}{4} & c_2 = 10 & \longrightarrow \quad n_2 = 2 \\ p_3 = \frac{1}{8} & c_3 = 110 & \longrightarrow \quad n_3 = 3 \\ p_4 = \frac{1}{8} & c_4 = 111 & \longrightarrow \quad n_4 = 3 \end{array}$$

We have $\overline{n} = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 3 \cdot \frac{1}{8} = 1.75$ binary symbol per state

$1,75 \geq \underbrace{H(S)}_{1,75}$ Sh/state

4- ary source : S

S	C : binary code
s_1	0
s_2	10
s_3	110
s_4	111



If T ≤ C : the channel can theoretically transmit the information

If we have a q-ary code where the average length $\overline{n}$ of codewords is such that $\overline{n}\, D_s \leq D_c$, then this code can be used for transmission.

Otherwise, how to encode the source words to make their transmission possible?

Source coding is used to eliminate redundant information WITHOUT LOSS !!!

$T = D_S\, H(S)$ Sh / sec. → source

$C = D_c \log_2 q$ Sh / sec → channel

if $T \leq C$: communication possible in terms of entropy

$\overline{n}\, D_S \leq D_c$: good code

$\overline{n}\, D_S \geq D_c$: bad code

Source coding

Toward Shannon's first theorem

Theorem 1. *The expected length $\overline{n}$ of the codewords of any uniquely decodable code is lower-bounded by:*

$$\frac{H(S)}{\log q} \leq \overline{n}.$$

$q = 2$ for binary code
$\log q = 1$

Condition of equality. The above inequality turns into an equality if $\sum_{i=1}^{Q} q^{-n_i} = 1$, that is, $p_i = q^{-n_i}$. This means that:

$$n_i = \frac{\log \frac{1}{p_i}}{\log q}.$$

Definition 2. *Any code where each codeword i is of length $n_i = \frac{\log \frac{1}{p_i}}{\log q}$ is absolutely optimal.*

$$\frac{H(S)}{\log q} \leq \overline{n}. \longrightarrow \frac{H(S)}{\log q} \stackrel{?}{=} \overline{n}$$

$$-\sum_{i=1}^{Q} p_i \frac{\log p_i}{\log q} = \sum_{i=1}^{Q} p_i n_i$$

$$\Rightarrow \boxed{n_i = \frac{-\log p_i}{\log q}}, \quad \forall i = 1, \dots, Q$$

$p_1 = \frac{1}{2}$	$c_1 = 0$	$\longrightarrow$	$n_1 = 1$
$p_2 = \frac{1}{4}$	$c_2 = 10$	$\longrightarrow$	$n_2 = 2$
$p_3 = \frac{1}{8}$	$c_3 = 110$	$\longrightarrow$	$n_3 = 3$
$p_4 = \frac{1}{8}$	$c_4 = 111$	$\longrightarrow$	$n_4 = 3$

$8 = 2^3$

$$\frac{-\log p_1}{\log 2} = \frac{-\log\left(\frac{1}{2}\right)}{\log 2} = 1 = n_1$$

$$\frac{-\log p_2}{\log 2} = \frac{-\log_2\left(\frac{1}{4}\right)}{\log_2} = 2 = n_2$$

Usually, the above equality condition is not satisfied because $n_i = \frac{\log \frac{1}{p_i}}{\log q}$ is not an integer. However, it is possible to construct a code such that:

$$\frac{\log \frac{1}{p_i}}{\log q} \leq n_i < \frac{\log \frac{1}{p_i}}{\log q} + 1.$$

Multiplying each member by p_i and summing over i, we obtain:

$$\frac{H(S)}{\log q} \leq \overline{n} < \frac{H(S)}{\log q} + 1.$$

Definition 3 (Shannon's code: predefined codeword lengths)**.** *We talk about a Shannon's code when:*

$$n_i = \left\lceil \frac{\log \frac{1}{p_i}}{\log q} \right\rceil.$$

Source coding

Shannon's first theorem: statement

The bounds that have just been established will allow us to demonstrate Shannon's first theorem, which reads as follows:

Theorem 2. *For any stationary source, there is a coding process to design a uniquely decodable code where the expected codeword length is as close to its lower bound as you want it to be.*

Proof in the case of a memoryless source. Consider the k^{th} extension of source S. In the case of a memoryless source:

$$\frac{kH(S)}{\log q} \leq \overline{n}_k < \frac{kH(S)}{\log q} + 1.$$

In this expression, $\overline{n}_k$ denotes the expected length of the codewords used to encode the k^{th} extension of source S. Dividing by k and calculating the limit as k tends to infinity leads to the result.

Binary coding techniques

Shannon's code: predefined codeword length

Shannon's first theorem provides an asymptotic property, but do not provide any practical method for doing so.

Shannon's coding technique consists of associating n_i q-ary symbols to each source state s_i, where:

$$n_i = \left\lceil \frac{\log \frac{1}{p_i}}{\log q} \right\rceil .$$

Binary coding techniques

Shannon's code: predefined codeword length

We consider a 5-symbol source $\{s_1, \ldots, s_5\}$ defined by probabilities:

$$
\begin{array}{lll}
p_1 = 0.35 & -\log_2 p_1 = 1.51 & \longrightarrow \quad n_1 = 2 \\
p_2 = 0.22 & -\log_2 p_2 = 2.18 & \longrightarrow \quad n_2 = 3 \\
p_3 = 0.18 & -\log_2 p_3 = 2.47 & \longrightarrow \quad n_3 = 3 \\
p_4 = 0.15 & -\log_2 p_4 = 2.73 & \longrightarrow \quad n_4 = 3 \\
p_5 = 0.10 & -\log_2 p_5 = 3.32 & \longrightarrow \quad n_5 = 4.
\end{array}
$$

We can easily get an instantaneous code that satisfies the above conditions on n_i using a tree. For instance:

$$
s_1 : 00 \qquad s_2 : 010 \qquad s_3 : 011 \qquad s_4 : 100 \qquad s_5 : 1010.
$$

This leads to $\overline{n} = 2.75$, to be compared to $H(S) = 2.19$ Sh/symb.

$$\overline{n}_{Sh} = (2 \times 0,35) + (3 \times 0,22) + (3 \times 0,18) + (3 \times 0,15) + (4 \times 0,10)$$

$$= 2,75 \text{ bit / state of } S$$

$$H(S) = -\sum_{i=1}^{5} p_i \log_2 p_i = 2,19 \text{ Sh / state of } S$$

Binary coding techniques

Shannon-Fano's code

Shannon-Fano's code is the first code that started to exploite the redundancy of a source. Its principle is now outlined.

1. Arrange the states of the system by decreasing probabilities.
2. Split the system states into 2 groups G_0 et G_1 with probabilities as close as possible without *modifying* their arrangement in 1.
3. Each group G_i is split into 2 sub-groups G_{i0} et G_{i1} with probabilities as close as possible to each other, again without modifying the state arrangement.
4. The procedure stops when each subgroup consists of a single element. The index of the group gives the codeword.

Binary coding techniques

Shannon-Fano's code

To design a Shannon-Fano's code, we proceed as follows:

0,5 → 0,57

0,5 → 0,43

state	p_i	step 1	step 2	step 3	code
s_1	0.35	0	0		00
s_2	0.22	0	1		01
s_3	0.18	1	0		10
s_4	0.15	1	1	0	110
s_5	0.10	1	1	1	111

This leads to $\overline{n} = 2.25$, to be compared to $H(S) = 2.19$ Sh/symb.

Binary coding techniques

Huffman's code

Huffman's method provides a compact instantaneous code of minimum average length. To achieve this, it exploits the following property.

Lemme 1. *For any source, there is an instantaneous code of minimum expected length that satisfies the following properties.*

1. *If* $P(S = s_i) > P(S = s_j)$*, then* $n_i \leq n_j$*.*
2. *The two longest words, therefore associated with the least likely states, have the same length and differ by only one bit.*

Huffman's method involves grouping the two least likely states together and then treating them as one by summing their probabilities. This technique is then repeated on the remaining states until only two remain.

Binary coding techniques

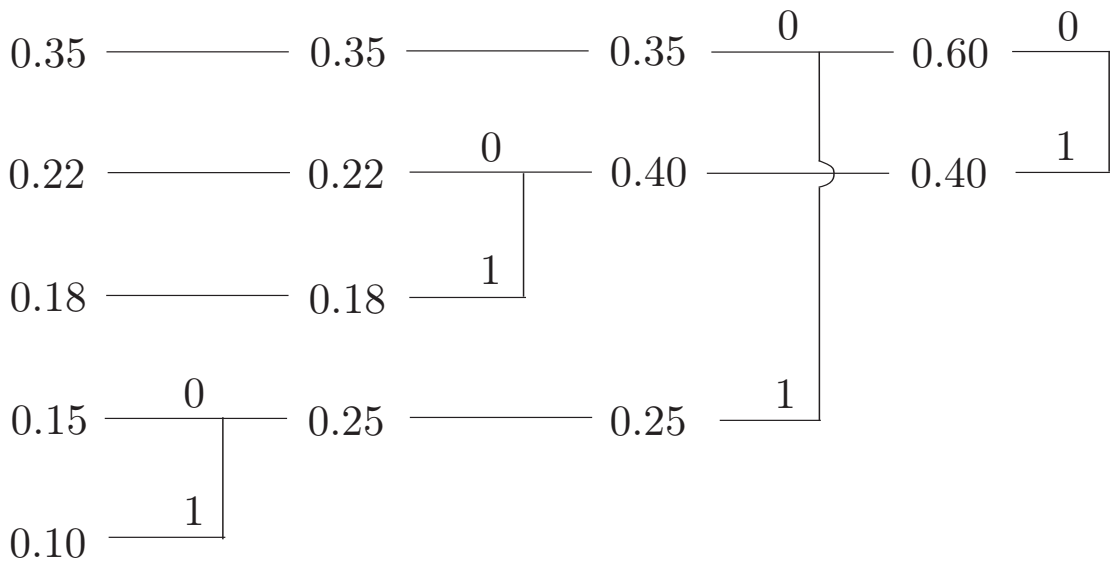
Huffman's code

A tree is built from the leaf nodes, which represent the states of the source.

1. At each step, the two least likely leaves are merged into one.
2. The procedure stops when the result is a single leaf consisting of all the symbols.
3. The reverse path of the tree provides the code words.

Binary coding techniques

Huffman's code



The reverse exploration of the tree provides the following code words:

$$s_1 : 00 \qquad s_2 : 10 \qquad s_3 : 11 \qquad s_4 : 010 \qquad s_5 : 011.$$

This leads to $\overline{n} = 2.25$, to be compared to $H(S) = 2.19$ Sh/symb.

Discrete-time source

General model

A discrete source S is characterized by its states $\{s_1, \ldots, s_Q\}$ and an emission mechanism. It is a discrete-time random process

$$S_1, \ldots, S_{i-1}, S_i, S_{i+1}, \ldots$$

characterized by joint laws:

$$P(S_1, \ldots, S_n), \forall n \in \mathbb{N}^*$$

▷ **model too general to give rise to tractable developments**

Discrete-time source

Complementary assumptions

For simplicity, assumptions need to be made about the source.

Property 1 (Stationary process)**.** *A random process S_i is said to be stationary if the laws that govern it are independent of the origin of time, that is,*

$$P(S_1 = s_{i_1}, \ldots, S_n = s_{i_n}) = P(S_{n_0+1} = s_{i_1}, \ldots, S_{n_0+n} = s_{i_n}),$$

for all positive n_0 and n.

Example. A memoryless source is characterized by independent and identically distributed S_i. This is a stationary process.

$$P(S_1 = s_{i_1}, \ldots, S_n = s_{i_n}) = P(S_{\times} = s_{i_1}) \ldots P(S_{\times} = s_{i_n}).$$

↪ stationary process

Memoryless source : S_i are i.i.d.

$$P(S_1 = s_{i_1}, \ldots, S_n = s_{i_n}) \overset{\text{independent}}{=} P(S_1 = s_{i_1}) P(S_2 = s_{i_2}) \cdots P(S_n = s_{i_n})$$

$$= P(S = s_{i_1}) P(S = s_{i_2}) \cdots P(S = s_{i_n})$$

identically distrib. — no time index

$\Rightarrow$ The source is stationnary.

In general :

$$P(S_1 = s_{i_1}, \ldots, S_n = s_{i_n}) = P(S_1 = s_{i_1}) P(S_2 = s_{i_2} \mid S_1 = s_{i_1})$$
$$\times P(S_3 = s_{i_3} \mid S_1 = s_{i_1}, S_2 = s_{i_2})$$
$$\times \ \ldots$$
$$\times P(S_n = s_{i_n} \mid S_1 = s_{i_1}, \ldots, S_{n-1} = s_{i_{n-1}})$$

If Markovian, then :

$$P(S_3 = s_{i_3} \mid S_1 = s_{i_1}, S_2 = s_{i_2}) = P(S_3 = s_{i_3} \mid S_2 = s_{i_2})$$

$\ldots$

$$P(S_n = s_{i_n} \mid S_1 = s_{i_1}, \ldots, S_{n-1} = s_{i_{n-1}}) = P(S_n = s_{i_n} \mid S_{n-1} = s_{i_{n-1}})$$

Discrete-time source

Markov source

Any source S emits symbols according to a distribution that can depend on all past symbols.

Definition 4 (Markov source)**.** *A source S is said to be Markovian if*

$$P(S_{n+1} = s_{i_{n+1}} | S_n = s_{i_n}, \ldots, S_1 = s_{i_1}) = P(S_{n+1} = s_{i_{n+1}} | S_n = s_{i_n})$$

for all $s_{i_1}, \ldots, s_{i_{n+1}}$ in $\mathcal{A}$.

As a direct consequence we have

$$P(S_1, \ldots, S_n) = P(S_1)\, P(S_2|S_1) \ldots P(S_n|S_{n-1})$$

Example : 

$P(S_1 = N)$

$P(S_2 = O \mid S_1 = N)$

$P(S_3 = T \mid S_2 = O)$

$P(S_1 = N, S_2 = O, S_3 = T) = P(S_1 = N) P(S_2 = O \mid S_1 = N) P(S_3 = T \mid S_2 = O)$

not time-invariant / stationnary

Discrete-time source

Markov source

Definition 5 (Time invariance). *A Markov source S is time-invariant if, for all $n \in \{1, 2, \ldots\}$, we have*

stationnary.

$$P(S_{n+1}|S_n) = P(S_2|S_1)$$

Such a source is entirely defined by the vector of initial probabilities $p|_{t=0}$ and the transition Π whose entries are

$$\Pi(i,j) = P(S_2 = s_j | S_1 = s_i)$$

i^{th} row, j^{th} column

Obviously, we have $\sum_{j=1}^{q} \Pi(i,j) = 1$ et $\Pi(i,j) \geq 0$.

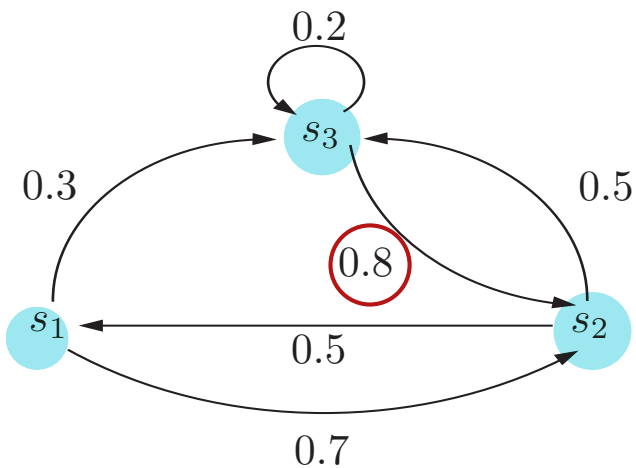
$P(S=N, S=O, S=T) = P(S=N)P(S=O|S=N)P(S=T|S=O)$

time-invariant / stationnary

Discrete-time source

Example of Markov source

Consider the following Markov source



$P(S=s_3 \mid S=s_1) = P(S=s_3 \mid S=s_2)$

The corresponding transition matrix can be written as:

$$\Pi(i,j) = P(S_2 = s_j | S_1 = s_i)$$

$$\Pi = \begin{pmatrix} 0 & 0.7 & 0.3 \\ 0.5 & 0 & 0.5 \\ 0 & 0.8 & 0.2 \end{pmatrix}$$

$i = 3$

$P(S_2 = s_2 \mid S_1 = s_3)$

$2 = j$

I need : Π and $\begin{matrix} P(S=s_1) \\ P(S=s_2) \\ P(S=s_3) \end{matrix} \Big| \; \Pi_{t=0}$

Π ↑ transition matrix

$\Pi_{t=0}$ ↑ initial distribution

<u>example</u> :

$P(S=s_1) = 0,2$

$P(S=s_2) = 0,3$

$P(S=s_3) = 0,5$

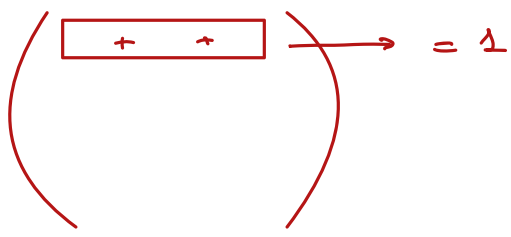
$$P(S=s_1, S=s_3, S=s_3) = P(S=s_1)\,P(S=s_3 \mid S=s_1)\,P(S=s_3 \mid S=s_3)$$

$$= 0,2 \times 0,3 \times 0,2$$

$$= 0,04 \times 0,3$$

$$= 0,012$$

<u>Remark</u> :

Π = 

because $\sum_i P(S=s_i \mid S=s_j) = 1$ (j fixed)

$$\boxed{\sum_i \Pi(j,i) = 1}$$

$$\Pi = \begin{pmatrix} 0 & 0.7 & 0.3 \\ 0.5 & 0 & 0.5 \\ 0 & 0.8 & 0.2 \end{pmatrix}$$

calculate $P_{t=m}$ given $P_{t=m-1}$

$$P(S_m = s_1) = \sum_{j=1,2,3} P(S_m = s_1, S_{m-1} = s_j)$$

$$= P(S_m = s_1 \mid S_{m-1} = s_1) P(S_{m-1} = s_1)$$
$$+ P(S_m = s_1 \mid S_{m-1} = s_2) P(S_{m-1} = s_2)$$
$$+ P(S_m = s_1 \mid S_{m-1} = s_3) P(S_{m-1} = s_3)$$

$$P_{t=m-1} = \left[P(S_{m-1} = s_1) \; P(S_{m-1} = s_2) \; P(S_{m-1} = s_3) \right]$$

$$P(S_m = s_1) = P_{t=m-1} \, \Pi(:, 1)$$

$$\boxed{P_{t=m} = P_{t=m-1} \, \Pi}$$

row-wise

$$P(S_0 = s_1) = 0,2$$
$$P(S_0 = s_2) = 0,3$$
$$P(S_0 = s_3) = 0,5$$

$$\Pi = \begin{pmatrix} 0 & 0.7 & 0.3 \\ 0.5 & 0 & 0.5 \\ 0 & 0.8 & 0.2 \end{pmatrix}$$

$$P_{t=0} = \left[0,2 \quad 0,3 \quad 0,5 \right]$$

$$P_{t=1} = \left[0,2 \quad 0,3 \quad 0,5 \right] \begin{pmatrix} 0 & 0,7 & 0,3 \\ 0,5 & 0 & 0,5 \\ 0 & 0,8 & 0,2 \end{pmatrix}$$

$$= \left[0,15 \quad 0,54 \quad 0,31 \right]$$

$P(S_1 = s_1)$ $P(S_1 = s_2)$ $P(S_2 = s_3)$

Discrete-time source

Markov source in steady state

Definition 6 (steady-state)**.** *Consider a Markov source S. If it exists, its steady state distribution is defined as:*

$$\lim_{n\to\infty} P(S_n = s_i)$$

for all $i \in \{1, \ldots, Q\}$.

it means : $p_{t=n} = p_{t=n-1}$ for $n \to \infty$

Let $p|_{t\to\infty}$ the steady-state distribution if it exists. Given that $p|_{t=n} = p|_{t=n-1}\,\Pi$, we have:

$$p|_{t\to\infty} = p|_{t\to\infty}\,\Pi$$

We say that $p|_{t\to\infty}$ is the steady-state distribution of S since initializing it with $p|_{t\to\infty}$ makes it stationary.

Drawback. The steady state defined in this way depends on the initial distribution $p|_{t=0}$. Other definitions exist.

Calculate the steady state distribution of :

$$\Pi = \begin{pmatrix} 0 & 0.7 & 0.3 \\ 0.5 & 0 & 0.5 \\ 0 & 0.8 & 0.2 \end{pmatrix}$$

$$P_\infty \Pi = P_\infty$$

$$P_\infty = [\, x \quad y \quad z \,] \qquad \text{⚠} \quad x + y + z = 1$$

$$\begin{cases} \text{①} & x = 0{,}5\, y \\ \text{②} & y = 0{,}7\, x + 0{,}8\, z \\ \text{③} & z = 0{,}3\, x + 0{,}5\, y + 0{,}2\, z \\ \text{④} & x + y + z = 1 \end{cases}$$

$$\begin{cases} \text{①} & 2x = y \\ \text{④} & x + y + z = 1 \Rightarrow z = 1 - x - y \\ \text{②} & 10\, y = 7x + 8z \end{cases}$$

with ② : $20\, x = 7x + 8z \Rightarrow z = \frac{13}{8} x$

$$x + 2x + \frac{13}{8} x = 1$$

$$\Rightarrow \quad 37\, x = 8 \quad \Rightarrow \quad x = \frac{8}{37}$$

$$y = \frac{16}{37}$$

$$z = \frac{13}{37}$$

$$P(S_\infty = s_1) = \frac{8}{37}$$

$$P(S_\infty = s_2) = \frac{16}{37}$$

$$P(S_\infty = s_3) = \frac{13}{37}$$

steady-state (stationary) distribution

Checking: $z = 0{,}3x + 0{,}5y + 0{,}2z$

$\hookrightarrow 8z = 3x + 5y$

$$3x + 5y = \frac{24}{37} + \frac{80}{37} = \frac{104}{37}$$

$$8z = \frac{104}{37}$$

$\rightarrow$ OK!

Discrete-time source

Entropy of a stationary source

Any source S emits symbols according to a law that can depend on the symbols that came before them. The definition of the entropy of S must take this into account.

Definition 7 (Entropy of a stationary source)**.** *The entropy of a stationary S source is defined as:*

$$H_0 \triangleq \lim_{n \to +\infty} H(S_n | S_1, \ldots, S_{n-1}).$$

This definition only makes sense if the limit exists.

Discrete-time source

Entropy of a stationary source

Validation of the definition. One need to check that the following limit exists:

$$\lim_{n \to +\infty} H(S_n | S_1, \ldots, S_{n-1})$$

We have:

$$0 \leq H(S_n | S_1, S_2, \ldots, S_{n-1}) \leq H(S_n | S_2, \ldots, S_{n-1}) \leq \ldots \leq H(S_n).$$

Since S is stationary, we can write:

$$H(S_n) = H(S_1) \qquad H(S_n | S_{n-1}) = H(S_2 | S_1) \qquad \ldots$$

The above inequality can be replaced by:

$$0 \leq H(S_n | S_1, \ldots, S_{n-1}) \leq H(S_{n-1} | S_1, \ldots, S_{n-2}) \leq \ldots \leq H(S_1).$$

The series $\{H(S_n | S_1, \ldots, S_{n-1})\}_{n \geq 1}$ is decreasing and bounded. It is therefore convergent, ensuring the validity of the definition in the stationary case.

Discrete-time source

Entropy of a stationary source

Example 1. In the case of a memoryless source, characterized by independent and identically distributed S_i, we have:

$$H_0 = H(S_1).$$

Example 2. If S denotes a time-invariant Markov source, its entropy is given by:

$$H_0 = H(S_2|S_1).$$

Definition 7 (Entropy of a stationary source)**.** *The entropy of a stationary S source is defined as:*

$$H_0 \triangleq \lim_{n \to +\infty} H(S_n | S_1, \ldots, S_{n-1}).$$

case 1 : Memoriless source , i.e., S_i are i.i.d.

$$H(S_n | S_1, \ldots, S_{n-1}) = H(S_n)$$

As S is stationnary $H(S_n) = H(S)$

$$H_0 \triangleq \lim_{n \to \infty} H(S_n | S_1, \ldots, S_{n-1})$$

$$= H(S)$$

case 2 : Markov source , stationnary state

$$H(S_n | S_1, \ldots, S_{n-1}) = f(\mathbb{P}(S_n | S_1, \ldots, S_{n-1}))$$

$$\overset{\text{Markov}}{=} f(\mathbb{P}(S_n | S_{n-1}))$$

$$= H(S_n | S_{n-1})$$

$$\overset{\text{stationnary}}{=} H(S_2 | S_1)$$

$$H_0 \triangleq \lim_{n \to \infty} H(S_n | S_1, \ldots, S_{n-1})$$

$$= H(S_2 | S_1)$$

UNIVERSITÉ CÔTE D'AZUR

Cédric Richard

Discrete source coding

Exercise 1

Indicate for each of the following codes whether it is regular, decodable, instantaneous and complete: $\mathcal{C}_1 = \{00, 01, 10, 11\}$, $\mathcal{C}_2 = \{0, 01, 11\}$, $\mathcal{C}_3 = \{0, 10, 11\}$, $\mathcal{C}_4 = \{0, 11, 111\}$.

Exercise 2

We consider a source S that can emit 5 symbols, each of which has a probability p_i in the table below. This table also provides two possible binary codings $\mathcal{C}_1$ and $\mathcal{C}_2$ of S. Indicate whether these codes are decodable and instantaneous. Calculate the average $\overline{n}_1$ and $\overline{n}_2$ lengths of their codewords. Compare to the minimum average codewords length $\overline{n}_{\min}$ required for S.

s_i	s_1	s_2	s_3	s_4	s_5
p_i	0.50	0.18	0.14	0.12	0.06
$\mathcal{C}_1$	0	10	11	101	1001
$\mathcal{C}_2$	00	10	11	010	011

Exercise 3

We consider a random variable X that can take n values distributed according to the following distribution: $P(X = x_i) = (1/2)^i$ for $1, 2, \ldots, n-1$, and $P(X = x_n) = (1/2)^{n-1}$. Determine the minimum average length $\bar{n}_{\min}(X)$. Propose a binary code using Huffman's method. Calculate the average length of its codewords. Discuss.

Exercise 4

A printer uses the following commands:

Raise the stylus (RS)
Press the stylus (PS)
move the stylus left (–X)
move the stylus right (+X)
move the stylus up (+Y)
move the stylus down (–Y).

Calculate the minimum average number of bits required for this set of commands if their probabilities are given by:

$$P_{\text{RS}} = P_{\text{PS}} = P_{\text{–X}} = 0.1 \qquad P_{\text{+X}} = 0.3 \qquad P_{\text{+Y}} = P_{\text{–Y}} = 0.2$$

Build a Shannon's binary code. Build a Huffman's binary code. Compare the two solutions.

Exercise 5

A high school has to communicate a list of undergraduate results for 2500 students. These results are as follows: 250 A, 375 B, 1125 C, 625 failed, 125 absent. Build a binary Huffman's code to compress the corresponding file. Calculate the average length of the codewords. Calculate the file size if the information are encoded using a fixed-length code with 8 bits. Evaluate the gain in file size achieved by using the Huffman's code.

Problem 1

We consider a code consisting of two words of length 2, two words of length 3 and one word of length 4.

1. Show that it exists a decodable binary code respecting these codeword lengths. Draw a possible code tree. Modify this tree in order the reduce the average codeword length.

2. We assign the following probabilities $\{0.50, 0.18, 0.14, 0.12, 0.06\}$ to the 5 states of the source. Associate these probabilities with the codewords proposed previously so as to minimize the average length of the codewords. Calculate the average codewords length and show that there exist binary codes with better performance.

3. Propose a binary code using Huffman's method. Compare the average length of its codewords to the one obtained in the previous question.

Problem 2

Consider a Markov source where $p = \frac{1}{10}$ and $q = \frac{2}{10}$.

$$
\begin{aligned}
P(S_n = 0 | S_{n-1} = 1) &= p \\
P(S_n = 1 | S_{n-1} = 1) &= 1 - p \\
P(S_n = 1 | S_{n-1} = 0) &= q \\
P(S_n = 0 | S_{n-1} = 0) &= 1 - q.
\end{aligned}
$$

1. Determine the stationary distribution of the source. Calculate the entropy of the source without taking the dependency of the states into account. Calculate in this case the minimum average length of binary codewords required to encode this source.

2. Calculate the entropy of the Markov source, which assumes that the dependency of successive states is taken into account. Calculate in this case the minimum average length of binary codewords to encode this Markov source.

3. Consider the extension of order 2 for S. Calculate its entropy. Calculate the minimum average length of binary codewords to encode this source. Propose a Huffman's binary code and calculate the average length of its codewords.

Exercise 4

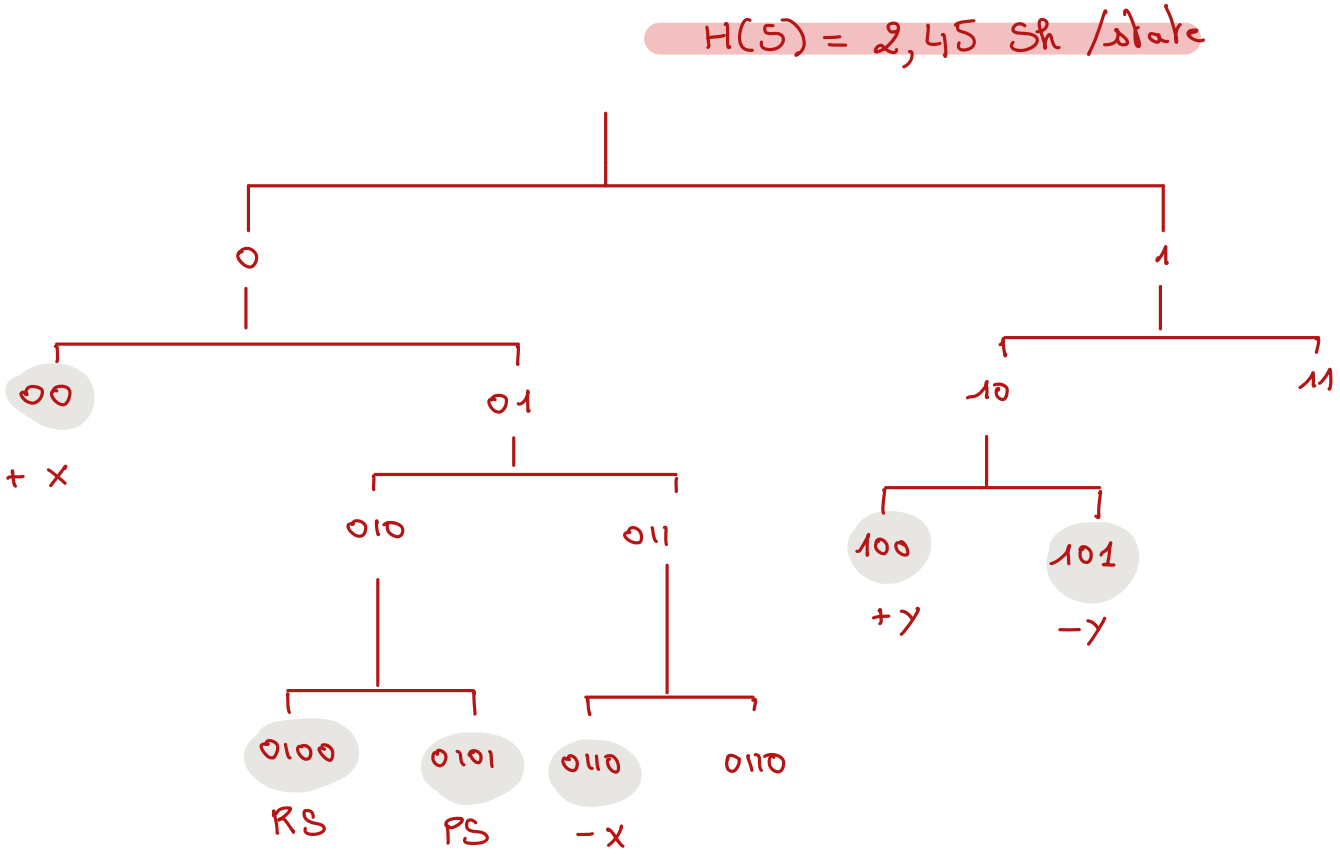
A printer uses the following commands:

Raise the stylus (RS)
Press the stylus (PS)
move the stylus left (–X)
move the stylus right (+X)
move the stylus up (+Y)
move the stylus down (–Y).

Calculate the minimum average number of bits required for this set of commands if their probabilities are given by:

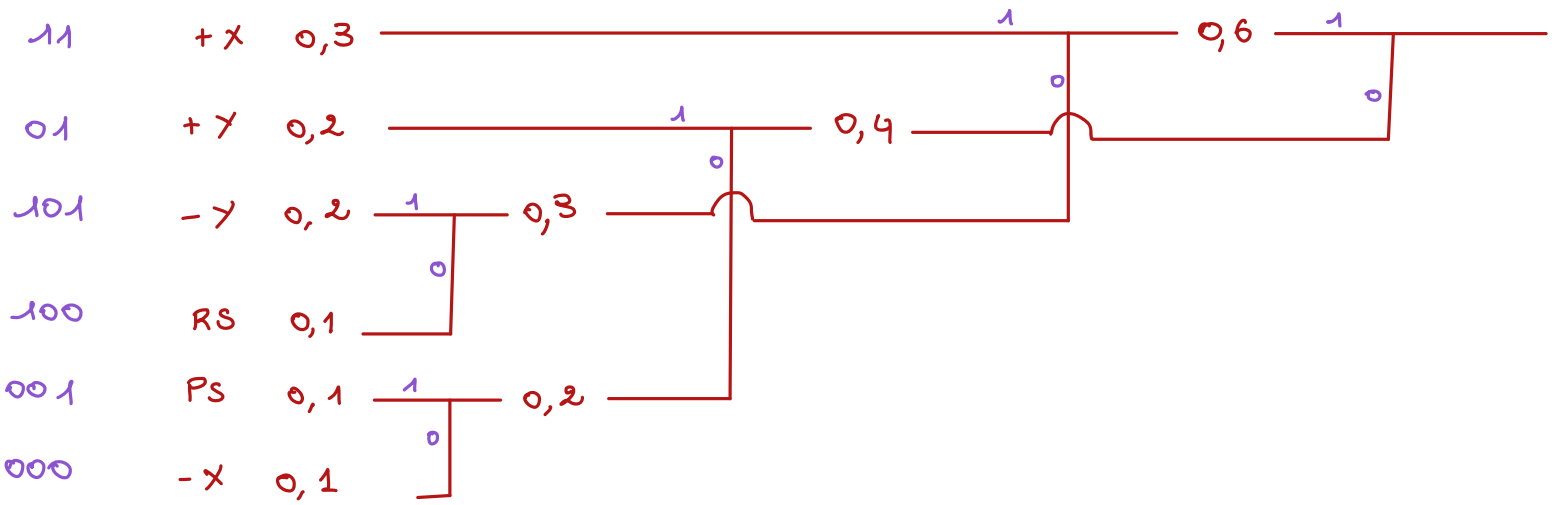
$$P_{RS} = P_{PS} = P_{-X} = 0.1 \qquad P_{+X} = 0.3 \qquad P_{+Y} = P_{-Y} = 0.2$$

Build a Shannon's binary code. Build a Huffman's binary code. Compare the two solutions.



$$\overline{n}_{SH} = (4 \times 0{,}3) + (2 \times 0{,}3) + (3 \times 0{,}4)$$
$$= 1{,}2 + 0{,}6 + 1{,}2$$
$$= 3 \text{ bit / state}$$

$P_{RS} = P_{PS} = P_{-X} = 0.1 \qquad P_{+X} = 0.3 \qquad P_{+Y} = P_{-Y} = 0.2$



$\overline{n}_{Huf} = (2 \times 0{,}5) + (3 \times 0{,}5) = 2{,}5$ bit / state

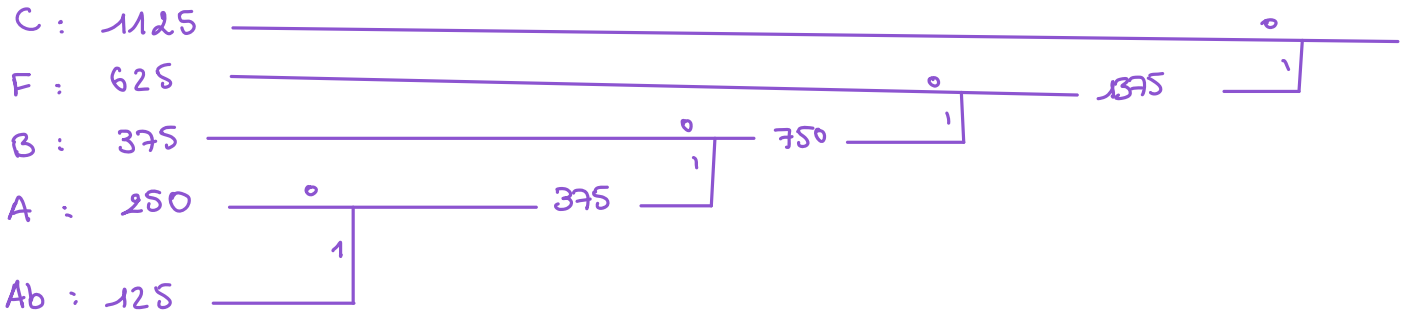
Shannon – Fano

00	+X	0,3	0	0	
01	+Y	0,2	0	1	
100	-Y	0,2	1	0	0
101	RS	0,1	1	0	1
110	PS	0,1	1	1	0
111	-X	0,1	1	1	1

$\overline{n}_{Shannon-Fano} = 2{,}5$ bit / state

Exercise 5

A high school has to communicate a list of undergraduate results for 2500 students. These results are as follows: 250 A, 375 B, 1125 C, 625 failed, 125 absent. Build a binary Huffman's code to compress the corresponding file. Calculate the average length of the codewords. Calculate the file size if the information are encoded using a fixed-length code with 8 bits. Evaluate the gain in file size achieved by using the Huffman's code.



C : 0 F : 10 B : 110 A : 1110 Abo : 1111

$$\bar{n} = \frac{1125 + 1250 + 1125 + 1000 + 500}{2500}$$

$$\bar{n}_{Huf} = 2 \text{ bit / state.}$$

$$File\_Size = 8 \times 2500 = 20\,000 \text{ bits}$$

2 : 8 = 1 : 4 compression ratio

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?

$$P(M=0 \mid T=0) = \frac{P(M=0, T=0)}{P(T=0)}$$

$P(M=i, T=j)$	sunny weather ($T=1$)	rainy weather ($T=0$)	$P(M)$
sunny weather ($M=1$)	5/8	1/16	11/16
rainy weather ($M=0$)	3/16	1/8	5/16
$P(T)$	13/16	3/16	

$$P(T=0) = P(T=0, M=0) + P(T=0, M=1)$$

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

$$P(T=0) = \frac{3}{16} \qquad P(T=1) = \frac{13}{16}$$

$$P(M=0) = \frac{5}{16} \qquad P(M=1) = \frac{11}{16}$$

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

$$\text{Error} = (M=0 \text{ and } T=1) \text{ or } (M=1 \text{ and } T=0)$$

$$P_{error}(M) = P(M=0, T=1) + P(M=1, T=0)$$

$$= \frac{3}{16} + \frac{1}{16}$$

$$= \frac{1}{4}$$

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

$$\text{Error} = (E=0 \text{ and } T=1) \text{ or } (E=1 \text{ and } T=0)$$

$$P_{error}(E) = P(E=1 \text{ and } T=0)$$

$$= P(T=0)$$

$$= \frac{3}{16} < P_{error}(M)$$

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 $I(E;T)$.

$$I(E,T) = H(E) - H(E|T)$$

$H(E) = 0$ because $P(E=0) = 0$, $P(E=1) = 1$

$$0 \leq H(E|T) \leq \underbrace{H(E)}_{\parallel 0} \Rightarrow H(E|T) = 0$$

So: $\boxed{I(E,T) = 0 \text{ Sh}}$

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

$$I(M,T) = H(M) - H(M|T)$$

$$= H(M) - \left[H(M|T=0)P(T=0) + H(M|T=1)P(T=1)\right]$$

$$= H\left(\frac{5}{16}, \frac{11}{16}\right) - \left[H\left(\frac{2}{3}, \frac{1}{3}\right) \times \frac{3}{16} + H\left(\frac{3}{13}, \frac{10}{13}\right) \times \frac{13}{16}\right]$$

$$\simeq 0,09 \text{ Sh}$$

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

We have $I(E,T) = 0$ and $I(M,T) > 0$

So, there is information on weather in meteorologist's forecast while there is no in student's forecast.

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 $\overline{n}$?

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