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.

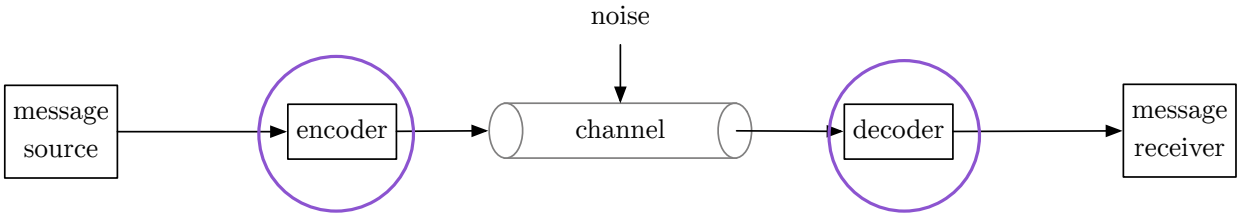
Following the basic concept, communication is the process of sending and receiving messages or transferring information from one part (sender) to another (receiver).

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

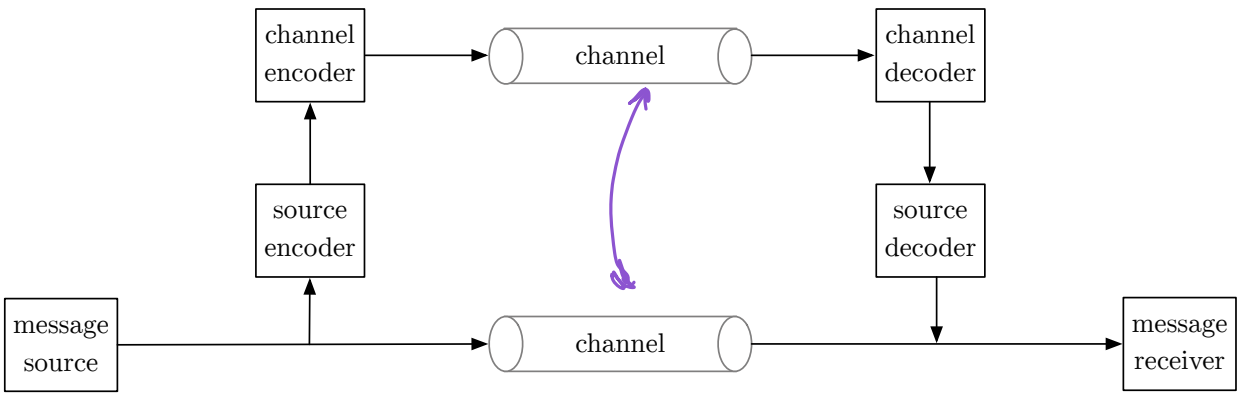
Shannon's communication model



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:

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

A, B independant

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

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

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

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

$\longrightarrow$ $h(A) = \log\left(\frac{1}{P(A)}\right)$

$f(p) = \log(p)$

A, B : two events

$\longrightarrow$ $P(A,B) = P(A)P(B)$ if A, B are independent

$$h(A,B) = f\left(\frac{1}{P(A,B)}\right) = f\left(\frac{1}{P(A)} \cdot \frac{1}{P(B)}\right)$$ because A, B indep

($\frac{1}{P(A)}$: P_1, $\frac{1}{P(B)}$: P_2)

I want

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

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

I want : $f(P_1 \cdot P_2) = f(P_1) + f(P_2)$

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

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

$\log_2$, $\log_{10}$, $\log_e = \ln$?

$\log_2$: Sh

$\log_{10}$: decit

$\log_e$: nat

$$\log_b \cdot = \frac{\ln \cdot}{\ln b} \quad , \quad b > 1$$

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. $f(p^n) = n f(p)$? $n \in \mathbb{N}$

$$f(p_1 \cdot p_2) = f(p_1) + f(p_2)$$

$$f(p^n) = f(p \cdot p^{n-1}) = f(p) + \underbrace{f(p^{n-1})}_{f(p) + f(p^{n-2})}$$

$$\ldots$$

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

$$= n f(p)$$

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

$$p = (p^{1/n})^n \quad , \quad n \in \mathbb{N}$$

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

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

$$\Rightarrow f(p^{1/n}) = \frac{1}{n} f(p) \quad , \quad n \in \mathbb{N}$$

3. $f\left(p^{m/n}\right) = \frac{m}{n} f(p)$ by combining the two previous equalities

$f(p^{m/n}) = ?$ with m and n integers

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

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

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

4. $f(p^q) = q\, f(p)$ where q is any positive rational number

trivial with 3. because $q = \frac{m}{n}$

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

$= q\, f(p)$

5. $f(p^r) = \lim_{n\to+\infty} f(p^{q_n}) = \lim_{n\to+\infty} q_n\, f(p) = r\, f(p)$ because rationals are dense in the reals

$\mathbb{Q}$ is dense in $\mathbb{R}$

$f(p^{q_n}) \overset{4.}{=} q_n f(p)$

$\underbrace{\lim_{n\to+\infty} f(p^{q_n})}_{f(p^x)} = \underbrace{\lim_{n\to\infty} q_n f(p)}_{x f(p)}$

$f(p^x) = x\, f(p)$

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

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

$0 < p, q < 1$

$$f(p) = f(q) \log_q p$$

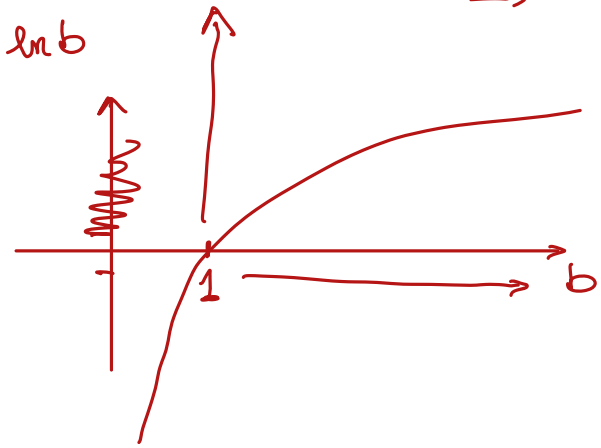
$$= f(q)\, \frac{\ln p}{\ln q}$$

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

$$\Rightarrow \quad f(p) = C \ln p \,, \quad C \in \mathbb{R}^-_*$$

as $f(p) \geq 0$, $C < 0$ because $\ln p < 0$ $\quad 0 < p \leq 1$

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



$$C = -\frac{1}{\ln b} \,, \quad b > 1$$

$$f(p) = -\log_b p \,, \quad b > 1$$

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:* (self info of A)

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

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)

$-\log_2(\cdot)$ → Shannon

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

$h(A)$

$P(A)$ small → A is very uncertain

$h(A)$ very high

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

$$= -\frac{\ln P(A)}{\ln 2}$$

$$h(A)\ [\text{nat}] = \underbrace{h(A)}_{[\text{Sh}]}\ \ln 2$$

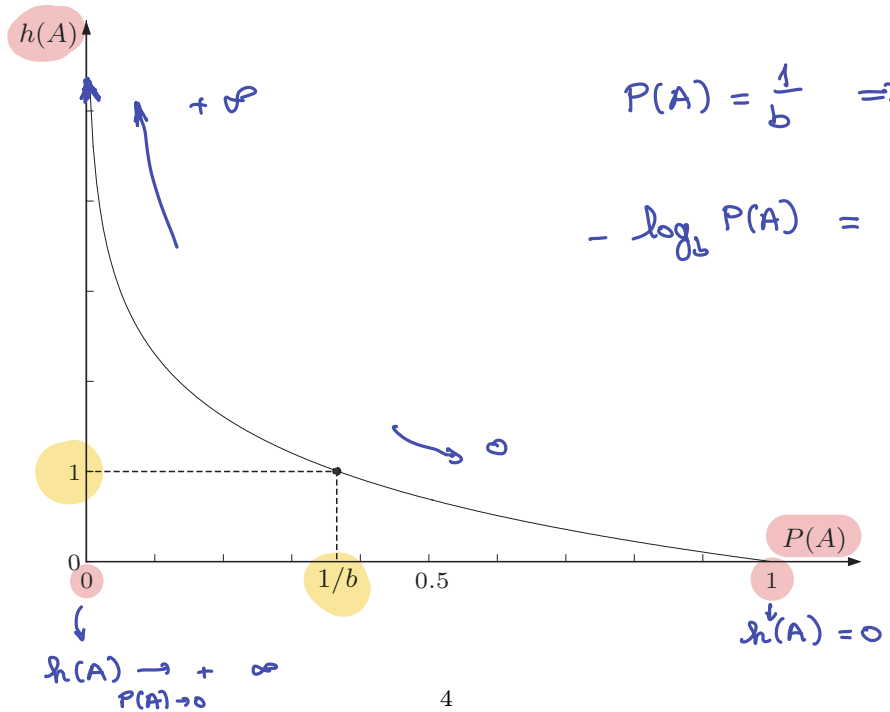
$$-\log_b P(A) = \frac{-\log_{b'} P(A)}{\ln b} \times \ln b'$$

$$h_{\text{Sh}}(A) = \frac{h_{\text{bit}}(A)}{\ln 2} \times \ln 10$$

Self-information

Information content

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



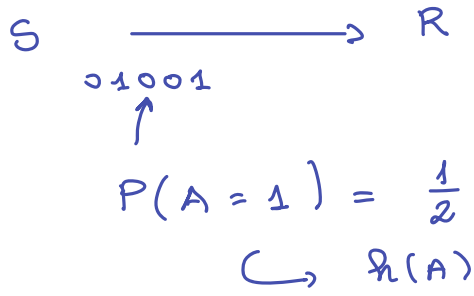
$P(A) = \frac{1}{b} \Rightarrow h(A) = 1$

$-\log_b P(A) = \frac{-\ln\left(\frac{1}{b}\right)}{\ln b} = \frac{\ln b}{\ln b} = 1$

$h(A)$ in Shannon

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

if $P(A) = \frac{1}{2} \Rightarrow h(A) = 1$ Sh



$P(A = 1) = \frac{1}{2}$

↳ $h(A) = 1$ Sh

$$h(A) = \frac{-\ln P(A)}{\ln 2}$$

SELF-INFORMATION

Information content

$P(A) = \frac{1}{2} \Rightarrow \ln P(A) = -\ln 2$

$\Rightarrow h(A) = \frac{\ln 2}{\ln 2} = 1$ Sh

bits : binary digit

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.

↳ binary unit

$h\left(\frac{1}{2}\right) = 1$ 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.

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

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

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.

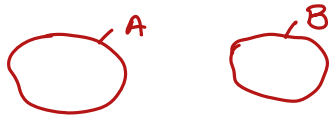
Definition 2. *Conditional information content of B given A is defined as:*

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

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

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

$A \cap B = \emptyset$: A and B independent

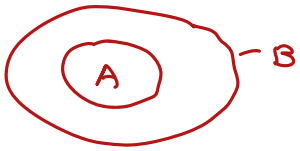


$$h(A,B) = h(A) + h(B)$$

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

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

$A \subset B$:



$A \Rightarrow B$

$P(B|A) = 1$

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

with $h(B|A) = -\log_2 \underbrace{P(B|A)}_{1}$

$= 0$ Sh

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

$B \subset A$: $B \Rightarrow A$, $P(A|B) = 1$

$h(A,B) = h(B)$ because $h(A|B) = 0$ Sh

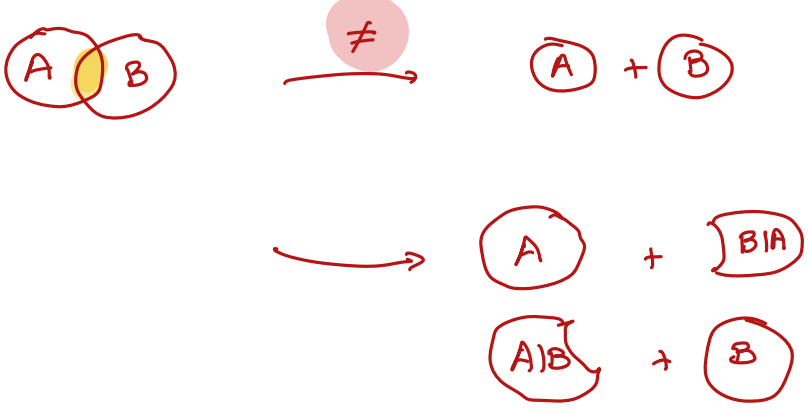
A, B two events

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

Bayes relationship : $P(B|A) = \frac{P(B)P(A|B)}{P(A)}$

$$h(A,B) = -\log_2 P(A,B)$$
$$= -\log_2 [P(A)P(B|A)] \quad // \quad -\log_2[P(B)P(A|B)]$$
$$= \underbrace{-\log_2 P(A)}_{h(A)} \underbrace{-\log_2 P(B|A)}_{h(B|A)}$$
$$= h(A) + h(B|A) \quad // \quad h(B)+h(A|B)$$

$$h(A,B) = h(A) + h(B|A) \rightarrow \text{quantity of info. provided by B given A}$$
$$= h(B) + h(A|B) \rightarrow \text{" provided by A given B}$$



If A and B indep.

then $P(A,B) = P(A)P(B)$ because $P(B|A) = P(B)$ and $P(A|B) = P(A)$

$$h(A,B) \overset{\text{indep.}}{=} -\log_2 \left(P(A) \times P(B) \right)$$

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

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

$$\Rightarrow \quad h(A|B) = h(A)$$

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

$\rightarrow$ if A and B indep.

1 15 : back

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

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

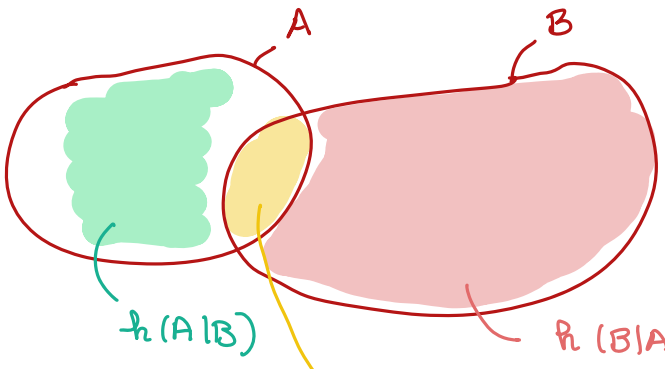
$$h(A,B) = h(A) + h(B|A) \quad \text{Sh}$$

$h(B|A)$ → quantity of info provided by B given A, i.e., when A known

---

if A and B independent:

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



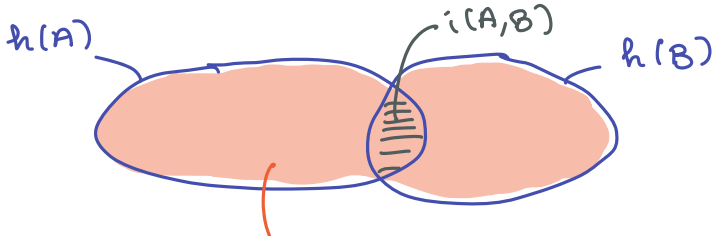
$i(A,B)$: information quantity shared by A and B

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

Remember that: $h(A,B) = h(A) + h(B|A)$

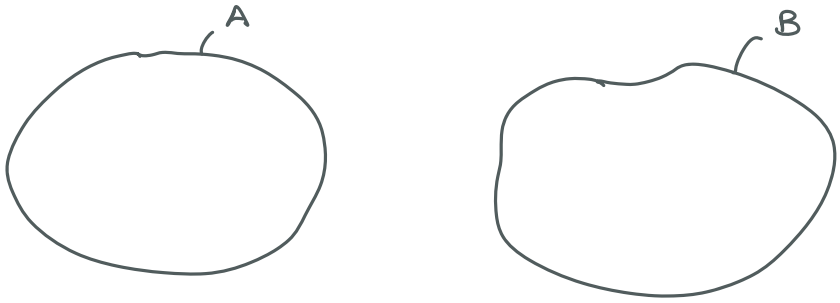
$$\Rightarrow h(B|A) = h(A,B) - h(A)$$

$$\Rightarrow \boxed{i(A,B) = h(A) + h(B) - h(A,B)}$$



h(A,B)

A and B independent :



$h(A|B) = h(A)$ $\qquad$ $h(B|A) = h(B)$

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

$i(A,B) = 0$

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

$= h(A) - h(A)$ because A, B indep

$= 0$ Sh

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

$h(A,B) = h(A)$ $\qquad$ and $\qquad h(B|A) = 0$

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

$= h(B)$

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:*

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

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

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

S ——— t, h, e, ␣, h, o, u, s, e, ␣, ...

S' ——— t, t, t, ... $P(S'=t)=1$, $P(S'\neq t)=0$ ⟶ $H(S')=0$

$\mathcal{A} = \{a, b, c, \ldots, z, 0, 1, \ldots, 9\}$

$P(S=a) \longrightarrow h(S=a) = -\log_2 P(S=a)$

$P(S=b) \longrightarrow h(S=b) = -\log_2 P(S=b)$

....

$P(S=z)$

$P(S=0)$

...

$P(S=9) \longrightarrow h(S=9) = -\log_2 P(S=9)$

$H(S) = E\{h(S)\}$

$\longrightarrow = P(S=a)\,h(S=a) + \ldots + P(S=9)\,h(S=9)$

= mean quantity of info delivered by the sour

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

[Sh / symbol
or character
or state of the source]
or ...

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:*

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

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

$P(S = s_i) = p_i \quad , \quad i = 1, \dots, n$

$$H(S) = \sum_{i=1}^{n} P(S = s_i)\, h(S = s_i) \quad \rightsquigarrow \text{expectation}$$

$$= \sum_{i=1}^{n} p_i\, h(S = s_i) \quad \text{with } h(S = s_i) = -\log_2 p_i$$

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

$$= -\sum_{i=1}^{n} p_i \log_2 p_i \quad \text{Sh / symb}$$

→ entropy of S

$S \in \{0, 1\}$: binary source

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

→ $H(S) = -p \log_2 p - (1-p) \log_2 (1-p)$ Sh/symb.

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

Let us plot $H(S)$ as a function of p.

$H(S) = -p \log_2 p - (1-p) \log_2 (1-p)$ Sh/symb.

Self-information : 1 event (A)

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

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

$$P(S = s_i) \triangleq p_i$$

$$H(S) = -\sum_{i=1}^{m} p_i \log p_i \quad \left(= \mathbb{E}\{h(S)\} \right)$$

Sh / event

2 events : $S = \{0, 1\}$

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

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

Extreme cases :

1. $p = 0$: $P(S=1) = 1$ and $P(S=0) = 0$

↳ S is stuck in state 1

S — 1 1 1 1 1 1 ....

$H(S) = 0$ Sh / state

because :

$$H(S) = \underbrace{-p \log_2 p}_{\substack{\| \\ 0 \\ \text{because} \\ \lim_{x \to 0} x \log x = 0}} \underbrace{- (1-p) \log_2 (1-p)}_{\substack{\| \\ 0 \text{ if } p=0}}$$

2. $p = 1$: $P(S=1) = 0$ and $P(S=0) = 1$

S — 0 0 0 0 0 ....

Ayun : $H(S) = 0$ Sh /state

3. $p = \frac{1}{2}$: $P(S=0) = P(S=1) = \frac{1}{2}$

$$H(S) = 2 \times \left(-\frac{1}{2} \log_2 \left(\frac{1}{2} \right) \right)$$

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

↳

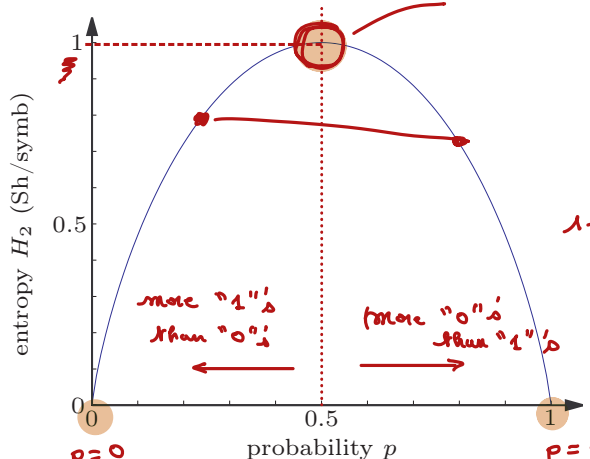
Entropy of a random variable

Example of a binary random variable

The entropy of a binary random variable is given by:

$$H(\mathcal{S}) = -p \log p - (1-p) \log(1-p) \triangleq H_2(p).$$

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



$P(S=1) = P(S=0) = \frac{1}{2}$

$H(S) = 1$ Sh/state.

$p = P(S=0)$

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

$p=0$ → S generates only "1"s

$p=1$ → S generates only "0"s

$H(S) = 0$

$$H_{max}(S) = \log_2 n \quad \text{Sh / state}$$

where n is the number of states of S

$\underline{n = 2}$ $\quad H_{max}(S) = \log_2 2 \quad \text{Sh / state}$

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

$\longrightarrow$ reached when $p_1 = \dots = p_n = \frac{1}{n}$

$\{p_1, \dots, p_n\}$: uniform distribution

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:*

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

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

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:

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

Gibbs

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

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

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

$$H_n(p_1, \ldots, p_n) \overset{?}{\leq} \log n$$

where $H = -\sum_{i=1}^{n} p_i \log_2 p_i$

you observe that if $p_i = \frac{1}{n}$ for all i

$\Rightarrow$ $H_{max} = \log_2 n$

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

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

$$1 - 1 = 0$$

$$\Rightarrow \boxed{\sum_{i=1}^{n} p_i \ln \frac{q_i}{p_i} \leq 0}$$

Gibbs inequality

$q_i = p_i \Rightarrow$ $= 0$

$\frac{1}{\ln b}$

$b > 1$

$$\Rightarrow \boxed{\sum_{i=1}^{n} p_i \log_b \frac{q_i}{p_i} \leq 0}$$

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

with Gibbs inequality: $\sum_{i=1}^{n} p_i \log_b \left(\frac{1}{n p_i} \right) \leq 0$

$$\Rightarrow - \sum_{i=1}^{n} p_i \log_b p_i - \underbrace{\sum_{i=1}^{n} p_i}_{1} \log_b n \leq 0$$

$\log_b n$

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

$$H(S) \leq \log_b n$$

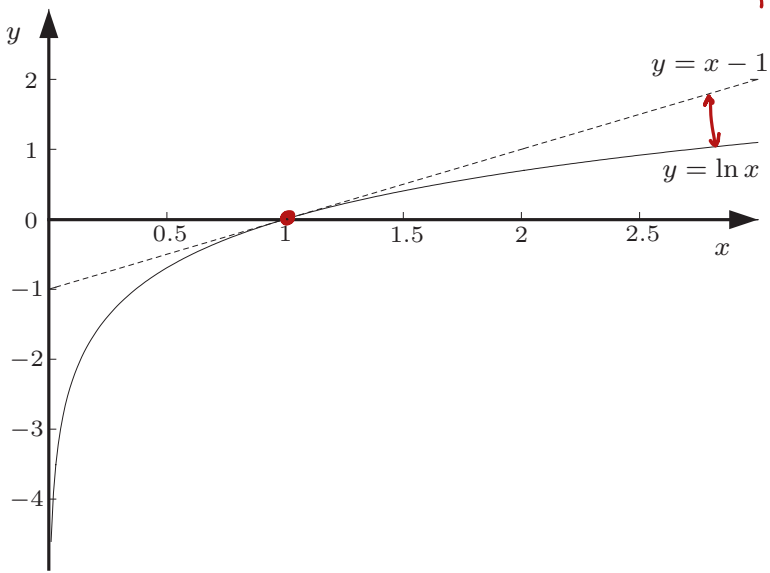
b : basis of $H(S)$ and $\log_b$

b : basis of $H(S)$ and $\log_b$

Entropy of a random variable

Notation and preliminary properties

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



analyze :
$f(x) = x - 1 - \ln x$
and show that
$f(x) \geq 0$
$\forall x > 0$

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:

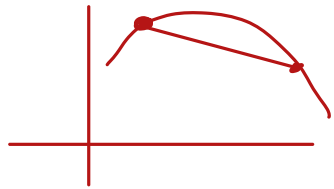
$$\begin{aligned} H(X') &= H(X) + p_k \log p_k - p_{k_1} \log p_{k_1} - p_{k_2} \log p_{k_2} \\ &= H(X) + p_{k_1}(\log p_k - \log p_{k_1}) + p_{k_2}(\log p_k - \log p_{k_2}). \end{aligned}$$

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.

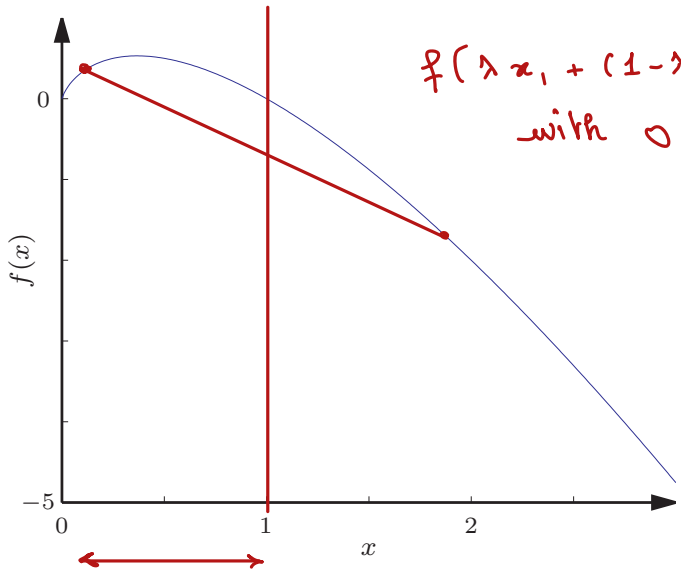
$H(S) = \sum_{i=1}^{n} f(p_i)$ with $p_i = P(S = s_i)$, $i = 1, \ldots, n$

↳ f concave ? $\quad 0 < p_i \leq 1$

Entropy of a random variable

Properties

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



$f(\lambda x_1 + (1-\lambda) x_2) \geq \lambda f(x_1) + (1-\lambda) f(x_2)$

with $0 < x_1, x_2 \leq 1$

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:*

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

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

Example. Case of two independent random variables

$$\hookrightarrow\ P(X=x_i, Y=y_j) = P(X=x_i)\,P(Y=y_j)$$

$$\Longrightarrow\ \boxed{H(X,Y) = H(X) + H(Y)}$$

X, Y two random variables

with $P(X = x_i, Y = y_j)$ for all i, j

$i \in \{1, \dots, n\}$

$j \in \{1, \dots, m\}$

$$Z = (X, Y)$$

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

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

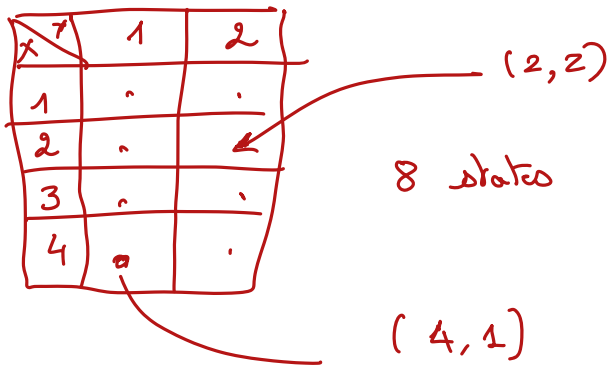
Sh / state of (X,Y)

or Sh / pair of states.

Rm : $H_{max}(X,Y) = \log_2 (n\,m)$

reached when $P(X=x_i, Y=y_j) = \frac{1}{n\,m}$

$\forall i, j$



8 states

$H_{max} = \log_2 8$

$= 3$ Sh / pair

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

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

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

$H(X|Y=y_{j_0})$: r.v. X; $Y=y_{j_0}$ fixed

$\sum_{i=1}^{n}$: Sh / state of X

X and Y independent: $P(X=x_i, Y=y_j) = P(X=x_i)P(Y=y_j)$

$\Rightarrow P(X=x_i | Y=y_j) = P(X=x_i)$

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

if A, B indep., then $P(A,B) = P(A)P(B)$

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

definition 6:

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

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

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

definition 7:

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

if X and Y indep.: $H(X|Y=y_j) = H(X)$

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

$\sum_{j=1}^{m} P(Y=y_j) = 1$

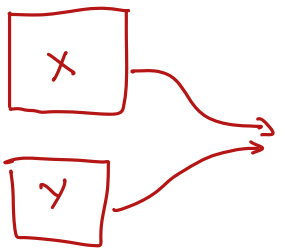
$= H(X)$

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

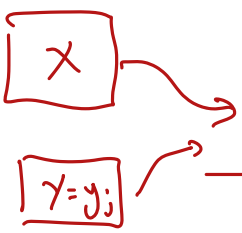
quantity of info. provided by X

knowing that $Y = y_j$

↳ fixed



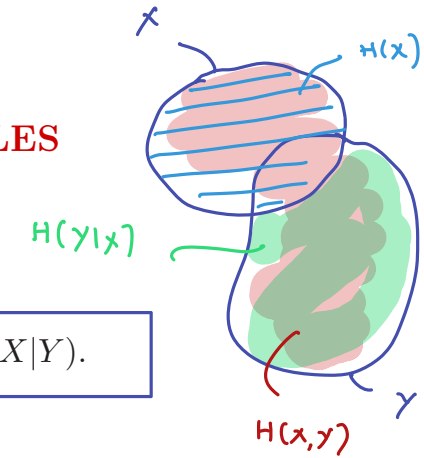
(x_1, y_2) (x_2, y_{10}) ....



(x_1, y_j) (x_2, y_j) ... (x_{10}, y_j)

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.

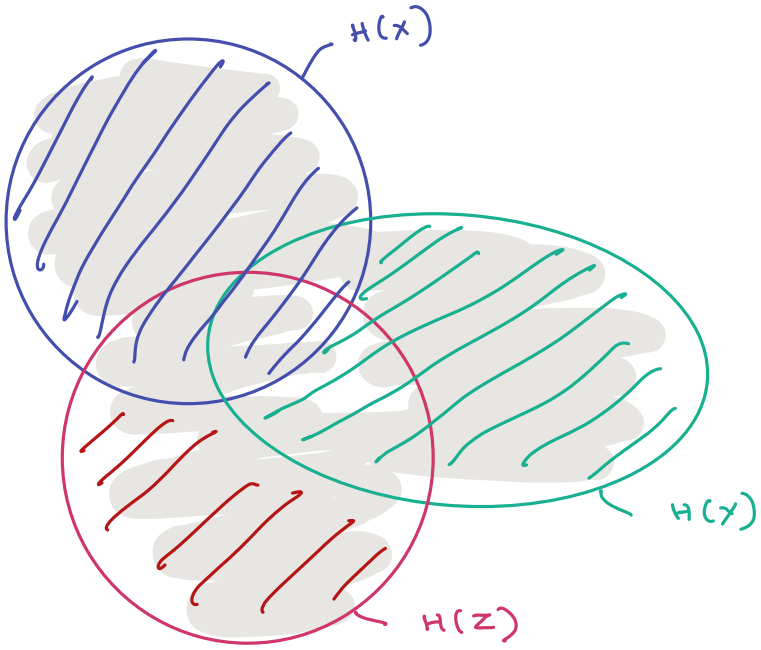
Property 5 (chain rule)**.** *The joint entropy of n random variables can be evaluated using the following chain rule:*

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

X, Y, Z

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

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

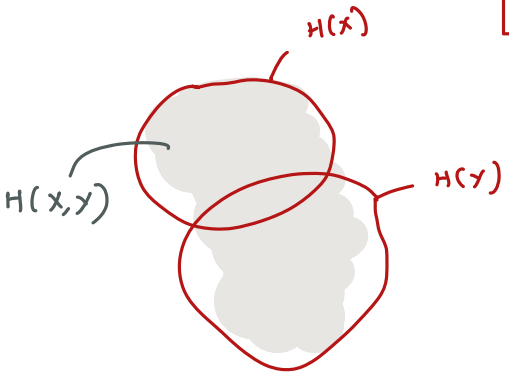


Pair of random variables

Relations between entropies

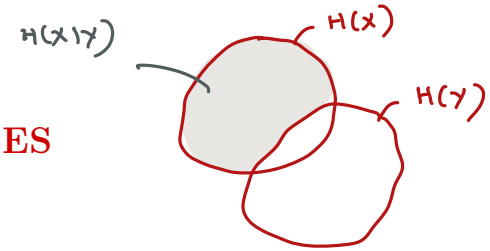
Each term of $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$ is positive. We can conclude that:

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



$$\underbrace{H(x,y)}_{\geq 0} = \underbrace{H(x)}_{\geq 0} + \underbrace{H(y|x)}_{\geq 0}$$

$$\Rightarrow H(x,y) \geq H(x)$$
$$H(x,y) \geq H(y|x)$$



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:

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

← 2 r.v.

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,*

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

← n r.v.

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

$X_1, \cdots, X_m$

$$H(X_1) \geq H(X_1 \mid X_2) \geq H(X_1 \mid X_2, X_3) \geq \cdots$$

Pair of random variables

Relations between entropies

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

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

① : entropy is positive always.

because $H(X)$ is a linear combination of $f(x) = -x \log x$ with $x \in]0,1]$, $f(x) \geq 0$ over $]0,1]$.

$$H(X) = -\sum_i p_i \log p_i$$

② conditionning decreases entropy (slide 21)

③ $\underbrace{H(X,Y)}_{\geq 0} = \underbrace{H(X)}_{\geq 0} + \underbrace{H(Y|X)}_{\geq 0} \Rightarrow H(X,Y) \geq H(X)$

④ $H(X,Y) = H(X) + H(Y|X)$

we know that : $H(Y|X) \leq H(Y)$

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

⑤ $H(X) \leq H(X,Y)$

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

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

Pair of random variables

Mutual information

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

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

or, equivalently,

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

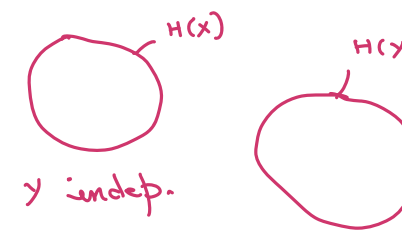
The mutual information quantifies the amount of information obtained about one random variable through observing the other random variable.

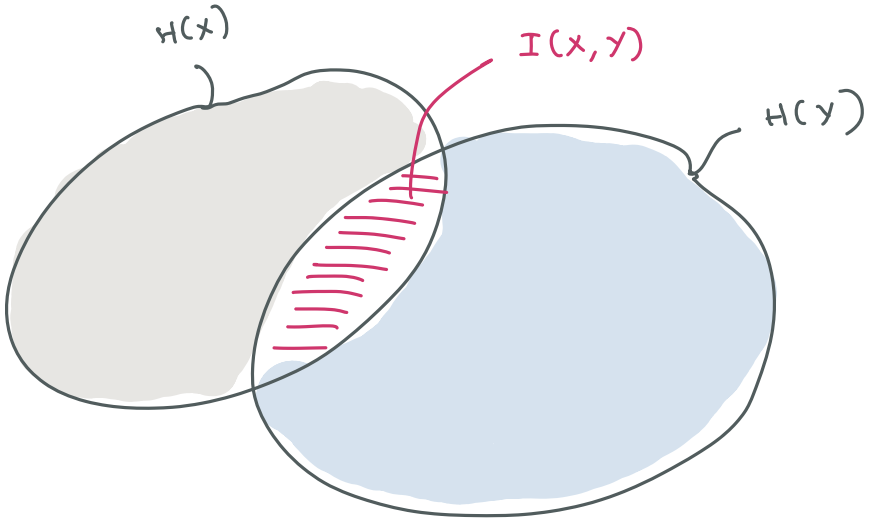
Exercise. Case of two independent random variables

$I(X,Y) = 0$ Sh

$I(X,Y) = H(X) - \underbrace{H(X|Y)}_{H(X) \text{ if } X \text{ and } Y \text{ indep.}}$

$= 0$





$$
\begin{aligned}
I(x,y) &\triangleq H(x) - H(x|y) \\
&\triangleq H(y) - H(y|x) \\
&\triangleq H(x,y) - H(x|y) - H(y|x)
\end{aligned}
$$

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

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

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.

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

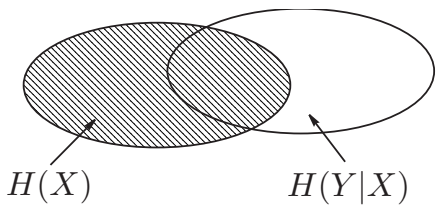
$= d\left(P(x,y),\ P(x)P(y)\right)$ if x and y are indep. $P(x,y) = P(x)P(y)$

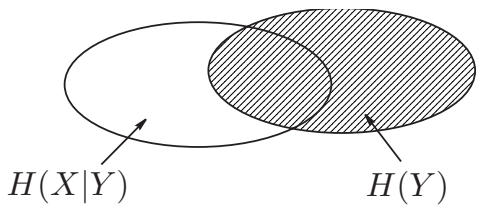
$\Rightarrow \quad d(P(x,y), P(x)P(y)) = 0$

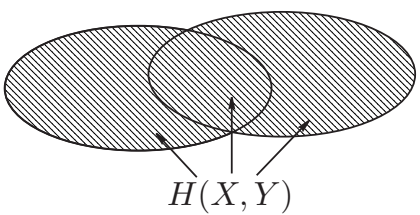
Pair of random variables

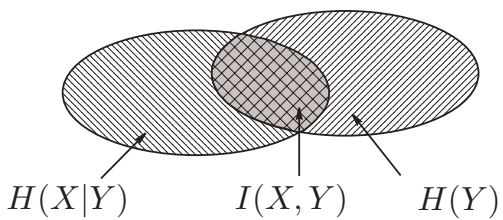
Venn diagram

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

Probability Calculation

Exercise 1

Consider a set of 52 playing cards in which we randomly choose cards with replacement. Calculate the probability to get a queen, a heart, the queen of hearts or the ace of spades, a queen or a spade, neither a queen nor a spade.

Exercise 2

We consider an urn containing 5 white balls, 4 red balls and 2 black balls. We randomly select one ball from this urn. Calculate the probability that it is white, that it is not white, that it is white or red. Now we select 3 balls with replacement. Calculate the probability to get a white ball first, next a red ball, and finally a black ball. Solve the same problem in the case where the balls are randomly selected without replacement.

Exercise 3

During a Poker game, you randomly select 5 cards among 52 playing cards. Calculate the probability of obtaining a pair (2 cards of the same height), three of a kind (3 cards of the same height), a flush (5 cards of the same suit), a full house (three of a kind and a pair), a four of a kind (4 cards of the same height).

Exercise 4

Consider two teams E_1 and E_2 playing football one against the other. The probability that E_1 wins or that the game ends in a draw are $1/2$ and $1/6$, respectively. During a tournament, these two teams play 5 games one against the other. Calculate the probability that E_1 wins all the games, that E_1 does not win at least once, that 2 games end in a draw.

Exercise 5

An urn I contains 2 black balls and 3 white balls, while an urn II contains 4 black balls and6 white balls. A ball is randomly selected in each urn. Calculate the probability of drawing 2 balls with the same color. Now, we assume that the ball randomly selected in urn I is placed in urn II before proceeding to the second draw. Calculate the probability of getting 2 balls of the same color.

Exercise 6

A person is randomly selected from a population where the proportion of cheaters is equal to p. It is asked someone to draw a card from a set of 52 playing cards. One admits that cheaters always get an ace. Calculate the probability that the selected person gets an ace. Calculate the probability that this person is a cheater if he gets an ace.

Exercise 7

We consider the roll of 2 dices. Let X denote the sum of points, and let Y denote the largest number of points obtained with one of these two dices. Study these two random variables.

Exercise 1

Consider a set of 52 playing cards in which we randomly choose cards with replacement. Calculate the probability to get a queen, a heart, the queen of hearts or the ace of spades, a queen or a spade, neither a queen nor a spade.

A = Q (Q : queen)

= $Q\heartsuit$ or $Q\clubsuit$ or $Q\spadesuit$ or $Q\diamondsuit$

$$P(A) = P(Q\heartsuit \text{ or } Q\clubsuit \text{ or } Q\spadesuit \text{ or } Q\diamondsuit)$$
$$= P(Q\heartsuit) + P(Q\clubsuit) + P(Q\spadesuit) + P(Q\diamondsuit)$$
$$= \frac{1}{52} + \frac{1}{52} + \frac{1}{52} + \frac{1}{52}$$
$$= \frac{4}{52}$$
$$= \frac{1}{13}$$

$B = \heartsuit$

= $1\heartsuit$ or $2\heartsuit$ or .... or $Q\heartsuit$ or $K\heartsuit$

$$P(B) = P(1\heartsuit \text{ or } 2\heartsuit \text{ or } \dots \text{ or } Q\heartsuit \text{ or } K\heartsuit)$$
$$= P(1\heartsuit) + \dots + P(K\heartsuit)$$
$$= 13 \times \frac{1}{52}$$
$$= \frac{1}{4}$$

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$C = Q\heartsuit$ or $1\spadesuit$

$$P(C) = P(Q\heartsuit \text{ or } 1\spadesuit)$$
$$= P(Q\heartsuit) + P(1\spadesuit) - \underbrace{P(Q\heartsuit \text{ and } 1\spadesuit)}_{0}$$
$$= \frac{1}{52} + \frac{1}{52} +$$
$$= \frac{1}{26}$$

D = Q or ♠

$$P(D) = P(Q \text{ or } ♠)$$

$$= P(Q) + P(♠) - P(\underbrace{Q \text{ and } ♠}_{Q♠})$$

$$= \frac{4}{52} + \frac{1}{4} - \frac{1}{52}$$

$$= \frac{3 + 13}{52}$$

$$= \frac{16}{52}$$

$$\overline{A \text{ and } B} = \bar{A} \text{ or } \bar{B}$$

$E = \bar{Q}$ and $\bar{♠}$

$$P(E) = P(\bar{Q} \text{ and } \bar{♠})$$

$$= P(\overline{Q \text{ or } ♠})$$

$$= 1 - P(Q \text{ or } ♠)$$

$$= 1 - \frac{16}{52}$$

$$= \frac{52 - 16}{52}$$

$$= \frac{36}{52}$$

Exercise 2

We consider an urn containing 5 white balls, 4 red balls and 2 black balls. We randomly select one ball from this urn. Calculate the probability that it is white, that it is not white, that it is white or red. Now we select 3 balls with replacement. Calculate the probability to get a white ball first, next a red ball, and finally a black ball. Solve the same problem in the case where the balls are randomly selected without replacement.

$$P(W) = \frac{5}{5+4+2} = \frac{5}{11}$$

$$P(\overline{W}) = 1 - P(W) = 1 - \frac{5}{11} = \frac{6}{11}$$

$$P(R \text{ or } W) = P(R) + P(W) - \underbrace{P(R \text{ and } W)}_{0}$$

$$= \frac{4}{11} + \frac{5}{11}$$

$$= \frac{9}{11}$$

with replacement :

$$P(W_1 \text{ and } R_2 \text{ and } B_3)$$

As selection is with replacement, all 3 events are independent

$$P(W_1 \text{ and } R_2 \text{ and } B_3) = P(W_1)\,P(R_2)\,P(B_3)$$

Rm : $P(A \text{ and } B) = P(A, B)$

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

$$= P(A)\,P(B) \text{ if}$$

A and B indep.

i.e, $P(B|A) = P(B)$

$$P(W_1 \text{ and } R_2 \text{ and } B_3) = P(W_1)\,P(R_2)\,P(B_3)$$

$$= \frac{5}{11} \times \frac{4}{11} \times \frac{2}{11}$$

$$= \frac{40}{(11)^3} \approx 0{,}03$$

without replacement :

$$P(W_1 \text{ and } R_2 \text{ and } B_3)$$

$$= P(W_1)\, P(R_2 \mid W_1)\, P(B_3 \mid W_1, R_2)$$

$$= \frac{5}{11} \times \frac{4}{10} \times \frac{2}{9}$$

$$= \frac{40}{990} = \frac{4}{99} \approx 0{,}04$$

Rm :

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

(W_1 : A, R_2 : B)

$$P(A, B, C) = P(A)\, P(B, C \mid A)$$

(B_3 : B, C)

$$= P(A)\, P(B \mid A)\, P(C \mid A, B)$$

$P(A, B)$ $\quad$ $P(C \mid A, B)$

$P(A, B, C)$

Exercise 5

An urn I contains 2 black balls and 3 white balls, while an urn II contains 4 black balls and6 white balls. A ball is randomly selected in each urn. Calculate the probability of drawing 2 balls with the same color. Now, we assume that the ball randomly selected in urn I is placed in urn II before proceeding to the second draw. Calculate the probability of getting 2 balls of the same color.

$$E = (B_I \text{ and } B_{II}) \text{ or } (W_I \text{ and } W_{II})$$

$$= P(B_I \text{ and } B_{II}) + P(W_I \text{ and } W_{II}) - \underbrace{P(\dots \cap \dots)}_{0}$$

$$= P(B_I)\,P(B_{II}) + P(W_I)\,P(W_{II})$$

$$= \frac{2}{5} \times \frac{4}{10} + \frac{3}{5} \times \frac{6}{10}$$

$$= \frac{8 + 18}{50} = \frac{26}{50} = \frac{13}{25} = 0,52$$

$$E' = (B_I \text{ and } B_{II}) \text{ or } (W_I \text{ and } W_{II})$$

$$= P(B_I \text{ and } B_{II}) + P(W_I \text{ and } W_{II})$$

$$= P(B_I)\,P(B_{II} \mid B_I) + P(W_I)\,P(W_{II} \mid W_I)$$

$$= \frac{2}{5} \times \frac{5}{11} + \frac{3}{5} \times \frac{7}{11}$$

$$= \frac{31}{55} \approx 0,56$$

Exercise 6

A person is randomly selected from a population where the proportion of cheaters is equal to p. It is asked someone to draw a card from a set of 52 playing cards. One admits that cheaters always get an ace. Calculate the probability that the selected person gets an ace. Calculate the probability that this person is a cheater if he gets an ace.

$$E = \text{ace}$$

$$= (\text{ace and cheater}) \text{ or } (\text{ace and } \overline{\text{cheater}})$$

$$P(E) = P(\text{ace} \mid \text{cheater})P(\text{cheater}) + P(\text{ace} \mid \overline{\text{cheater}})P(\overline{\text{cheater}})$$

$$= 1 \cdot p + \frac{4}{52} \cdot (1 - p)$$

$$= \frac{52p + 4 - 4p}{52} = \frac{48p + 4}{52}$$

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

$$\Rightarrow P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

$$P(\text{cheater} \mid \text{ace}) = \frac{P(\text{ace} \mid \text{cheater})P(\text{cheater})}{P(\text{ace})}$$

$$= \frac{P(\text{ace} \mid \text{cheater})P(\text{cheater})}{P(\text{ace} \mid \overline{\text{cheater}})P(\overline{\text{cheater}}) + P(\text{ace} \mid \text{cheater})P(\text{cheater})}$$

$$= \frac{1 \cdot p}{\frac{48p + 4}{52}}$$

$$= \frac{52p}{48p + 4} = \frac{13p}{21p + 1}$$

15 : 15

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 $p_k = N_k/N$.

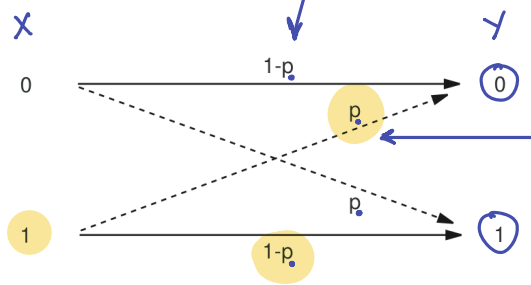
1. Determine the self-information of any element of $\mathcal{E}_k$. Calculate the average amount of information needed to determine any element in $\mathcal{E}_k$.
2. Calculate the average amount of information needed to characterize any element of $\mathcal{E}$. By noticing that we can split the identification 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 $\mathcal{E}_k$.

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 $I(X,Y)$.

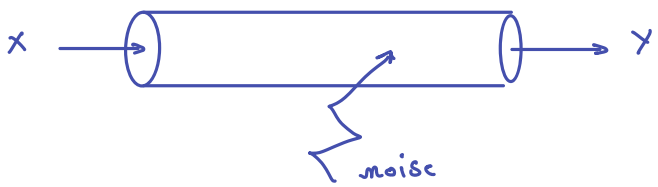
$P(X=0) = \frac{1}{4}$

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



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

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



$H(X) = H\left(\frac{1}{4}, \frac{3}{4}\right) =$

$H(Y) = H\left(\frac{1+2p_0}{4}, \frac{3-2p_0}{4}\right)$

$H(X,Y) = H\left(\frac{3}{4}p_0, \frac{3}{4}(1-p_0), \frac{1}{4}(1-p_0), \frac{1}{4}p_0\right)$

$H(Y|X) = H(p_0, 1-p_0)$

$H(X|Y) = ?$

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

$$H(X|Y) = \underbrace{H(X|Y=0)}_{\downarrow} P(Y=0) + \underbrace{H(X|Y=1)}_{\downarrow} P(Y=1)$$

$$P(X=0|Y=0) = \frac{P(Y=0|X=0)P(X=0)}{P(Y=0)}$$

$$= \frac{(1-p_0)\frac{1}{4}}{\frac{1+2p_0}{4}} = \frac{1-p_0}{1+2p_0}$$

$$P(X=1|Y=0) = 1 - \frac{1-p_0}{1+2p_0} = \frac{1+2p_0-1+p_0}{1+2p_0}$$

$$= \frac{3p_0}{1+2p_0}$$

$$P(X=0|Y=1) = \frac{P(Y=1|X=0)P(X=0)}{P(Y=1)}$$

$$= \frac{p_0 \times \frac{1}{4}}{\frac{3-2p_0}{4}} = \frac{p_0}{3-2p_0}$$

$$P(x=1 \mid y=1) = 1 - \frac{p_0}{3-2p_0}$$

$$= \frac{3-2p_0-p_0}{3-2p_0}$$

$$= \frac{3-3p_0}{3-2p_0}$$

$H(x) = H\left(\frac{1}{4}, \frac{3}{4}\right) = 0,81$ Sh/state of x

$H(y) = H\left(\frac{1+2p_0}{4}, \frac{3-2p_0}{4}\right) = H(0,3 \; ; \; 0,7) = 0,88$

$H(x,y) = H\left(\frac{3}{4}p_0, \frac{3}{4}(1-p_0), \frac{1}{4}(1-p_0), \frac{1}{4}p_0\right)$

$= H(0,075 \; ; \; 0,675 \; ; \; 0,225 \; , \; 0,025)$

$= 1,28$ Sh / pair of states (x,y)

$H(y|x) = H(p_0, 1-p_0) = 0,47$ Sh/state of y

$I(x,y) = H(y) - H(y|x) = 0,88 - 0,47 = 0,41$ Sh/pair of state

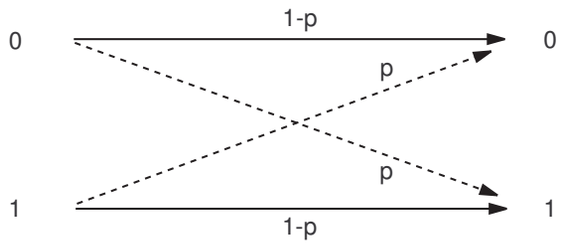


Figure 1: Imperfect channel.

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.

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 m 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?

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 m 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.

X has n states. The maximum entropy is:

$$H_{max}(X) = \log_2 n \quad \text{Sh/state}$$

$H_{max}(X)$ is reached if X uniformly distributed.

2. Calcule the entropy of Y.

$$H(Y) = \log_2 m \quad \text{Sh/state}$$

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.

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

① : entropy always positive or zero

$$H = -\sum_i p_i \log p_i = \sum_i f(p_i)$$

with $\begin{cases} f(x) = -x \log_2 x \text{, with } x \in]0,1] \\ f(0) = 0 \end{cases}$

one can show that $f(x) \geq 0$ for all $x \in [0,1]$

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

③ X and Y have the same number of states.
Y is uniformly distributed

$$\Rightarrow H(X) \leq \underbrace{H(Y)}_{\log_2 m}$$

④ $H(X,Y) = H(Y) + \underbrace{H(X|Y)}_{\geq 0} \Rightarrow H(X,Y) \geq H(Y)$

⑤ $H(X,Y) = H(Y) + H(X|Y) \leq H(Y) + H(X)$

we know that $H(X|Y) \leq H(X)$

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

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.

① several reasons to have $0 = H(X|Y)$

* There exists a function f : $X = f(Y)$
* $P(X = x_1) = 1$ and $P(X = x_i) = 0$ for all $i \neq 1$

② equality if X and Y are independent : $H(X|Y) = H(X)$

③ equality if X uniformly distributed (as Y is)

④ we have : $H(X,Y) = H(Y) + H(X|Y)$

$\Rightarrow H(X,Y) = H(Y)$ if $H(X|Y) = 0$ condition, see ①

⑤ we have : $H(X,Y) = H(Y) + H(X|Y)$

$\Rightarrow H(X,Y) = H(Y) + H(X)$ if $H(X|Y) = H(X)$ condition, see ②

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 $I(X;Y)$.

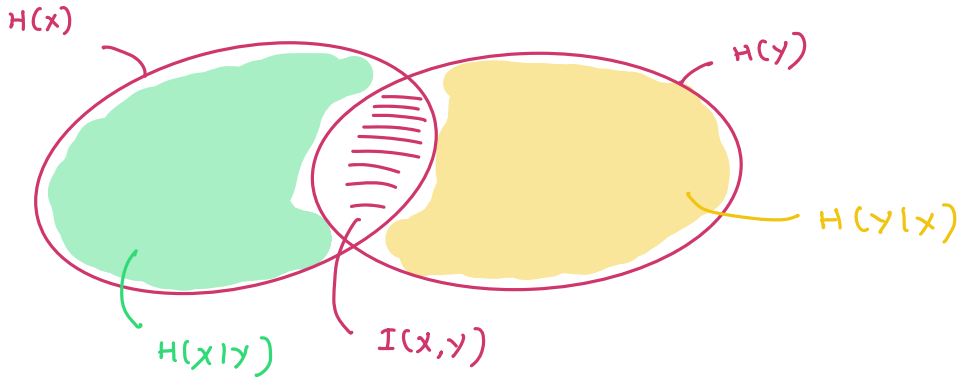
$P(Y = y_i) = \sum_{j=1}^{3} P(X = x_j, Y = y_i)$

sum →

$P(X;Y)$	y_1	y_2	y_3	y_4	$P(X = x_j)$
x_1	1/24	1/12	1/6	1/24	$\frac{1+2+4+1}{24} = \frac{1}{3}$
x_2	1/6	1/8	1/24	1/6	$\frac{4+3+1+4}{24} = \frac{1}{2}$
x_3	1/24	1/24	1/24	1/24	$\frac{4}{24} = \frac{1}{6}$
$P(Y = y_i)$	1/4	1/4	1/4	1/4	

sum ↓

Y is uniformly distributed



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

$$= \log_2 4 - H(Y|X) \quad \text{because } Y \text{ uniformly distributed}$$

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

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

$\boxed{H(Y|X=x_1)}$ we need $P(Y=y_j \mid X=x_1)$ for all $j = 1, \dots, 4$

$$P(Y=y_j \mid X=x_1) = \frac{P(X=x_1,\ Y=y_j)}{P(X=x_1)}$$

$P(Y=y_1 \mid X=x_1) = (1/24)/(1/3) = 1/8$

$P(Y=y_2 \mid X=x_1) = (1/12)/(1/3) = 1/4$

$P(Y=y_3 \mid X=x_1) = (1/6)/(1/3) = 1/2$

$P(Y=y_4 \mid X=x_1) = (1/24)/(1/3) = 1/8$

$$H(Y|X=x_1) = -\sum_{i=1}^{4} P(Y=y_i|X=x_1) \log_2 P(Y=y_i|X=x_1)$$

$$= 1{,}75 \text{ Sh / state (of } Y)$$

$$H(Y|X=x_2) = H(1/3,\ 1/4,\ 1/3,\ 1/12)$$

$$= 1{,}85 \text{ Sh / state}$$

$$H(Y|X=x_3) = H(1/4,\ 1/4,\ 1/4,\ 1/4)$$

$$= 2 \text{ Sh / state}$$

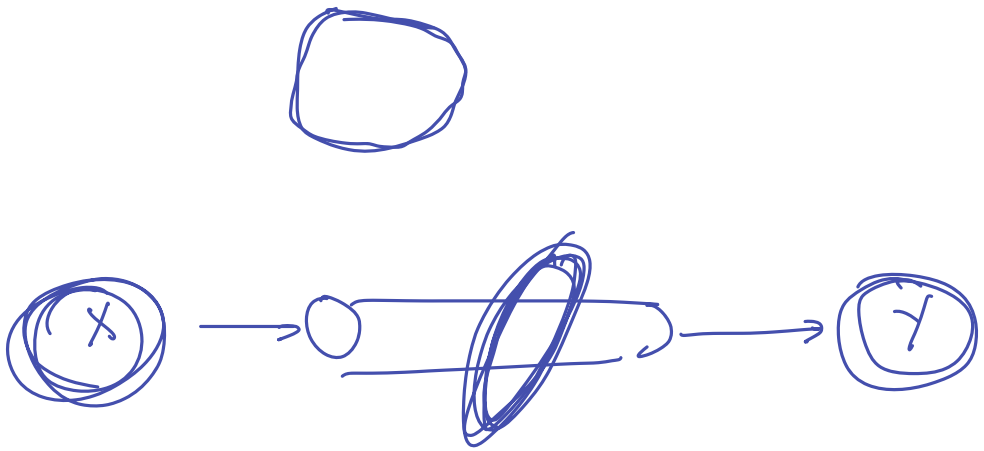
$$H(Y|X) = \frac{1,75}{3} + \frac{1,85}{2} + \frac{2}{6} = \frac{3,5 + 5,55 + 2}{6}$$

$$= \frac{11,05}{6}$$

$$= 1,84 \text{ Sh / state}$$

$$I(X,Y) = 2 - 1,84 = 0,16 \text{ Sh / state of } (X,Y)$$

$\log 12 \approx 3,6$ Sh



$I(X,Y)$ is max

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 $\overline{n}$?

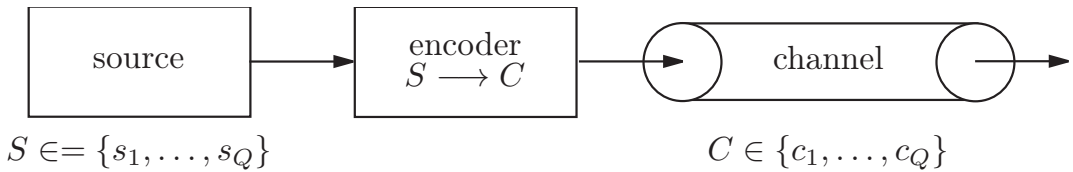
Information Theory and Coding

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: $\mathcal{C} = \{c_1, \ldots, c_q\}$.



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 !!!

Discrete-time source

General model

A discrete source S is defined by an alphabet $\mathcal{A} = \{s_1, \ldots, s_Q\}$ 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.

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

Discrete-time source

Complementary assumptions

Again, for the sake of simplicity, the following ergodicity assumption can be made.

Property 2 (Ergodic process)**.** *A stationary random process S_i is ergodic if, for every $k = 1, 2, \ldots$, for every set of indices $i_1, \ldots, i_k$ and for any bounded function $f(\cdot)$ from $\mathcal{A}^k$ into $\mathbb{R}$, we have:*

$$\frac{1}{n}\sum_{k=1}^{n} f(S_{i_1}, \ldots, S_{i_k}) \xrightarrow{a.s.} E\{f(S_{i_1}, \ldots, S_{i_k})\}.$$

Interest. An ergodic process can be studied by observing any long enough trajectory.

Discrete-time source

Markov source

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

Definition 1 (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

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

Discrete-time source

Markov source

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

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

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

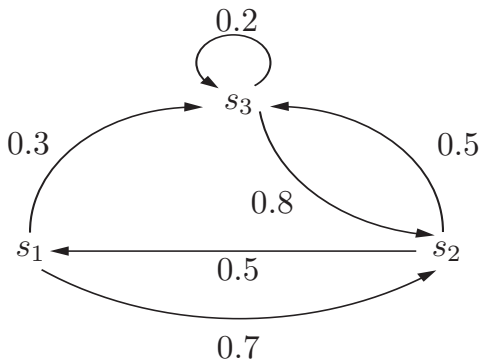
$$\Pi(i, j) = P(S_2 = s_j|S_1 = s_i)$$

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

Discrete-time source

Example of Markov source

Consider the following Markov source



The corresponding transition matrix can be written as:

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

Discrete-time source

Markov source in steady state

Definition 3 (steady-state - version 1)**.** *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 $i \in \{1, \ldots, Q\}$.

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.

Discrete-time source

Markov source in steady state

Definition 4 (steady-state - version 2)**.** *Consider a Markov source S. If it exists, its steady state distribution is defined as:*

$$\lim_{n\to\infty} P(S_n = s_i | S_1 = j)$$

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

Main interest. The asymptotic behavior of S is independent of the initial distribution.

Discrete-time source

m-th order Markov source

A Markov source is characterized by a memory of size $m = 1$. This can be generalized to memory sizes $m > 1$.

Definition 5 (m-th order Markov source)**.** *A source S is an m-th order Markov source if:*

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

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

Remark. Any m-th order Markov source S can be expressed as a 1-st order Markov source by considering an m-th order extension of S.

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 6 (Entropy of a stationary source - version 1)**.** *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

Definition 7 (Entropy of a stationary source - version 2)**.** *The entropy of a stationary S source is defined as:*

$$H_0 \triangleq \lim_{n \to +\infty} \frac{H(S_1, \ldots, S_n)}{n}.$$

Both definitions are equivalent in the case of stationary sources. Indeed, it results from the following equality:

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

that $H(S_1, \ldots, S_n)/n$ is the arithmetic mean of the n first terms of the series $H(S_1)$, $H(S_2|S_1)$, $\ldots$, $H(S_n|S_1, \ldots, S_{n-1})$. Cesaro's theorem yields the expected result.

Cesaro's theorem. If $a_n \underset{n \to \infty}{\longrightarrow} a$, then $\frac{1}{n} \sum_{k=1}^{n} a_k \underset{n \to \infty}{\longrightarrow} a$

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:

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

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

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.
>
> **Separator.** A symbol of the alphabet is used as a word separator.
>
> **Without prefix.** A code is called a prefix code or an instantaneous code if no codeword is a prefix of any other codeword.

Exercise. Characterize codes A to G.

Toward Shannon's first theorem

Kraft's inequality

We propose to design uniquely decodable codes, and more particularly instantaneous codes, that are as compact as possible.

Kraft's inequality provides a necessary and sufficient condition on the existence of instantaneous codes for given codeword lengths.

Theorem 1 (Kraft's inequality)**.** *Let $n_1, \ldots, n_Q$ be candidate codeword lengths to encode the Q symbols of a source with a q-ary alphabet. A necessary and sufficient condition for the existence of an instantaneous code with these codeword lengths is given by:*

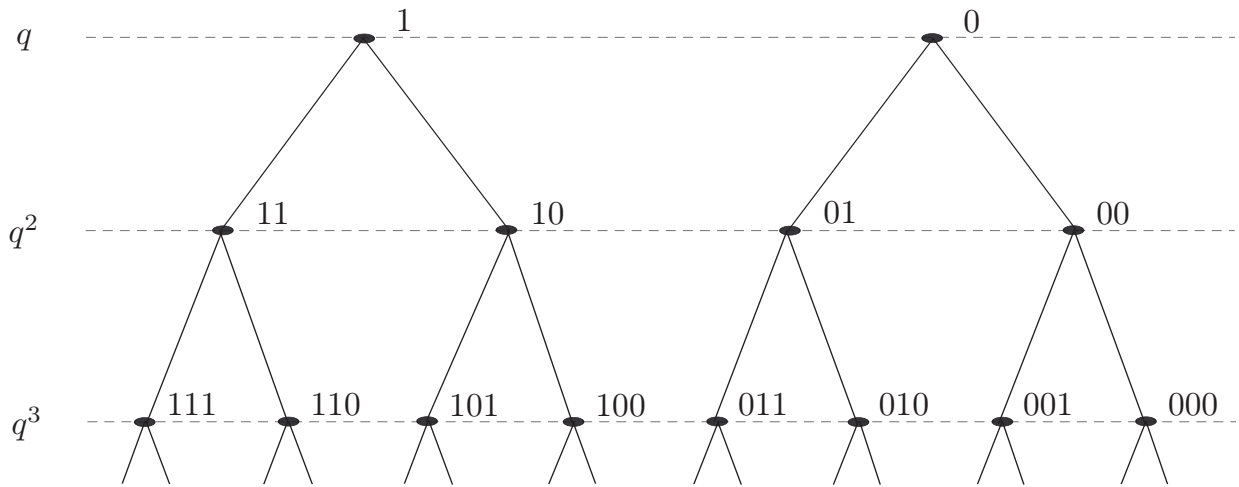
$$\sum_{i=1}^{Q} q^{-n_i} \leq 1.$$

Remark. The same necessary and sufficient condition was established by McMillan for uniquely decodable codes, previously to Kraft's inequality.

Toward Shannon's first theorem

Kraft's inequality

Proof. The following representation, in the case of a binary code, makes the proof clear.



Toward Shannon's first theorem

Kraft's inequality

Let $n_1 \leq \ldots \leq n_Q$. Consider a q-ary tree of height n_Q. This tree then has q^{n_Q} leaf nodes.

Necessary condition. The prefix condition requires that a codeword of length n_i excludes $q^{n_Q - n_i}$ leaf nodes. Therefore, the total number of excluded leaf nodes must satisfy:

$$\sum_{i=1}^{Q} q^{n_Q - n_i} \leq q^{n_Q}.$$

Sufficient condition.

First, we select a node with depth n_1, which excludes $q^{n_Q - n_1}$ leaf nodes. However, there are still available leaf nodes because, using Kraft's inequality, we know that

$$q^{n_Q - n_1} < q^{n_Q}$$

On the way to one of the available leaf nodes, we select a node with depth n_2, ...

Toward Shannon's first theorem

McMillan's inequality

Kraft's inequality implies that McMillan's inequality is sufficient since any prefix code is uniquely decodable.

Necessary condition. Consider the following expansion:

$$\left(\sum_{i=1}^{Q} r_i\, q^{-i}\right)^N = \sum_{n=1}^{NQ} \nu(n)\, q^{-n}$$

where $\nu(n) = \sum_{i_1+\ldots i_N=n} r_{i_1} \ldots r_{i_N}$. Interpreting r_i as the number of codewords of length i, $\nu(n)$ corresponds to the number of messages of length n. The unique decodability condition implies that $\nu(n) \leq q^n$. Then we have:

$$\sum_{i=1}^{Q} r_i\, q^{-i} \leq (NQ)^{\frac{1}{N}},$$

which leads to the result by considering the limit of the upper bound when N tends to infinity.

Toward Shannon's first theorem

McMillan's inequality

Definition 8 (Complete code)**.** *A code is complete if:*

$$\sum_{i=1}^{Q} q^{-n_i} = 1.$$

Toward Shannon's first theorem

McMillan's inequality

As an example, McMillan's inequality is applied to different codes.

	code A	code B	code C
s_1	00	0	0
s_2	01	100	10
s_3	10	110	110
s_4	11	111	11
$\sum_{i=1}^{4} 2^{-n_i}$	1	7/8	9/8

Codes A and B are uniquely decodable, the first one being complete. Code C is not uniquely decodable.

Toward Shannon's first theorem

Consequences of McMillan's inequality

Let S be a memoryless source with Q symbols. Let p_i be the probability of symbol s_i, encoded to a q-ary codeword of length n_i. By setting:

$$q_i = \frac{q^{-n_i}}{\sum_{j=1}^{Q} q^{-n_j}},$$

then applying Gibb's inequality with p_i and q_i, we obtain:

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

Applying McMillan's inequality to the right-hand side member of the inequality yields:

$$H(S) - \overline{n} \log q \leq \log \sum_{i=1}^{Q} q^{-n_i} \leq 0,$$

where $\overline{n} = \sum_{i=1}^{Q} p_i \, n_i$ is the expected length of the codewords.

Toward Shannon's first theorem

Consequences of McMillan's inequality

Theorem 2. *The expected length $\overline{n}$ of the codewords of any uniquely decodable code is lower-bounded by:*

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

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

Toward Shannon's first theorem

Consequences of McMillan's inequality

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

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

Shannon's first theorem

Statement and demonstration

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

Theorem 3. *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.

Premier théorème de Shannon

énoncé et démonstration

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

$$\frac{H(S_1, \ldots, S_k)}{k \log q} \leq \frac{\overline{n}_k}{k} < \frac{H(S_1, \ldots, S_k)}{k \log q} + \frac{1}{k}.$$

In this expression, $\overline{n}_k$ denotes la longueur denotes the expected length of the codewords used to encode the k^{th} extension of source S.

In the case of a stationary source, we know that $\lim_{k\to\infty} H(S_1, \ldots, S_k)$ exists. Denoting this limit by H_0 yields:

$$\lim_{k\to\infty} \frac{\overline{n}_k}{k} = \frac{H_0}{\log q}.$$

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.

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:

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

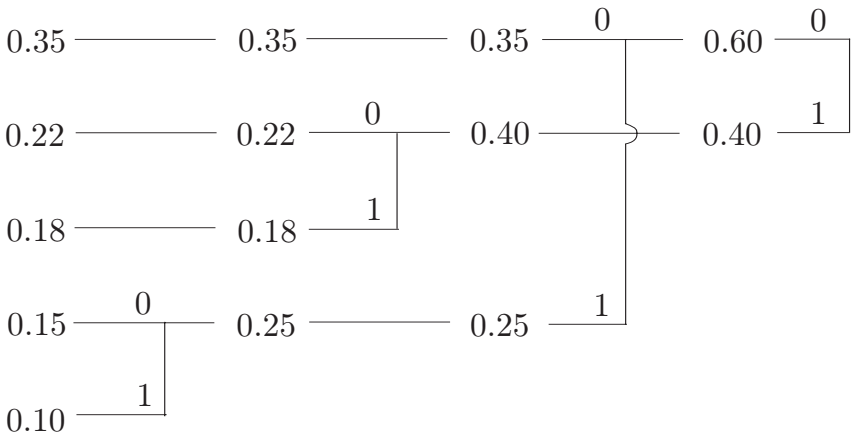
Huffman's code

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.

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.

Information Theory and Coding

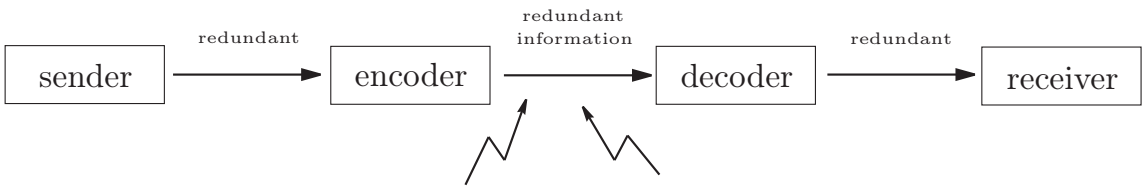
Discrete channels

Cédric RICHARD
Université Côte d'Azur

Channel coding

Motivations

In a real system, the message received by the recipient can differ from that emitted by the source due to perturbations. We talk about *noisy channel*.



Channel coding consists of introducing redundancy into the message

$\longrightarrow$ prevent the loss of information due to the channel

Discrete channel models

General model

A discrete channel is a stochastic system that accepts, as an input, symbol sequences defined on an alphabet $\mathcal{X}$, and outputting sequences of symbols defined on an alphabet $\mathcal{Y}$.

Inputs and outputs are linked by a probabilistic model:

$$P(Y_1 = y_1, \ldots, Y_m = y_m | X_1 = x_1, \ldots, X_n = x_n)$$

$\triangleright$ **model too general to give rise to simple derivations**

Discrete channel models

Properties

For the sake of simplicity, assumptions are made about the model.

Property 1 (Causal channel)**.** *A channel is causal if:*

$$P(Y_1 = y_1, \ldots, Y_m = y_m | X_1 = x_1, \ldots, X_n = x_n) \\ = P(Y_1 = y_1, \ldots, Y_m = y_m | X_1 = x_1, \ldots, X_m = x_m)$$

for all m and n such that $m \leq n$.

Consequence. By summing both equality members with respect to $Y_1, \ldots, Y_{m-1}$, we check:

$$P(Y_m = y_m | X_1 = x_1, \ldots, X_n = x_n) = P(Y_m = y_m | X_1 = x_1, \ldots, X_m = x_m)$$

$\longrightarrow$ any output is independent of future inputs

Discrete channel models

Properties

Property 2 (Memoryless causal channel)**.** *A channel is said to be memoryless if, for all $k \geq 2$, we have:*

$$P(Y_k = y_k | X_1 = x_1, \ldots, X_k = x_k, Y_1 = y_1, \ldots, Y_{k-1} = y_{k-1}) \\ = P(Y_k = y_k | X_k = x_k).$$

Consequence. The conditional law governing the channel behavior is entirely determined by the instantaneous conditional laws:

$$P(Y_1 = y_1, \ldots, Y_m = y_m | X_1 = x_1, \ldots, X_n = x_n) = \prod_{k=1}^{m} P(Y_k = y_k | X_k = x_k).$$

$\longrightarrow P(Y_k = y_k | X_k = x_k)$ may be time-dependent.

Discrete channel models

Properties

Noticing that $P(Y_k = y_k | X_k = x_k)$ may depend on the time instant k, we introduce the following property:

Property 3 (Stationary memoryless channel)**.** *A memoryless channel is stationary if, for all $k \geq 1$, we have:*

$$P(Y_k = y_k | X_k = x_k) = P(Y = y_k | X = x_k).$$

Notation. We denote by $(\mathcal{X}, \mathcal{Y}, \Pi)$ any discrete memoryless channel, where Π is the transition matrix defined as:

$$\Pi(i, j) = P(Y = y_j | X = x_i)$$

Discrete channel models

Symmetric channel

A channel is *symmetric* if the rows of its transition matrix all have the same entries up to a permutation, as well as its columns.

Examples. The following transition matrices correspond to symmetric channels.

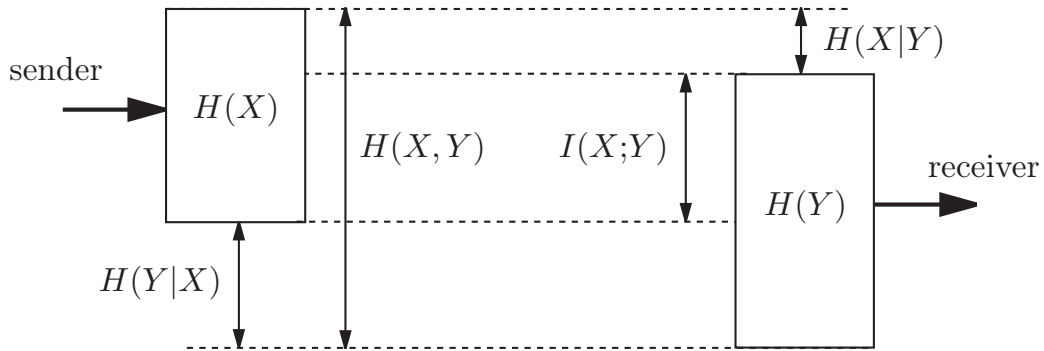
$$\Pi = \begin{pmatrix} p & q & 1-p-q \\ q & 1-p-q & p \\ 1-p-q & p & q \end{pmatrix},$$

$$\Pi = \begin{pmatrix} p & 1-p-q & q \\ q & 1-p-q & p \end{pmatrix},$$

with p and q in $[0, 1]$.

Memoryless channel capacity

Introduction



$H(X)$ is the amount of information transmitted through a noiseless channel

$H(X|Y)$ is the amount information required to determine the entry

$I(X;Y)$ is the amount of information transmitted through the channel.

Memoryless channel capacity

Definition

Definition 1. *The information capacity per symbol of a channel is defined as:*

$$C \triangleq \max_{P(X=x)} I(X;Y).$$

Caution. We check that $I(X,Y)$ is a concave function of the law of X. Indeed, by writing $f(x) = -x \log x$, we note that it is a sum of concave functions:

$$\begin{aligned} I(X;Y) &= \sum_i \sum_j p(i,j) \log \frac{p(i,j)}{p(i)\,p(j)} \\ &= \sum_i \sum_j p_i\, p_i(j) \log \frac{p_i(j)}{\sum_i p_i\, p_i(j)} \\ &= \sum_i p_i \left(\sum_j p_i(j) \log p_i(j) \right) + \sum_j f\left(\sum_i p_i p_i(j) \right). \end{aligned}$$

Memoryless channel capacity

Capacity calculation

In the general case, calculating the capacity of a channel is complicated. However, in the case of a symmetric channel, the calculation is easy.

Theorem 1. *The capacity of a symmetric channel* $(\mathcal{X}, \mathcal{Y}, \Pi)$ *is equal to* $I(X;Y)$ *in the case where* X *is governed by a uniform law.*

Proof. The entropy $H(Y|X = x_i) = -\sum_j p_i(j) \log p_i(j)$ is independent of i since the rows i of Π all have the same entries. As a consequence, $H(Y|X)$ is independent of the law of X.

It is easy to check that Y is governed by a uniform law if X is. Indeed,

$$p_j = \sum_i p_i \, p_i(j) = \frac{1}{q} \sum_i p_i(j)$$

is independent of j since the columns of Π all have the same entries. □

Channel capacity calculation

Examples

Noiseless binary channel. The channel outputs are identical to the channel inputs. As a consequence, we have $I(X;Y) = H(X)$ because $H(X|Y) = 0$.

$$C = 1 \text{ Sh/symb}$$

Disfunctional binary channel. This channel always reproduces the same output regardless of the input. Consequently, mutual information $I(X;Y)$ is zero because $H(Y) = H(Y|X) = 0$.

$$C = 0 \text{ Sh/symb}$$

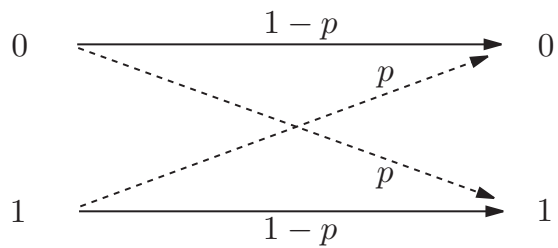
Channel capacity calculation

Binary symmetric channel

The transition matrix of a symmetric binary channel is given by

$$\Pi = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$

which is represented schematically as follows:



Channel capacity calculation

Binary symmetric channel

In order to evaluate the information capacity of this channel, let us calculate first the mutual information $I(X;Y)$:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - P(X=0)H(Y|X=0) - P(X=1)H(Y|X=1). \end{aligned}$$

A simple calculation shows that $H(Y|X=x) = H_2(p)$, where $x \in \{0,1\}$, which leads to:

$$I(X;Y) = H(Y) - H_2(p) \leq \log 2 - H_2(p).$$

As a consequence we have:

$$C = 1 - H_2(p) \text{ Sh/symb}$$

Channel capacity calculation

Binary symmetric channel