

# ELEMENT OF INFORMATION THEORY

O. Le Meur  
[olemeur@irisa.fr](mailto:olemeur@irisa.fr)

Univ. of Rennes 1  
<http://www.irisa.fr/temics/staff/lemeur/>

October 2010

## VERSION:

- 2009-2010: Document creation, done by OLM;
- 2010-2011: Document updated, done by OLM:
  - Slide in introduction: The human communication;
  - Slide on Spearman's rank correlation coefficient;
  - Slide on Correlation does not imply causation;
  - Slide to introduce the amount of information;
  - Slide presenting a demonstration concerning the joint entropy  $H(X, Y)$ .

## ELEMENTS OF INFORMATION THEORY

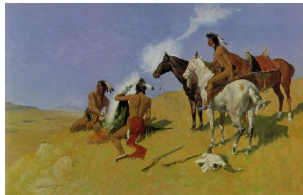
- 1 Introduction
- 2 Statistical signal modelling
- 3 Amount of information
- 4 Discrete source
- 5 Discrete channel
- 6 Shannon's theorem
- 7 Summary

# Information Theory

- 1 Introduction
  - Goal and framework of the communication system
  - Some definitions
- 2 Statistical signal modelling
- 3 Amount of information
- 4 Discrete source
- 5 Discrete channel
- 6 Shannon's theorem
- 7 Summary

## The human communication...

- Visual signal: the smoke signal is one of the oldest forms of communication in recorded history;



Painting by Frederic Remington.

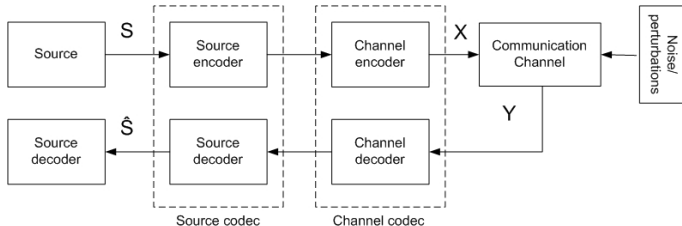
Smoke signals were also employed by Hannibal, the Carthaginian general who lived two centuries BC

- Sonor signal: in Ancient China, soldiers stationed along the Great Wall would alert each other of impending enemy attack by signaling from tower to tower. In this way, they were able to transmit a message as far away as 480 km (300 miles) in just a few hours.

A common point, there is a code...

## Goal and Framework of the communication system

- To transmit an information at the minimum rate for a given quality;
- Seminal work of Claude Shannon (1948)[Shannon,48].



### Ultimate goal

The source and channel codec must be designed to ensure a good transmission of a message given a minimum bit rate or a minimum level of quality.

## Goal and framework of the communication system

Three major research axis:

- 1 **Measure:** Amount of information carried by a message.
- 2 **Compression:**
  - Lossy vs lossless coding...
  - Mastering the distortion  $d(S, \hat{S})$
- 3 **Transmission:**
  - Channel and noise modelling
  - Channel capacity

## Some definitions

### Definition

Source of information: something that produces messages!

### Definition

Message: a stream of symbols taking their values in a predefined alphabet.

### Example

Source: a camera

Message: a picture

Symbols: RGB coefficients

alphabet =  $(0, \dots, 255)$

### Example

Source: book

Message: a text

Symbols: letters

alphabet =  $(a, \dots, z)$

## Some definitions

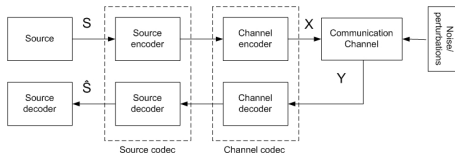
### Definition

*Source Encoder:* the goal is to transform  $S$  in a binary signal  $X$  of size as small as possible (eliminate the redundancy).

*Channel Encoding:* the goal is to add some redundancy in order to be sure to transmit the binary signal  $X$  without errors.

### Definition

$$\text{Compression Rate: } \sigma = \frac{\text{Nb bits of input}}{\text{Nb bits of output}}$$



# Information Theory

- 1 Introduction
- 2 Statistical signal modelling
  - Random variables and probability distribution
  - Joint probability
  - Conditional probability and Bayes rule
  - Statistical independence of two random variables
  - Correlation coefficient
- 3 Amount of information
- 4 Discrete source
- 5 Discrete channel
- 6 Shannon's theorem

## Random variables and probability distribution

The transmitted messages are considered as a random variable with a finite alphabet.

### Definition (Alphabet)

An alphabet  $\mathcal{A}$  is a set of data  $\{a_1, \dots, a_N\}$  that we might wish to encode.

### Definition (Random Variable)

A discrete random variable  $X$  is defined by an alphabet  $\mathcal{A} = \{x_1, \dots, x_N\}$  and a probability distribution  $\{p_1, \dots, p_N\}$ , i.e.  $p_i = P(X = x_i)$ .

Remark: a symbol is the outcome of a random variable.

### Properties

- $0 \leq p_i \leq 1$ ;
- $\sum_{i=1}^N p_i = 1$  also noted  $\sum_{x \in \mathcal{A}} p(x)$ .
- $p_i = P(X = x_i)$  is equivalent to  $P_X(x_i)$  and  $P_X(i)$ .

## Joint probability

### Definition (Joint probability)

Let  $X$  and  $Y$  be discrete random variables defined by alphabets  $\{x_1, \dots, x_N\}$  and  $\{y_1, \dots, y_M\}$ , respectively.

$A$  and  $B$  are the events  $X = x_i$  and  $Y = y_j$ ,

$P(X = x_i, Y = y_j)$  is the joint probability also called  $P(A, B)$  or  $p_{ij}$ .

### Properties of the joint probability density function (pdf)

- $\sum_{i=1}^N \sum_{j=1}^M P(X = x_i, Y = y_j) = 1$ ,
- If  $A \cap B = \emptyset$ ,  $P(A, B) = P(X = x_i, Y = y_j) = 0$ ,
- Marginal probability distribution of  $X$  and  $Y$ :
  - $P(A) = P(X = x_i) = \sum_{j=1}^M P(X = x_i, Y = y_j)$  is the probability of the event  $A$ ,
  - $P(B) = P(Y = y_j) = \sum_{i=1}^N P(X = x_i, Y = y_j)$  is the probability of the event  $B$ .

## Joint probability

### Example

Let  $X$  and  $Y$  be discrete random variables defined by alphabets  $\{x_1, x_2\}$  and  $\{y_1, y_2, y_3, y_4\}$ , respectively.

The sets of events of  $(X, Y)$  can be represented in a joint probability matrix:

$X, Y$	$y_1$	$y_2$	$y_3$	$y_4$
$x_1$	$(x_1, y_1)$	$(x_1, y_2)$	$(x_1, y_3)$	$(x_1, y_4)$
$x_2$	$(x_2, y_1)$	$(x_2, y_2)$	$(x_2, y_3)$	$(x_2, y_4)$

## Joint probability

### Example

Let  $X$  and  $Y$  be discrete random variables defined by alphabets  $\{R, NR\}$  and  $\{S, Su, A, W\}$ , respectively.

$X, Y$	$S$	$Su$	$A$	$W$
$R$	0.15	0.05	0.15	0.20
$NR$	0.10	0.20	0.10	0.05

## Joint probability

### Example

Let  $X$  and  $Y$  be discrete random variables defined by alphabets  $\{R, NR\}$  and  $\{S, Su, A, W\}$ , respectively.

$X, Y$	$S$	$Su$	$A$	$W$
$R$	0.15	0.05	0.15	0.20
$NR$	0.10	0.20	0.10	0.05

Questions:

- Does  $(X, Y)$  define a pdf?  $\Rightarrow$  Yes,  $\sum_{i=1}^2 \sum_{j=1}^4 P(X = x_i, Y = y_j) = 1$ ;
- Is it possible to define the marginal pdf of  $X$ ?  $\Rightarrow$  Yes,  $X = \{R, NR\}$ ,  
 $P(X = R) = \sum_{j=1}^4 P(X = R, Y = y_j) = 0.55$ ,  
 $P(X = NR) = \sum_{j=1}^4 P(X = NR, Y = y_j) = 0.45$ .

## Conditional probability and Bayes rule

Notation: the conditional probability of  $X = x_i$  knowing that  $Y = y_j$  is written as  $P(X = x_i | Y = y_j)$ .

### Definition (Bayes rule)

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

### Properties

- $\sum_{k=1}^N P(X = x_k | Y = y_j) = 1$ ;
- $P(X = x_i | Y = y_j) \neq P(Y = y_j | X = x_i)$ .

## Conditional probability and Bayes rule

### Example

Let  $X$  and  $Y$  be discrete random variables defined by alphabets  $\{R, NR\}$  and  $\{S, Su, A, W\}$ , respectively.

$X, Y$	$S$	$Su$	$A$	$W$
$R$	0.15	0.05	0.15	0.20
$NR$	0.10	0.20	0.10	0.05

Question:

What is the conditional probability distribution of  $P(X = x_i | Y = S)$ ?

## Conditional probability and Bayes rule

### Example

Let  $X$  and  $Y$  be discrete random variables defined by alphabets  $\{R, NR\}$  and  $\{S, Su, A, W\}$ , respectively.

$X, Y$	$S$	$Su$	$A$	$W$
$R$	0.15	0.05	0.15	0.20
$NR$	0.10	0.20	0.10	0.05

Question:

What is the conditional probability distribution of  $P(X = x_i | Y = S)$ ?

$P(Y = S) = \sum_{i=1}^2 P(Y = y_i) = 0.25$ , and from Bayes

$P(X = R | Y = S) = \frac{0.15}{0.25}$  and  $P(X = NR | Y = S) = \frac{0.10}{0.25}$

## Statistical independence of two random variables

### Definition (Independence)

Two discrete random variables are independent if the joint pdf is equal to the product of the marginal pdfs:

$$P(X = x_i, Y = y_j) = P(X = x_i) \times P(Y = y_j) \quad \forall i \text{ and } j.$$

Remark: If  $X$  and  $Y$  independent,  $P(X = x_i | Y = y_j) = P(X = x_i)$  (From Bayes).



While independence of a set of random variables implies independence of any subset, **the converse is not true**. In particular, random variables can be pairwise independent but not independent as a set.

## Linear correlation coefficient (Pearson)

### Definition (linear correlation coefficient)

The degree of linearity between two discrete random variables  $X$  and  $Y$  is given by the linear correlation coefficient  $r$ :

$$r(X, Y) \stackrel{\text{def}}{=} \frac{\text{COV}(X, Y)}{\sigma_X \sigma_Y}$$

$\text{COV}(X, Y)$  is the covariance given by  $\frac{1}{N} \sum (X - \bar{X})(Y - \bar{Y})$

$\sigma_X$  is the standard deviation given by  $\sqrt{\frac{\sum (X - \bar{X})^2}{N}}$ .

### Properties

- $r(X, Y) \in [-1, 1]$ ;  $|r(X, Y)| \approx 1$  means that  $X$  and  $Y$  are very correlated;
- $r(T \times X, Y) = r(X, Y)$ , where  $T$  is a linear transform.

## Linear correlation coefficient (Pearson)

Relationship between the regression line and the linear correlation coefficient:

### Regression line

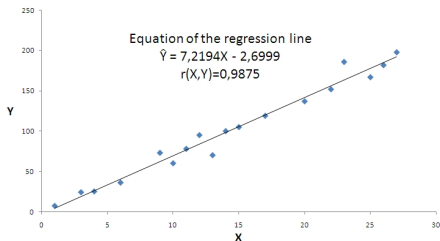
The goal is to find  $\hat{Y} = aX + b$ .  $a$  and  $b$  deduced from the minimization of the MSE:

$$J = \operatorname{argmin}_{(a,b)} \sum (Y - \hat{Y})^2$$

$$\frac{\partial J}{\partial a} = 0 \Rightarrow a = \frac{\operatorname{COV}(X,Y)}{\sigma_X^2}$$

$$\frac{\partial J}{\partial b} = 0 \Rightarrow b = \bar{Y} - a\bar{X}$$

The slope  $a$  is equal to  $r$ , only when  $\sigma_X = \sigma_Y$  (standardised variables).



## Spearman's rank correlation coefficient

### Definition (Spearman's rank correlation coefficient)

The degree of linearity between two discrete random variables  $X$  and  $Y$  is given by the linear correlation coefficient  $r$ :

$$\rho(X, Y) \stackrel{\text{def}}{=} 1 - \frac{6 \sum d_i^2}{n^2(n-1)}$$

where  $d_i = x_i - y_i$ . The  $n$  data  $X_i$  and  $Y_i$  are converted to ranks  $x_i$  and  $y_i$ .

To illustrate the difference between Pearson and Spearman correlation coefficient:

$$(0, 1), (10, 100), (101, 500), (102, 2000)$$

- Pearson correlation is 0.456;
- Spearman correlation is 1.

## Correlation does not imply causation

**Correlation does not imply causation** is a phrase used in science and statistics to emphasize that correlation between two variables does not **automatically** imply that one causes the other.

### Example (Ice cream and drowning)

As ice cream sales increase, the rate of drowning deaths increases sharply.  
Therefore, **ice cream causes drowning**.

The aforementioned example fails to recognize the importance of time in relationship to ice cream sales. Ice cream is sold during the summer months at a much greater rate, and it is during the summer months that people are more likely to engage in activities involving water, such as swimming. The increased drowning deaths are simply caused by more exposure to water based activities, not ice cream.

From [http://en.wikipedia.org/wiki/Correlation\\_does\\_not\\_imply\\_causation](http://en.wikipedia.org/wiki/Correlation_does_not_imply_causation)

# Information Theory

- 1 Introduction
- 2 Statistical signal modelling
- 3 Amount of information**
  - Self-Information
  - Entropy definition
  - Joint information, joint entropy
  - Conditional information, conditional entropy
  - Mutual information
  - Venn's diagram
- 4 Discrete source
- 5 Discrete channel
- 6 Shannon's theorem

## Claude Shannon (1916-1981) and communication theory



For Shannon a message is **very informative** if the chance of **its occurrence is small**. If, in contrast, a message is very predictable, then it has a small amount of information. One is not surprised to receive it.

A measure of the amount of information in a message.

## Self-Information

Let  $X$  be a discrete random variable defined by the alphabet  $\{x_1, \dots, x_N\}$  and the probability density  $\{p(X = x_1), \dots, p(X = x_N)\}$ .

How to measure the amount of information provided by an event  $A, X = x_i$ ?

Definition (Self-Information proposed by Shannon)

$$I(A) \stackrel{\text{def}}{=} -\log_2 p(A) \Leftrightarrow I(X = x_i) \stackrel{\text{def}}{=} -\log_2 p(X = x_i) \text{ Unit: bit/symbol.}$$

Properties

- $I(A) \geq 0$ ;
- $I(A) = 0$  if  $p(A) = 1$ ;
- if  $p(A) < p(B)$  then  $I(A) > I(B)$ ;
- $p(A) \rightarrow 0, I(A) \rightarrow +\infty$ .

The self-information can be measured in **bits, nits, or hartleys** depending on the **base of the logarithm**.

## Shannon entropy

Shannon introduced the entropy rate, a quantity that measured a source's information production rate.

### Definition (Shannon Entropy)

The entropy of a discrete random variable  $X$  defined by the alphabet  $\{x_1, \dots, x_N\}$  and the probability density  $\{p(X = x_1), \dots, p(X = x_N)\}$  is given by:

$$H(X) = E\{I(X)\} = - \sum_{i=1}^N p(X = x_i) \log_2(p(X = x_i)), \text{ unit: bit/symbol.}$$

The entropy  $H(X)$  is a measure of the amount of *uncertainty*, a *measure of surprise* associated with the value of  $X$ .

Entropy gives the average number of bits per symbol to represent  $X$

### Properties

- $H(X) \geq 0$ ;
- $H(X) \leq \log_2 N$  (equality for a uniform probability distribution).

## Shannon entropy

### Example

- Example 1: The value of  $p(0)$  is highly predictable, the entropy (amount of *uncertainty*) is zero.

	$p(0)$	$p(1)$	$p(2)$	$p(3)$	$p(4)$	$p(5)$	$p(6)$	$p(7)$	Entropy (bits/symbol)
Example1	1.0	0	0	0	0	0	0	0	0.00

## Shannon entropy

### Example

- Example 1: The value of  $p(0)$  is highly predictable, the entropy (amount of *uncertainty*) is zero;
- Example 2: This is a probability distribution of Bernoulli  $\{p, 1 - p\}$ .

	$p(0)$	$p(1)$	$p(2)$	$p(3)$	$p(4)$	$p(5)$	$p(6)$	$p(7)$	Entropy (bits/symbol)
Example1	1.0	0	0	0	0	0	0	0	0.00
Example2	0	0	0.5	0.5	0	0	0	0	1.00

## Shannon entropy

### Example

- Example 1: The value of  $p(0)$  is highly predictable, the entropy (amount of *uncertainty*) is zero;
- Example 2: This is a probability distribution of Bernoulli  $\{p, 1 - p\}$ ;
- Example 3: Uniform probability distribution ( $P(X = x_i) = \frac{1}{M}$ , with  $M = 4$ ,  $i \in \{2, \dots, 5\}$ ).

	$p(0)$	$p(1)$	$p(2)$	$p(3)$	$p(4)$	$p(5)$	$p(6)$	$p(7)$	Entropy (bits/symbol)
Example1	1.0	0	0	0	0	0	0	0	0.00
Example2	0	0	0.5	0.5	0	0	0	0	1.00
Example3	0	0	0.25	0.25	0.25	0.25	0	0	2.00

## Shannon entropy

### Example

- Example 1: The value of  $p(0)$  is highly predictable, the entropy (amount of *uncertainty*) is zero;
- Example 2: This is a probability distribution of Bernoulli  $\{p, 1 - p\}$ ;
- Example 3: Uniform probability distribution ( $P(X = x_i) = \frac{1}{M}$ , with  $M = 4$ ,  $i \in \{2, \dots, 5\}$ );
- Example 4: -;

	$p(0)$	$p(1)$	$p(2)$	$p(3)$	$p(4)$	$p(5)$	$p(6)$	$p(7)$	Entropy (bits/symbol)
Example1	1.0	0	0	0	0	0	0	0	0.00
Example2	0	0	0.5	0.5	0	0	0	0	1.00
Example3	0	0	0.25	0.25	0.25	0.25	0	0	2.00
Example4	0.06	0.23	0.30	0.15	0.08	0.06	0.06	0.06	2.68

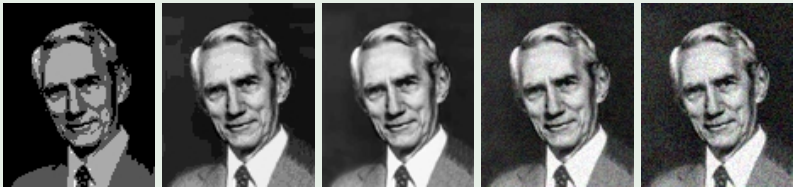
## Shannon entropy

### Example

- Example 1: The value of  $p(0)$  is highly predictable, the entropy (amount of *uncertainty*) is zero;
- Example 2: This is a probability distribution of Bernoulli  $\{p, 1 - p\}$ ;
- Example 3: Uniform probability distribution ( $P(X = x_i) = \frac{1}{M}$ , with  $M = 4$ ,  $i \in \{2, \dots, 5\}$ );
- Example 4: -;
- Example 5: Uniform probability distribution ( $P(X = x_i) = \frac{1}{N}$ , with  $N = 8$ ,  $i \in \{0, \dots, 7\}$ )

	$p(0)$	$p(1)$	$p(2)$	$p(3)$	$p(4)$	$p(5)$	$p(6)$	$p(7)$	Entropy (bits/symbol)
Example1	1.0	0	0	0	0	0	0	0	0.00
Example2	0	0	0.5	0.5	0	0	0	0	1.00
Example3	0	0	0.25	0.25	0.25	0.25	0	0	2.00
Example4	0.06	0.23	0.30	0.15	0.08	0.06	0.06	0.06	2.68
Example5	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	3.00

## Example (Shannon entropy)



(a)  $H=1.39\text{bit/s}$  (b)  $H=3.35\text{bit/s}$  (c)  $H=6.92\text{bit/s}$  (d)  $H=7.48\text{bit/s}$  (e)  $H=7.62\text{bit/s}$

- (c) Original picture (256 levels);
- (a) Original picture (4 levels);
- (b) Original picture (16 levels);
- (d) Original picture + uniform noise;
- (e) Original picture + uniform noise (more than previous one).

## Joint information / joint entropy

### Definition (Joint information)

Let  $X$  and  $Y$  be discrete random variables defined by alphabet  $\{x_1, \dots, x_N\}$  and  $\{y_1, \dots, y_M\}$ , respectively.

The joint information of two events  $(X = x_i)$  and  $(Y = y_j)$  is defined by

$$I(X = x_i, Y = y_j) = -\log_2(p(X = x_i, Y = y_j)).$$

### Definition (Joint entropy)

The joint entropy of the two discrete random variables  $X$  and  $Y$  is given by:

$$H(X, Y) = E\{I(X = x_i, Y = y_j)\}.$$

$$H(X, Y) = -\sum_{i=1}^N \sum_{j=1}^M p(X = x_i, Y = y_j) \log_2(p(X = x_i, Y = y_j))$$

## Joint information / joint entropy

Demonstrate that the joint entropy  $H(X, Y)$  is equal to  $H(X) + H(Y)$ , if  $X$  and  $Y$  are statistically independent:

$$\begin{aligned}
 H(X, Y) &= - \sum_{x \in X} \sum_{y \in Y} P(x, y) \log P(x, y) \\
 &= - \sum_{x \in X} \sum_{y \in Y} P(x)P(y) \log P(x)P(y) \\
 &= - \sum_{x \in X} \sum_{y \in Y} P(x)P(y) [\log P(x) + \log P(y)] \\
 &= - \sum_{x \in X} \sum_{y \in Y} P(x)P(y) \log P(x) - \sum_{y \in Y} \sum_{x \in X} P(x)P(y) \log P(y) \\
 &= - \sum_{x \in X} P(x) \log P(x) \sum_{y \in Y} P(y) - \sum_{y \in Y} P(y) \log P(y) \sum_{x \in X} P(x) \\
 &= - \sum_{x \in X} P(x) \log P(x) - \sum_{y \in Y} P(y) \log P(y) \\
 H(X, Y) &= H(X) + H(Y)
 \end{aligned}$$

## Joint information/joint entropy

In the same vein, you should know how to demonstrate that:

$$H(X, Y) \leq H(X) + H(Y) \text{ (Equality if } X \text{ and } Y \text{ independant).}$$

This property is called sub-Additivity. It means that two systems, considered together, can never have more entropy than the sum of the entropy in each of them.

## Conditional information/Conditional entropy

### Definition (Conditional information)

The conditional information ...  $I(X = x_i | Y = y_j) = -\log_2(p(X = x_i | Y = y_j))$

### Definition (Conditional entropy)

The conditional entropy of  $Y$  given the random variable  $X$ :

$$H(Y|X) = \sum_{i=1}^N p(X = x_i) H(Y|X = x_i)$$

$$H(Y|X) = \sum_{i=1}^N \sum_{j=1}^M p(X = x_i, Y = y_j) \log_2 \frac{1}{p(Y = y_j | X = x_i)}$$

The conditional entropy  $H(Y|X)$  is the amount of uncertainty remaining about  $Y$  after  $X$  is known.

#### Remarks:

- We always have  $H(Y|X) \leq H(Y)$  (The knowledge reduces the uncertainty);
- Entropy chain rule:  $H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$  (From Bayes);

## Mutual information

### Definition (Mutual information)

- The Mutual information of two events  $X = x_i$  and  $Y = y_j$  is defined as  $I(X = x_i; Y = y_j) = -\log_2 \frac{p(X=x_i)p(Y=y_j)}{p(X=x_i, Y=y_j)}$
- The Mutual information of two random variables  $X$  and  $Y$  is defined as

$$I(X; Y) = -\sum_{i=1}^N \sum_{j=1}^M p(X = x_i, Y = y_j) \log_2 \frac{p(X=x_i)p(Y=y_j)}{p(X=x_i, Y=y_j)}$$

The mutual information  $I(X; Y)$  measures the information that  $X$  and  $Y$  share...



Be careful  
 to the  
 notation,  
 $I(X, Y) \neq$   
 $I(X; Y)$

### Properties

- Symmetry:  $I(X = x_i; Y = y_j) = I(Y = y_j; X = x_i)$ ;
- $I(X; Y) \geq 0$ ; zero if and only if  $X$  and  $Y$  are independent variables;
- $H(X|X) = 0 \Rightarrow H(X) = I(X; X) \Rightarrow I(X; X) \geq I(X; Y)$ .

## Mutual information

### Mutual information and dependency

The mutual information can be expressed as :

$$I(X; Y) = D(p(X = x_i, Y = y_j) || p(X = x_i)p(Y = y_j))$$

where,

- $p(X = x_i, Y = y_j)$  joint pdf of  $X$  and  $Y$ ;
- $p(X = x_i)$  and  $p(Y = y_j)$  marginal probability distribution of  $X$  and  $Y$ , respectively;
- $D(.||.)$  the divergence of Kullback-Leibler.

Remarks regarding the KL-divergence:

- $D(p||q) = -\sum_{i=1}^N p(X = x_i) \log_2 \frac{p(X=x_i)}{p(Q=q_i)}$ ,  $Q$  random variable  $\{q_1, \dots, q_N\}$ ;
- This is a measure of divergence between two pdfs, not a distance;
- $D(p||q) = 0$ , if and only if the two pdfs are strictly the same.

## Example (Mutual information)

Mutual information between the original picture of C. Shannon and the following ones:



(a)  $I=2.84$



(b)  $I=2.54$



(c)  $I=2.59$



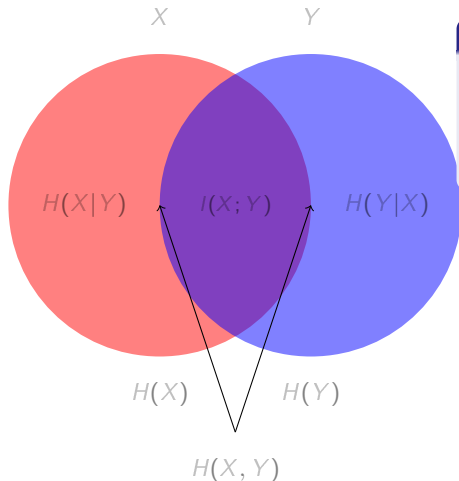
(d)  $I=2.41$



(e)  $I=1.97$

- (a) Original picture + uniform noise;
- (b) Original picture + uniform noise (more than previous one);
- (c) Huffman;
- (d) Einstein;
- (e) A landscape.

## Venn's diagram



We retrieve:

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

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

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

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

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

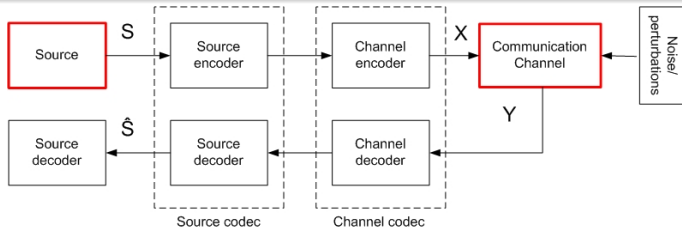
# Information Theory

- 1 Introduction
- 2 Statistical signal modelling
- 3 Amount of information
- 4 Discrete source**
  - Introduction
  - Parameters of a discrete source
  - Discrete memoryless source
  - Extension of a discrete memoryless source
  - Discrete source with memory (Markov source)
- 5 Discrete channel
- 6 Shannon's theorem

## Introduction

### Remind of the goal

- To transmit an information at the minimum rate for a given quality;
- Seminal work of Claude Shannon (1948)[Shannon,48].



## Parameters of a discrete source

### Definition (Alphabet)

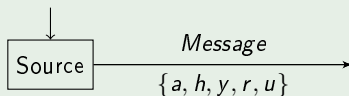
An alphabet  $\mathcal{A}$  is a set of data  $\{a_1, \dots, a_N\}$  that we might wish to encode.

### Definition (discrete source)

A source is defined as a discrete random variable  $S$  defined by the alphabet  $\{s_1, \dots, s_N\}$  and the probability density  $\{p(S = s_1), \dots, p(S = s_N)\}$ .

### Example (Text)

Alphabet =  $\{a, \dots, z\}$



## Discrete memoryless source

### Definition (Discrete memoryless source)

A discrete source  $S$  is memoryless if the symbols of the source alphabet are independent and identically distributed:

$$p(S = s_1, \dots, S = s_N) = \prod_{i=1}^N p(S = s_i)$$

### Remarks:

- Entropie:  $H(S) = -\sum_{i=1}^N p(S = s_i) \log_2 p(S = s_i)$  bit;
- Particular case of a uniform source:  $H(S) = \log_2 N$ .

## Extension of a discrete memoryless source

Rather than considering individual symbols, more useful to deal with blocks of symbols.

Let  $S$  be a discrete source with an alphabet of size  $N$ . The output of the source is grouped into blocks of  $K$  symbols. The new source, called  $S^K$ , is defined by an alphabet of size  $N^K$ .

**Definition (Discrete memoryless source,  $K^{\text{th}}$  extension of a source  $S$ )**

If the source  $S^K$  is the  $K^{\text{th}}$  extension of a source  $S$ , the entropy per extended symbols of  $S^K$  is  $K$  times the entropy per individual symbol of  $S$ :

$$H(S^K) = K \times H(S)$$

Remark:

the probability of a symbol  $s_i^K = (s_{i_1}, \dots, s_{i_K})$  from the source  $S^K$  is given by  $p(s_i^K) = \prod_{j=1}^K p(s_{i_j})$ .

## Discrete source with memory (Markov source)

### Discrete memoryless source

This is not realistic!

Successive symbols are not completely independent of one another...

- in a picture: a pel ( $S_0$ ) depends statistically on the *previous* pels.



200	210	207	205	200	202
201	205	199	199	200	201
202	203	203	201	200	204
200	210	207	205	200	202

-	-	-	-	-	-
-	-	$s_5$	$s_4$	$s_3$	$s_2$
-	-	$s_1$	$s_0$	-	-
-	-	-	-	-	-

This dependence is expressed by the conditionnal probability

$$p(S_0|S_1, S_2, S_3, S_4, S_5).$$

$$p(S_0|S_1, S_2, S_3, S_4, S_5) \neq p(S_0)$$

- in the langage (french):  $p(S_k = u) \leq p(S_k = e)$ ,  
 $p(S_k = u|S_{k-1} = q) \gg p(S_k = e|S_{k-1} = q)$ ;

## Discrete source with memory (Markov source)

### Definition (Discrete source with memory)

A discrete source with memory of order  $N$  ( $N^{\text{th}}$  order Markov) is defined as:

$$p(S_k | S_{k-1}, S_{k-2}, \dots, S_{k-N})$$

The entropy is given by:

$$H(S) = H(S_k | S_{k-1}, S_{k-2}, \dots, S_{k-N})$$

### Example (One dimensional Markov model)

The pel value  $S_0$  depends statistically only on the pel value  $S_1$ .



$Q$



85	85	170	0	255
85	85	85	170	255

$$H(X) = 1.9 \text{ bit/symb}, H(Y) = 1.29, H(X, Y) = 2.15, H(X|Y) = 0.85$$

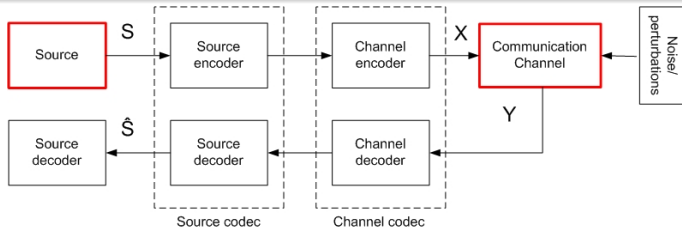
# Information Theory

- 1 Introduction
- 2 Statistical signal modelling
- 3 Amount of information
- 4 Discrete source
- 5 Discrete channel**
  - Introduction
  - Parameters of a discrete channel
  - Memoryless channel
  - Channel with memory
  - A taxonomy
- 6 Shannon's theorem

## Introduction

### Remind of the goal

- To transmit an information at the minimum rate for a given quality;
- Seminal work of Claude Shannon (1948)[Shannon,48].



## Parameters of a discrete channel



### Definition (Channel)

A channel transmits as best as it can symbols  $X_t$ . Symbols  $Y_t$  are produced (possibly corrupted). The subscript  $t$  indicates the time.

The channel is featured by the following conditional probabilities:

$$p(Y_t = y_j | X_t = x_i, X_t = x_{i-1} \dots)$$

The output symbol can be dependent on several input symbols.

### Capacity

The channel is characterized by its capacity  $C$  bits per symbol. It is an upper bound on the bit rate that can be accommodated by the channel.

$$C = \max_{p(X)} I(X; Y)$$

## Memoryless channel

### Definition (Memoryless channel)

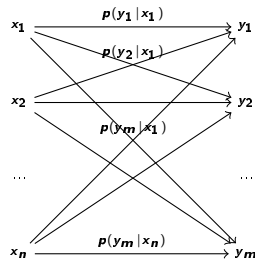
The channel is memoryless when the conditional probability  $p(Y|X)$  is independent of previously transmitted channel symbols:  $p(Y = y_j | X = x_i)$ .

Notice that the subscript  $t$  is no more required...

Transition matrix  $\Pi$ , size  $(N, M)$ :

$$\Pi(X, Y) = \begin{pmatrix} p(Y=y_1|X=x_1) & \dots & p(Y=y_m|X=x_1) \\ \vdots & \ddots & \vdots \\ p(Y=y_1|X=x_n) & \dots & p(Y=y_m|X=x_n) \end{pmatrix}$$

Graphical representation:



## Ideal Channel

### Ideal Channel or noiseless channel

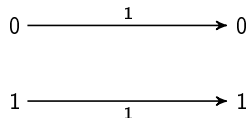
Perfect correspondance between the input and the output. Each transmitted bit is received without error.

$$p(Y = y_j | X = x_i) = \delta_{x_i y_j}$$

$$\text{with, } \delta_{xy} = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$$

The channel capacity is  $C = \max_{p(x)} I(X; Y) = 1 \text{ bit}$ .

$$\Pi(X, Y) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



## Binary Symmetric Channel (BSC)

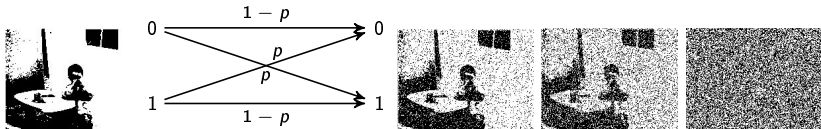
### Definition (Binary Symmetric Channel (BSC))

A binary symmetric channel has binary input and binary output ( $X = \{0, 1\}$ ,  $Y = \{0, 1\}$ ).

$$C = 1 - H(p) \text{ bits.}$$

$p$  the probability that the symbol is modified at the output.

$$\Pi(X, Y) = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$



From left to right: original binary signal, result for  $p = 0.1$ ,  $p = 0.2$  and  $p = \frac{1}{2}$  (the channel has a null capacity!).

## Binary Erasure Channel (BEC)

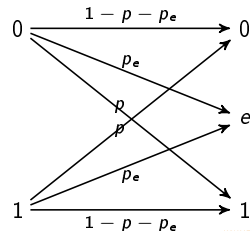
### Definition (Binary Erasure Channel (BEC))

A binary erasure channel has binary input and ternary output ( $X = \{0, 1\}$ ,  $Y = \{0, 1, e\}$ ). The third symbol is called an erasure (complete loss of information about an input bit).

$$C = 1 - p_e - p \text{ bits.}$$

$p_e$  is the probability to receive the symbol  $e$ .

$$\Pi(X, Y) = \begin{pmatrix} 1 - p - p_e & p_e & p \\ p & p_e & 1 - p - p_e \end{pmatrix}$$

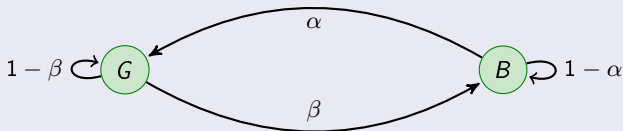


## Channel with memory

### Gilbert-Elliott model

The channel is a binary symmetric channel with memory determined by a two states Markov chain. A channel has two states:  $G$ , good state and  $B$ , bad state.

$\alpha$  and  $\beta$  are the transition probabilities



Transition matrix to go from one state to another:  $S_t = \{G, B\}$ .

$$S_{t-1} \begin{pmatrix} 1 - \beta & \alpha \\ \beta & 1 - \alpha \end{pmatrix}$$

## Discrete channel: a taxonomy

### Discrete channel

- Lossless channel:  $H(X|Y) = 0$ , the input is defined by the output;
- Deterministic channel:  $H(Y|X)$ , the output defined the input;
- Distortionless channel:  $H(X) = H(Y) = I(X; Y)$ , deterministic and lossless;
- Channel with a null capacity:  $I(X; Y) = 0$ .

# Information Theory

- 1 Introduction
- 2 Statistical signal modelling
- 3 Amount of information
- 4 Discrete source
- 5 Discrete channel
- 6 Shannon's theorem**
  - Source code
  - Kraft inequality
  - Higher bound of entropy
  - Source coding theorem
  - Rabboni-Jones extension
  - Channel coding theorem
  - Source/Channel theorem

- 7 Summary

## Source code

### Definition (Source code)

A source code  $C$  for a random variable  $X$  is a mapping from  $x \in \mathcal{X}$  to  $\{0, 1\}^*$ . Let  $c_i$  denotes the code word for  $x_i$  and  $l_i$  denote the length of  $c_i$ .

$\{0, 1\}^*$  is the set of all finite binary string.

### Definition (Prefix code)

A code is called a prefix code (instantaneous code) if no code word is a prefix of another code word

Not required to wait for the whole message to be able to decode it.

## Kraft inequality

### Definition (Kraft inequality)

A code  $C$  is instantaneous if it satisfies the following inequality:

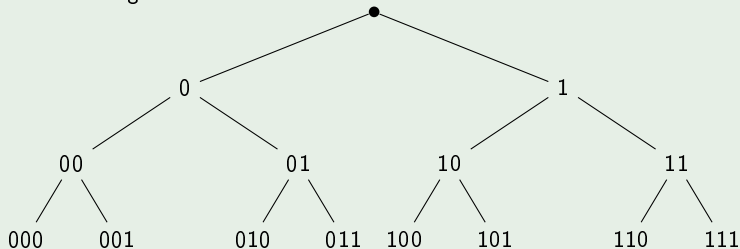
$$\sum_{i=1}^N 2^{-l_i} \leq 1$$

with,  $l_i$  the length of code word length  $i$

## Kraft inequality

### Example (Illustration of the Kraft inequality using a coding tree)

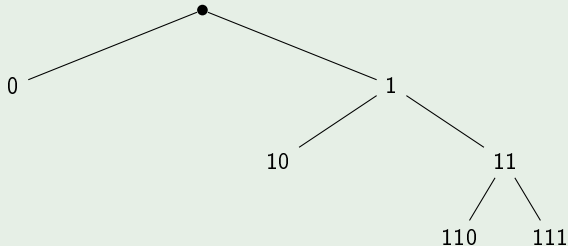
The following tree contains all three-bit codes:



## Kraft inequality

### Example (Illustration of the Kraft inequality using a coding tree)

The following tree contains a prefix code. We decide to use the code word 0 and 10.



The remaining leaves constitute a prefix code:

$$\sum_{i=1}^4 2^{-l_i} = 2^{-1} + 2^{-2} + 2^{-3} + 2^{-3} = 1$$

## Higher bound of entropy

Let  $S$  a discrete source defined by the alphabet  $\{s_1, \dots, s_N\}$  and the probability density  $\{p(S = s_1), \dots, p(S = s_N)\}$ .

### Definition (Higher bound of entropy)

$$H(S) \leq \log_2 N$$

### Interpretation

- the entropy is limited by the size of the alphabet;
- a source with a uniform pdf provides the highest entropy.

## Source coding theorem

Let  $S$  a discrete source defined by the alphabet  $\{s_1, \dots, s_N\}$  and the probability density  $\{p(S = s_1), \dots, p(S = s_N)\}$ . Each symbol  $s_i$  is coded with a length  $l_i$  bits:

Definition (Source coding theorem or First Shannon's theorem)

$$H(S) \leq \bar{l}_C \text{ with } \bar{l}_C = \sum_{i=1}^N p_i l_i$$

The entropy of the source gives the limit of **the lossless compression**. We can not encode the source with less than  $H(S)$  bit per symbol. **The entropy of the source is the lower-bound**.

Warning....

$\{l_i\}_{i=1, \dots, N}$  must satisfy Kraft's inequality.

Remarks:

- $\bar{l}_C = H(S)$ , when  $l_i = -\log_2 p(X = x_i)$ .

## Source coding theorem

### Definition (Source coding theorem (bis))

Whatever the source  $S$ , there exist an instantaneous code  $C$ , such that

$$H(S) \leq \bar{l}_C < H(S) + 1$$

The upper bound is equal to  $H(S) + 1$ , simply because the Shannon information gives a fractionnal value.

## Source coding theorem

### Example

Let  $X$  a random variable with the following probability density. The optimal code lengths are given by the self-information:

$X$	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
$P(X = x_i)$	0.25	0.25	0.2	0.15	0.15
$I(X = x_i)$	2.0	2.0	2.3	2.7	2.7

The entropy  $H(X)$  is equal to 2.2855 bits. The source coding theorem gives:  

$$2.2855 \leq \bar{l} < 3.2855$$

## Rabani-Jones extension

Symbols can be coded in blocks of source samples instead of one at a time (block coding). In this case, further bit-rate reductions are possible.

### Definition (Rabani-Jones extension)

Let  $S$  be an ergodic source with an entropy  $H(S)$ . Consider encoding blocks of  $N$  source symbols at a time into binary codewords.

For any  $\delta > 0$ , it is possible to construct a code that the average number of bits per original source symbol  $\bar{l}_C$  satisfies:

$$H(S) \leq \bar{l}_C < H(S) + \delta$$

Remarks:

- Any source can be **losslessly** encoded with a code very close to the source entropy in bits;
- There is a high benefit to increase the value  $N$ ;
- But, the number of symbols in the alphabet becomes very high. Example: block of  $2 \times 2$  pixels (coded on 8 bits) leads to  $256^4$  values per block...

## Channel coding theorem

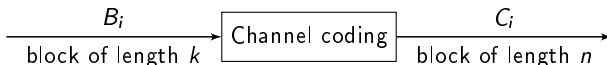
Let a discrete memoryless channel of capacity  $C$ . The channel coding transform the messages  $\{b_1, \dots, b_k\}$  into binary codes having a length  $n$ .

### Definition (Transmission rate)

The transmission rate  $R$  is given by:

$$R \stackrel{\text{def}}{=} \frac{k}{n}$$

$R$  is the amount of information stemming from the symbols  $b_i$  per transmitted bits.





## Shannon's theorem

Source/Channel theorem

Let a noisy channel having a capacity  $C$  and a source  $S$  having an entropy  $H$ .

### Definition (Source/Channel theorem)

- if  $H < C$  it is possible to transmit information nearly without error. Shannon showed that it was possible to do that by making a source coding followed by a channel coding;
- if  $H \geq C$ , the transmission cannot be done with an arbitrarily small probability.

# Information Theory

- 1 Introduction
- 2 Statistical signal modelling
- 3 Amount of information
- 4 Discrete source
- 5 Discrete channel
- 6 Shannon's theorem
- 7 Summary**

## YOU MUST KNOW

Let  $X$  a random variable defined by  $\mathcal{X} = \{x_1, \dots, x_N\}$  and the probabilities  $\{p_{x_1}, \dots, p_{x_N}\}$ .  
 Let  $Y$  a random variable defined by  $\mathcal{Y} = \{y_1, \dots, y_N\}$  and the probabilities  $\{p_{y_1}, \dots, p_{y_N}\}$ .

- $\sum_{i=1}^N p_{x_i} = 1$
- Independence:  $p(X = x_1, \dots, X = x_N) = \prod_{i=1}^N p(X = x_i)$
- Bayes rule:  $p(X = x_i | Y = y_j) = \frac{p(X=x_i, Y=y_j)}{p(Y=y_j)}$
- Self information:  $I(X = x_i) = -\log_2 p(X = x_i)$
- Mutual information:  

$$I(X; Y) = -\sum_{i=1}^N \sum_{j=1}^M p(X = x_i, Y = y_j) \log_2 \frac{p(X=x_i)p(Y=y_j)}{p(X=x_i, Y=y_j)}$$
- Entropy:  $H(X) = -\sum_{i=1}^N p(X = x_i) \log_2 p(X = x_i)$
- Conditional entropy of  $Y$  given  $X$ :  $H(Y|X) = \sum_{i=1}^N p(X = x_i) H(Y|X = x_i)$
- Higher Bound of entropy:  $H(X) \leq \log_2 N$
- Limit of the lossless compression :  $H(X) \leq \bar{I}_C, \bar{I}_C = \sum_{i=1}^N p_{x_i} l_i$

Suggestion for further reading...

[Shannon,48] C.E. Shannon. A Mathematical Theory of Communication. Bell System Technical Journal, 27, 1948.