ELEMENT OF INFORMATION THEORY

O. Le Meur
olemeur@irisa.fr

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

September 2009
ELEMENTS OF INFORMATION THEORY

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

2 Statistical signal modelling

3 Amount of information

4 Discrete source

5 Shannon’s theorem

6 Summary
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.
Three major research axes:

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 theirs values in a predefined alphabet.

**Example**
Source: a camera  
Message: a picture  
Symbols: RGB coefficients  
alphabet = (0, ..., 255)

**Example**
Source: book  
Message: a text  
Symbols: letters  
alphabet = (a, ..., z)
**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**

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

3 Amount of information

4 Discrete source

5 Shannon’s theorem

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

**Definition (Alphabet)**

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

**Definition (Random Variable)**

A discrete random variable $X$ is defined by an alphabet $\mathcal{A} = \{x_1, \ldots, x_N\}$ and a probability distribution $\{p_1, \ldots, p_N\}$, 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 $P_X(i)$. 
Definition (Joint probability)

Let \( X \) and \( Y \) be discrete random variables defined by alphabets \( \{x_1, \ldots, x_N\} \) and \( \{y_1, \ldots, y_M\} \), 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 \( B \).
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:

<table>
<thead>
<tr>
<th>$X, Y$</th>
<th>$y_1$</th>
<th>$y_2$</th>
<th>$y_3$</th>
<th>$y_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$(x_1, y_1)$</td>
<td>$(x_1, y_2)$</td>
<td>$(x_1, y_3)$</td>
<td>$(x_1, y_4)$</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$(x_2, y_1)$</td>
<td>$(x_2, y_2)$</td>
<td>$(x_2, y_3)$</td>
<td>$(x_2, y_4)$</td>
</tr>
</tbody>
</table>
Example

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

<table>
<thead>
<tr>
<th>$X, Y$</th>
<th>$S$</th>
<th>$Su$</th>
<th>$A$</th>
<th>$W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td>0.15</td>
<td>0.05</td>
<td>0.15</td>
<td>0.20</td>
</tr>
<tr>
<td>$NR$</td>
<td>0.10</td>
<td>0.20</td>
<td>0.10</td>
<td>0.05</td>
</tr>
</tbody>
</table>
Example

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

<table>
<thead>
<tr>
<th>$X, Y$</th>
<th>$S$</th>
<th>$Su$</th>
<th>$A$</th>
<th>$W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td>0.15</td>
<td>0.05</td>
<td>0.15</td>
<td>0.20</td>
</tr>
<tr>
<td>$NR$</td>
<td>0.10</td>
<td>0.20</td>
<td>0.10</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Questions:
- Does $(X, Y)$ define a pdf? => 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 $X$? => 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**

- $\sum_{k=1}^{N} P(X = x_k|Y = y_j) = 1$;
- $P(X = x_i|Y = y_j) \neq P(Y = y_j|X = x_i)$. 
Let $X$ and $Y$ be discrete random variables defined by alphabets $\{R, \text{NR}\}$ and $\{S, \text{Su}, A, W\}$, respectively.

<table>
<thead>
<tr>
<th></th>
<th>$S$</th>
<th>$\text{Su}$</th>
<th>$A$</th>
<th>$W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td>0.15</td>
<td>0.05</td>
<td>0.15</td>
<td>0.20</td>
</tr>
<tr>
<td>$\text{NR}$</td>
<td>0.10</td>
<td>0.20</td>
<td>0.10</td>
<td>0.05</td>
</tr>
</tbody>
</table>

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

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

\[
\begin{array}{c|cccc}
X, Y & S & Su & A & W \\
\hline
R & 0.15 & 0.05 & 0.15 & 0.20 \\
NR & 0.10 & 0.20 & 0.10 & 0.05 \\
\end{array}
\]

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.
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. Shannon’s theorem

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

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

**Definition (Self-Information proposed by Shannon)**

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

**Properties**

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

#### Definition (Shannon Entropy)

The entropy of a discrete random variable $X$ defined by the alphabet \{\(x_1, \ldots, x_N\)\} and the probability density \(\{p(X = x_1), \ldots, p(X = x_N)\}\) 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).
Example 1: The value of $p(0)$ is highly predictable, the entropy (amount of uncertainty) is zero.

<table>
<thead>
<tr>
<th></th>
<th>p(0)</th>
<th>p(1)</th>
<th>p(2)</th>
<th>p(3)</th>
<th>p(4)</th>
<th>p(5)</th>
<th>p(6)</th>
<th>p(7)</th>
<th>Entropy (bits/symbol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example1</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
</tbody>
</table>
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 $\{p, 1-p\}$.

<table>
<thead>
<tr>
<th></th>
<th>$p(0)$</th>
<th>$p(1)$</th>
<th>$p(2)$</th>
<th>$p(3)$</th>
<th>$p(4)$</th>
<th>$p(5)$</th>
<th>$p(6)$</th>
<th>$p(7)$</th>
<th>Entropy (bits/symbol)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Example 1</strong></td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>Example 2</strong></td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
</tr>
</tbody>
</table>
## 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 \( M = 4, \ i \in \{2, \ldots, 5\} \).

<table>
<thead>
<tr>
<th></th>
<th>( p(0) )</th>
<th>( p(1) )</th>
<th>( p(2) )</th>
<th>( p(3) )</th>
<th>( p(4) )</th>
<th>( p(5) )</th>
<th>( p(6) )</th>
<th>( p(7) )</th>
<th>Entropy (bits/symbol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>Example 2</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>Example 3</td>
<td>0</td>
<td>0</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>2.00</td>
</tr>
</tbody>
</table>
Introduction
Statistical signal modelling
Amount of information
Discrete source
Shannon’s theorem
Summary

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 \( M = 4, i \in \{2, \ldots, 5\} \);
- Example 4: ─;

<table>
<thead>
<tr>
<th>Example</th>
<th>p(0)</th>
<th>p(1)</th>
<th>p(2)</th>
<th>p(3)</th>
<th>p(4)</th>
<th>p(5)</th>
<th>p(6)</th>
<th>p(7)</th>
<th>Entropy (bits/symbol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example1</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>Example2</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>Example3</td>
<td>0</td>
<td>0</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>2.00</td>
</tr>
<tr>
<td>Example4</td>
<td>0.06</td>
<td>0.23</td>
<td>0.30</td>
<td>0.15</td>
<td>0.08</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>2.68</td>
</tr>
</tbody>
</table>
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 \( (P(X = x_i) = \frac{1}{M}, \text{ with } M = 4, i \in \{2, \ldots, 5\}) \).
- Example 4: \(-\).
- Example 5: Uniform probability distribution \( (P(X = x_i) = \frac{1}{N}, \text{ with } N = 8, i \in \{0, \ldots, 7\}) \).

<table>
<thead>
<tr>
<th></th>
<th>( p(0) )</th>
<th>( p(1) )</th>
<th>( p(2) )</th>
<th>( p(3) )</th>
<th>( p(4) )</th>
<th>( p(5) )</th>
<th>( p(6) )</th>
<th>( p(7) )</th>
<th>Entropy (bits/symbol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example1</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>Example2</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>Example3</td>
<td>0</td>
<td>0</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>2.00</td>
</tr>
<tr>
<td>Example4</td>
<td>0.06</td>
<td>0.23</td>
<td>0.30</td>
<td>0.15</td>
<td>0.08</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>2.68</td>
</tr>
<tr>
<td>Example5</td>
<td>( \frac{1}{8} )</td>
<td>( \frac{1}{8} )</td>
<td>( \frac{1}{8} )</td>
<td>( \frac{1}{8} )</td>
<td>( \frac{1}{8} )</td>
<td>( \frac{1}{8} )</td>
<td>( \frac{1}{8} )</td>
<td>( \frac{1}{8} )</td>
<td>3.00</td>
</tr>
</tbody>
</table>
Joint information/joint entropy

**Definition (Joint information)**

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

The joint information of two events $(X = x_i)$ and $(Y = y_i)$ 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:

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

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

Remark: $H(X, Y) \leq H(X) + H(Y)$ (Equality if $X$ and $Y$ independent).
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 \( X \):

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

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

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);
- Sub-Additivity: \( H(X, Y) \leq H(X) + H(Y) \)
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...

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

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 \( p(Y = y_i) \) 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 \( \{q_1, ..., q_N\} \);
- 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.
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 Shannon’s theorem

6 Summary
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 $\{a_1, \ldots, a_N\}$ that we might wish to encode.

**Definition (discrete source)**

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

**Example (Text)**

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

Message $= \{a, h, y, r, u\}$
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:

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

Remarks:

- Entropy: $H(S) = - \sum_{i=1}^{N} p(S = s_i) \log_2 p(S = s_i)$ bit;
- Particular case of a uniform source: $H(S) = \log_2 N$. 
Rather than considering individuals 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, $K^{th}$ extension of a source $S$)**

If the source $S^K$ is the $K^{th}$ 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 $s^K_i = (s_{i1}, ..., s_{iK})$ from the source $S^K$ is given by $p(s^K_i) = \prod_{j=1}^{K} p(s_{ij})$. 

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.

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 language (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 \( N \) (\( N^{th} \) order Markov) is defined as:

\[
p(S_k|S_{k-1}, S_{k-2}, ..., S_{k-N})
\]

The entropy is given by:

\[
H(S) = H(S_k|S_{k-1}, S_{k-2}, ..., S_{k-N})
\]

Example (One dimensional Markov model)

The pel value \( S_0 \) depends statistically only on the pel value \( S_1 \).

\[
\begin{array}{cccccc}
85 & 85 & 170 & 0 & 255 \\
85 & 85 & 85 & 170 & 255 \\
\end{array}
\]

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

Statistical signal modelling

Amount of information

Discrete source

Shannon’s theorem

Kraft inequality

Higher bound of entropy

Source coding theorem

Rabbani-Jones extension

Channel coding theorem

Source/Channel theorem

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.
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 $i$
Example (Illustration of the Kraft inequality using a coding tree)

The following tree contains all three-bit codes:
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.

```
  0
    /
   / 1
  /    /
10 11  
   /    /
  110 111
```

The remaining leaves constitute a prefix code:

\[ \sum_{i=1}^{4} 2^{-i} = 2^{-1} + 2^{-2} + 2^{-3} + 2^{-3} = 1 \]
Let $S$ a discrete source defined by the alphabet $\{s_1, \ldots, s_N\}$ and the probability density $\{p(S = s_1), \ldots, p(S = s_N)\}$.

**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 $\{s_1, ..., s_N\}$ and the probability density $\{p(S = s_1), ..., p(S = s_N)\}$. Each symbol $s_i$ is coded with a length $l_i$ bits:

**Definition (Source coding theorem or First Shannon’s theorem)**

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

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

**Warning....**

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

**Remarks:**

- $\overline{l}_C = H(S)$, when $l_i = -\log_2 p(X = x_i)$. 
Definition (Source coding theorem (bis))

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

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

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

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

<table>
<thead>
<tr>
<th>$X$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(X = x_i)$</td>
<td>0.25</td>
<td>0.25</td>
<td>0.2</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>$I(X = x_i)$</td>
<td>2.0</td>
<td>2.0</td>
<td>2.3</td>
<td>2.7</td>
<td>2.7</td>
</tr>
</tbody>
</table>

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

$2.2855 \leq \bar{l} < 3.2855$
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 (Rabbani-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...
Let a discrete memoryless channel of capacity $C$. The channel coding transform the messages $\{b_1, ..., b_k\}$ into binary codes having a length $n$.

**Definition (Transmission rate)**

The transmission rate $R$ is given by:

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

$R$ is the amount of information stemming from the symbols $b_i$ per transmitted bits.
**Channel coding theorem**

**Example (repetition coder)**

The coder is simply a device repeating \( r \) times a particular bit. Below, example for \( r = 2 \). \( R = \frac{1}{3} \).

\[
\begin{align*}
B_t & \quad |011| \\
\text{Channel coding} & \\
C_t & \quad |00111111|
\end{align*}
\]

This is our basic scheme to communicate with others! We repeat the information...

**Definition (Channel coding theorem or second Shannon’s theorem)**

- \( \forall R < C, \forall p_e > 0 \): it is possible to transmit information nearly without error at any rate below the channel capacity;

- if \( R \geq C \), all codes will have a probability of error greater than a certain positive minimal level, and this level increases with the rate.
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.
YOU MUST KNOW

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

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

- $\sum_{i=1}^{N} p_{x_i} = 1$
- Independence: $p(X = x_1, ..., X = x_N) = \prod_{i=1}^{N} p(X = x_i)$
- 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...