Information Theory - General Information Sources

Eric Filiol

ESIEA - Laval
Laboratoire de cryptologie et de virologie opérationnelles
$(C+V)^O$
filiol@esiea.fr

2013 - 2014

Introduction

- The model of zero-memory source is clearly unrealistic in many cases (natural languages for example) and was a first hand approach.
- The aim is now to extend previous results to a wider class of sources.
- Let us recall that a source $\mathcal{S}$ is an object that emits symbols from a finite alphabets Σ according to a random mechanism.
  - Let $(X_1, X_2, \ldots, X_n)$ a sequence of random elements of Σ emitted by $\mathcal{S}$ where X_i denotes the i^{th} emitted.
- The n-stage entropy $H(X_1, X_2, \ldots, X_n)$ is well defined. We say that $\mathcal{S}$ has entropy $H(\mathcal{S}) = H$ if

$$\lim_{n \to +\infty} \frac{H(X_1, X_2, \ldots, X_n)}{n} \text{ exists and is equal to } H$$

- If $\mathcal{S}$ is just a zero-memory source then this definition reduces to the previous one (left as an exercice).

The Entropy of a General Source

- We would also take the limit of $H(X_n|X_1, X_2, \ldots, X_{n-a})$ as n tends towards the infinity

General Source Entropy

If the source $\mathcal{S}$ is such that

$$\lim_{n \to +\infty} H(X_n|X_1, X_2, \ldots, X_{n-a}) \text{ exists}$$

then

$$\lim_{n \to +\infty} \frac{H(X_1, X_2, \ldots, X_n)}{n}$$

exists and the two limits are equal.

- Note that the converse of this theorem does not hold (see exercices).
- The proof is left as an exercice (hint : use the *Arithmetic mean lemma*).

Agenda

Agenda

Definition of a Stationary Source

- Let us consider non-memoryless sources. To have some control over the amount of dependence between successive elements we are going to consider a reduced case.
- A natural restriction is that of *stationary source* :
  - For any positive integers n and h and for any sequence $(s_1, s_2, \ldots, s_n)$ of symbols from Σ and for any non-negative indices $(i_1, i_2, \ldots, i_n)$ we must have

$$P(X_{i_1} = s_1, \ldots, X_{i_n} = s_n) = P(X_{i_1+h} = s_1, \ldots, X_{i_n+h} = s_n)$$

- This strong restriction means that the probability that a given sequence (e.g a single letter, a bigram of letters...) is emitted is independent of time.
- For natural languages it is a plausible model.

The Entropy of Stationary Sources

Stationary Source Entropy

Any stationary source $\mathcal{S}$ has an entropy which is bounded above by

$$H(\mathcal{S}) \leq \frac{1}{n} H(X_1, X_2, \ldots, X_n)$$

for any positive integer n. Moreover $\lim_{n \to +\infty} H(X_n | X_1, X_2, \ldots, X_{n-1})$ exists and equals the entropy of $\mathcal{S}$.

- The proof depends on a particular property of subadditive functions. A function $\mathbb{Z}^+ \to \mathbb{R}$ is subadditive if, $\forall (x, y) \in \mathbb{Z}^+ \times \mathbb{Z}^+$ we have $f(x + y) \leq f(x) + f(y)$.

Fundamental Lemma on Subadditive Functions (Hille & Phillips, 1957)

If f is a subadditive function on $\mathbb{Z}^+$ then

$$\lim_{n \to +\infty} \frac{f(n)}{n} = \inf_{n \geq 1} \frac{f(n)}{n} \text{ exists and is finite}$$

Agenda

Source Typical Message

- Suppose that we have a source which is a mathematical model of any natural language (e.g. English).
- While there is a probability of emitting a string of 20 consecutive A's, this probability is very low and any typical source will output around 20 % of E's, frequent spaces, few Z's...
- Shannon's idea was then to divide source outputs into two distinct groups : output that are "typical" and those which are not.
- Wlog we will focus on memoryless sources to make the statement and the proof more easily understood.
- Many applications are based on this output message classification (for example bruteforce attack in cryptanalysis).
- Statistical interpretation (compare with the three-σ rule).

Typical Message Theorem

Typical Message Theorem

Let $\mathcal{S}$ be a memoryless source with entropy H. Then, given any $\epsilon > 0$, the set $\Sigma^{(N)}$ of sequences of length N can be divided into two classes :

- a set Π_N such that, if X_N denotes the random output of length N of $\mathcal{S}$ then
$$P(X_n \in \Pi_N) < \epsilon.$$
- The remainder, $\Sigma^{(N)} - \Pi_N$, all of whose members σ_n have "high" probability satisfying the inequality
$$2^{-N.H-A.N^{\frac{1}{2}}} \leq P(X_N = \sigma_N) \leq 2^{-N.H+A.N^{\frac{1}{2}}}$$

where A is a positive constant.

Consequences

- According to the previous Theorem, $\Sigma^{(N)}$ consists of a set oft low probability or atypical sequences (namely Π_N) and a disjoint set of high probability or typical sequences each of which has a probability of occurence approximately $\frac{1}{2}^{NH}$.
- As an immediate consequence, we have the following theorem :

Number of Typical Sequences

The number of typical sequences of length N emitted by a memoryless source of entropy H is

$$2^{NH+o(N)} \text{ as } N \to \infty$$

- Proof left as an exercice.

Agenda

Asymptotic Equipartition Property (AEP)

- The aim is to extend the idea of "typical sequence" to more general sources : dividing the set Σ^N of all possible N-sequences into $T(\Sigma^N)$ (typical group) and $\Pi(\Sigma^N)$ (low-probability group).
- In $T(\Sigma^N)$, each N-sequence should have roughly the same probability 2^{-NH} where H is the source entropy.
- The total probability of $\Pi(\Sigma^N)$ is arbitrary small for sufficiently large N.
- We say that the source $\mathcal{S}$ with entropy H has the *Asymptotic Equipartition Property (AEP)* if for each value of N it is possible to partition the set Σ^N into $T(\Sigma^N)$ and $\Pi(\Sigma^N)$ as stated before.
- Unfortunately, we cannot achieve anything approaching this for general stationary sources (see example : not all stationary sources have the AEP).
- We need one more theoretical tool : the concept of *ergodicity*.

Ergodicity

- For any sequence $s = (s_1, s_2, \ldots, s_a)$ of symbols from Σ and any output $X = (X_1, X_2, \ldots)$ of a source $\mathcal{S}$, we define the *frequency* $f_N(s, X)$ to be the number of times s occurs in the first N terms of the sequence X.
- $s = 001 \qquad X = 0010110011011001$ then $f_{16}(s, X) = 3$.
- Thus $\mathcal{S}$ is said to be *ergodic* if it is stationary and if, for any finite sequence $s = (s_1, s_2, \ldots, s_a)$, we have

$$P(\lim_{N \to +\infty} \frac{f_N(s, X)}{N} = P(X_1 = s_1, X_2 = s_2, \ldots, X_a = s_a)) = 1$$

- Ergodicity demands that $\frac{f_N(s,X)}{N}$ converge (in probability) with probability 1 towards the constant $P(X_1 = s_1, X_2 = s_2, \ldots, X_a = s_a)$.
- Since $\mathcal{S}$ is stationary, $P(X_1 = s_1, X_2 = s_2, \ldots, X_a = s_a)$ can be replaced by $P(X_{b+1} = s_1, X_{b+2} = s_2, \ldots, X_{b+a} = s_a)$ for any integer $b > 0$.

Ergodic Sources

Shannon - McMillan Theorem

Ergodic Sources have the AEP. In other words, let $\mathcal{S}$ be an ergodic source with entropy H. Then for any $\epsilon > 0$, there exists $N_0(\epsilon) \in \mathbb{N}$ such that if $N > N_0(\epsilon)$, the set Σ^N of possible N-sequences of the source alphabet decomposes into two sets Π and T satisfying

$$P(X_N \in \Pi) < \epsilon$$

$$2^{-N(H+\epsilon)} < P(X_N = \sigma) < 2^{-N(H-\epsilon)}$$

for any N-sequence $\sigma \in T$.

- Very technical proof due to (Billingsley, 1965) (omitted here).
- Proving that a source is ergodic is generally non-trivial and generally it is far easier to prove that a source is non-ergodic.

Agenda

Introduction

- Markov sources as the most realistic simple models for natural languages.
- A source $\mathcal{S}$ with alphabet $\Sigma = \{s_1, \ldots, s_m\}$ is said to be a *Markov source* if, letting $X = X_1, X_2, \ldots$ represent the source output, then for each $n \geq 1$ and any collection of source symbols $x_{n+1}, x_1, x_2, \ldots, x_n$, the following property holds

$$\begin{aligned} P(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) &= \\ P(X_{n+1} = x_{n+1} | X_n = x_n) &= p_{n+1,n} \end{aligned}$$

- This means that the probability distribution of X_{n+1}, given the past history, depends on the most recent input X_n only.
- We generalize to m-markov sources easily with

$$P(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) = P(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_{n-m+1} = x_{n-m+1})$$

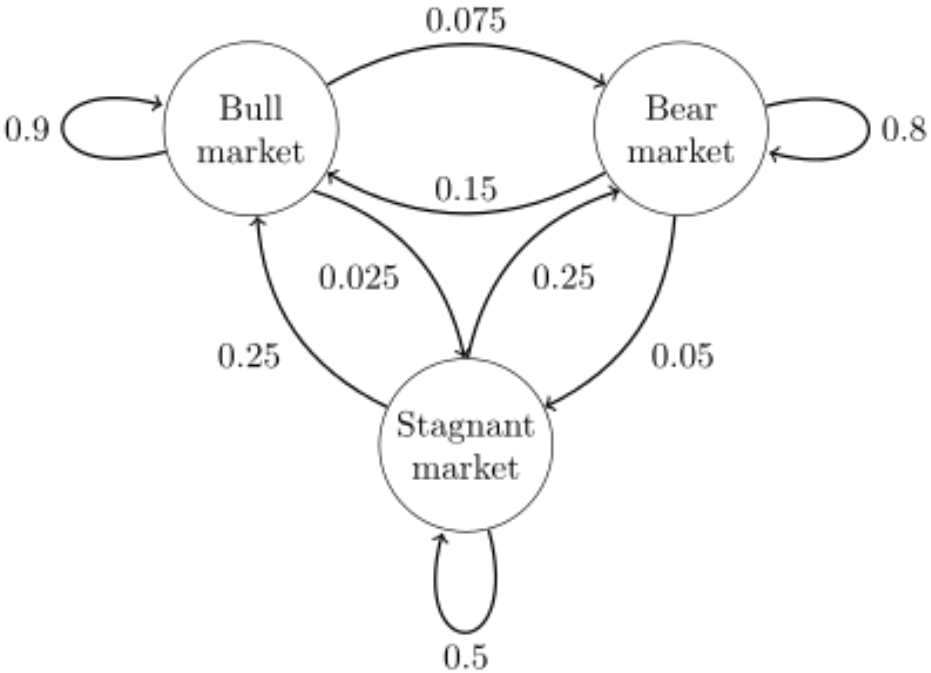
Introduction (2)

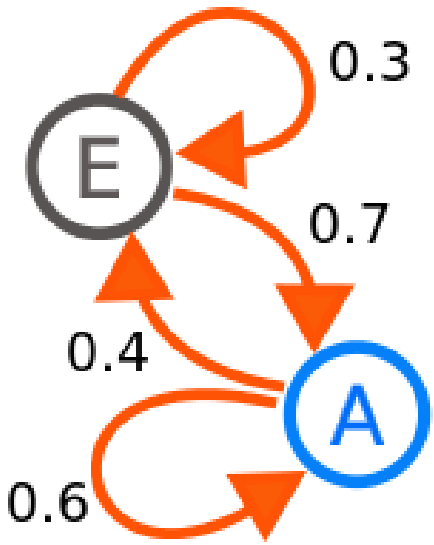
- To define things thoroughly we must have $p_{n+1,n} \geq 0$ and $\sum_j p_{ij} = 1 \quad (1 \leq i \leq m)$.
- The matrix P whose (i,j)-th entry is p_{ij} is the *transition matrix* of the source.
- To specify the source completely, we must also give the initial probabilities

$$\pi_i = P(X_1 = s_i) \quad (1 \leq i \leq m)$$

- In fact, in the context of stochastic processes, it is equivalent to *finite-state Markov chains*.

Examples of Markov Sources/Processes





- Give the transition matrix for these two markov sources.

Steady-state Distribution of $\mathcal{S}$

- Let us now work out the entropy of Markov sources. We first consider

$$H(X_{n+1}|X_n) = \sum_{j=1}^{m} H(X_{n+1}|X_n = s_j).P(X_n = s_j)$$

where

$$H(X_{n+1}|X_n = s_j) = -\sum_{k=1}^{m} p_{jk} \log_2(p_{jk}) = H_j$$

- Remind that the terms $P(X_n = s_j)$ depend on the initial distribution $\{\pi_i\}$.
- We define the *absolute probability* $a_j(n) = P(X_n = s_j)$. Using the Markov property and basic conditional probability, we have the following recurrence relation :

$$a_j(n+1) = \sum_i P(X_{n+1} = s_j|X_n = s_i).P(X_n = s_i) = \sum_i p_{ij} a_i(n),$$

$$a_j(1) = \pi_j$$

Steady-state Distribution of $\mathcal{S}$ (2)

- Whenever $\lim_{n\to\infty} a_j(n) = a_j$ exists for all j with $1 \leq j \leq m$, we say that $(a_1, \ldots, a_m)$ is the steady-state distribution of $\mathcal{S}$.
- If we write

$$H(X_{n+1}|X_n) = \sum_{j=1}^{m} H_j a_j(n)$$

and

$$\lim_{n\to\infty} H(X_{n+1}|X_n) = \sum_{j=1}^{m} a_j H_j \text{ exists}$$

- The right-hand side is precisely the entropy of the Markov source $\mathcal{S}$. We have proved the result :

Markov Source Entropy Theorem

If a Markov source $\mathcal{S}$ has steady-state distribution $(a_k : 1 \leq k \leq m)$, then its entropy H is given by

$$H = \sum_i a_i \sum_j p_{ij} \log_2(p_{ij}).$$

Theorem Application

- In order to use the previous Theorem, we need to be able to find a steady-state distribution of the source whenever it exists.
- Consider the recurrence relation on $a_j(n+1)$. Since the sum is finite for $n \to \infty$, the steady state must satisfy the set of homogenous equations

$$a_j = \sum_{i=1}^{m} p_{ij} a_i \quad (1 \leq j \leq m) \quad (1)$$

- Since a is a probability distribution, we must also have

$$\sum_{i=1}^{m} a_i = 1 \quad (2)$$

- For general Markov source, this system of equations may not have a unique solution.
- However if P is *irreducible*, those equations can be used to compute the entropy.

Markov Source with Irreducible Matrix

- The matrix P is said to be *irreducible* if for each i and j there exists n such that $(P^n)[i,j] > 0$
- This can interpreted, for the source $\mathcal{S}$ as, for any positive integers n, k

$$(P^n)[i,j] = P(X_{n+k} = s_j | X_k = s_i)$$

- In other words, P is irreducible if the source $\mathcal{S}$ is such that for each pair of symbols s_i and s_j there is a non-zero probability that at some time following the appearance of s_i, the source will emit s_j.
- It is the case for all natural languages modelled as a Markov source.

Irreducible Markov Sources (Grimmet & Stirzaker, 1982)

If a Markov source $\mathcal{S}$ is irreducible then equations (1) and (2) have a unique non-negative solution $v = (v_1, v_2, \ldots, v_m)$ which is called the *stationary distribution* of the source.

Markov Source with Irreducible Matrix (2)

- From Grimmet & Stirzaker's result we can prove the following theorem :

Theorem : Irreducible Markov Sources Entropy

An irreducible Markov source has an entropy H given by

$$H = \sum_{i=1}^{m} v_i H_i$$

where $v = (v_1, v_2, \ldots, v_m)$ is the stationary distribution and

$$H_i = \sum_j p_{ij} \log_2(p_{ij}).$$

- Exercice : prove the theorem.

Irreducible Ergodic Markov Source

Theorem : Irreducible Ergodic Markov Source

If $\mathcal{S}$ is an irreducible Markov source and if the unique stationary distribution is taken as the initial distribution, then $\mathcal{S}$ is an ergodic source which therefore has the AEP property.

- In practical situations, it is very unlikely that the unique stationary distribution will be exactly the initial distribution.
- However we can let run the source for a long time before starting the time clock. We this can expect the "initial distribution" to be close to the steady-state distribution (if it exists).
- Exercices.

Agenda

Extension of the Noiseless Coding Theorem

- Let us present a brief sketch of the way in which Shannon's two major theorems can be extended from the case of memoryless sources to that of any source which has the AEP and in particular to ergodic sources.
- We wish to encode the output of such a source $\mathcal{S}$ into an alphabet of size D.
- For a given $\epsilon > 0$, let N be large enough for the number T_N to satisfy $T_N \leq 2^{N(H+\epsilon)}$.
- Encode each of these by a distinct string of length r of symbols from the source alphabet of size D. Since there D^r such strings, this is achieved when $D^r > T_n$, that is when $r \log_2(D) \geq N(H + \epsilon)$.
- The other, atypical sequences from $\mathcal{S}$ are encoded by first prefixing a fixed string σ_0 of length r that was not used in the encoding of the typical sequences, and then encoding them by a string of lenth N.
- Exercice : give the average length l_N of this encoding (use δ as being the probability of an atypical sequence). Prove that it is a compact encoding.

Ergodic Sources Connected to a BSC

- We consider an ergodic source $\mathcal{S}$ of entropy H linked to a BSC of capacity C.
- Provided that $H < C$ we can find R such that $H < R < C$.
- By Shannon's noisy coding theorem, there exists a sequence of codes $(\mathcal{C}_n : 1 \leq N < \infty)$ such that $\mathcal{C}_n$ has 2^{Rn} codewords of length n and error probability that tends towards 0 as $n \to \infty$.
- We take N large enough so that T_N is $\approx 2^{NH}$. Then since $H < R$, we can encode them by the $\lfloor 2^{NR} \rfloor$ codewords of $\mathcal{C}_n$ and encode the remainder of the atypical N-strings arbitrarily with low probability error (refer to (Khinchin, 1957) for a detailed proof).

Agenda

Essential Bibliography

A few papers are available on the Moodle repository for this lecture.

- Billingsley, P. (1965). *Ergodic Theory and Information*, Wiley.
- Grimmet, G.R., Stirzaker, D. R. (1982). *Probability and Random Processes*, Oxford University Press.
- Khinchin, A. Ya. (1957). *Mathematical Foundations of Information Theory*, Dover.
- Shannon, C.E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656.
- Welsh, D. (1988). *Codes and Cryptography*, Oxford Science Publishing.