Information Theory - Noiseless Coding Theorem for Memoryless Sources

Eric Filiol

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

2013 - 2014

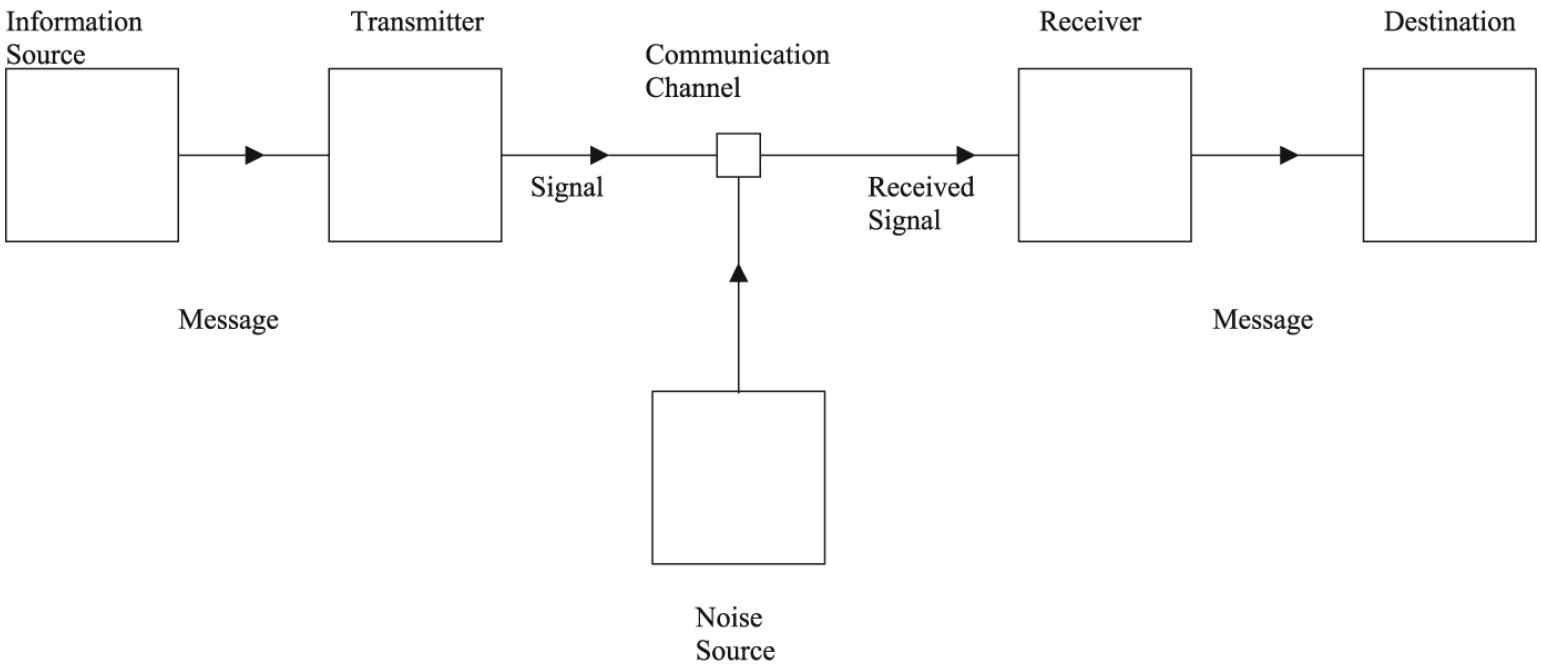
Introduction : What is an Information Source ?

- An *information source* $\mathcal{S}$ is a stream of symbols from some finite alphabet.
- There is usually some (more or less complicated) random mechanism based on the statistics of the situation being modelled.
- We are going to focus on a simple case of source : zero-memory or *memoryless* source :
  - If X_i denotes the i^{th} symbol produced by the source, then for each symbol a_j we set $P(X_i = a_j) = p_j$.
  - Probability p_j is independent from i (hence from all previous or future symbols emitted).
  - Then $X_1, X_2, \ldots$ is a sequence of identically distributed, independent random variables.
  - The entropy of a memoryless source is given by

$$H = -\sum_j p_j log_2(p_j)$$

over all $p_j \neq 0$.

Noiseless Coding Problem



Agenda

Agenda

Introduction

- Suppose we have a memoryless source $\mathcal{S}$ which emits symbols from a set $W = \{w_1, w_2, \ldots, w_m\}$ with probabilities $\{p_1, p_2, \ldots, p_m\}$ respectively.
- The quantities $w_i \in W$ are called the *source words*.
- Let Σ be an alphabet of D symbols, how can we encode the quantities $w_i \in W$, using symbols from Σ in the most economic way ?
- The cost aspect directly relates to critical issues like bandwidth or storage capacities, for exemples.
- The "optimality" of the encoding obviously relates to the **average** length of encoded source words.
- The Morse code exemple.

Source Word Encoding or Code

- An *encoding* or *code* is a map f from W into Σ^* (the collection of finite strings of symbols from Σ).
- A *message* is any finite string of source words $m = w_{i_1} \dots w_{i_k}$.
- The extension of f to W^* is defined by $f(m) = f(w_{i_1})f(w_{i_2}) \dots f(w_{i_k})$
- A code f is *uniquely decipherable* if any finite string from Σ^* is the image of at most one message.
- The string $f(w_i)$ is called the *codewords* and the integers $|f(w_i)|$ *word lengths* of f.
- The *average length* of the code f is $\langle f \rangle$ defined by

$$\langle f \rangle = \sum_{i=1}^{m} p_i |f(w_i)|$$

Source Word Encoding or Code (2)

- A code f is *instantaneous* of a *prefix code* if there do not exist distinct w_i and w_j such that $f(w_i)$ is a prefix of $f(w_j)$.
  - If $(x, y) \in \Sigma^* x \Sigma^*$ then is x is a prefix of y if there exists $z \in \Sigma^*$ such that $xz = y$.
- Instantaneous codes are clearly uniquely decipherable.
- Much stronger property : an instantaneous code can be decoded "on line" without looking into the future.
- Not every uniquely decipherable is instantaneous.
- Instead of using the map f, we will identify a code with the collection $\mathcal{C}$ of codewords.
- Examples.

Agenda

Kraft's Inequality

- The concept of uniquely decipherable code is much more difficult that that of instantaneous code.
- Is it possible to restrict our attention to instantaneous code in our search for uniquely decipherable codes having minimal average length ?
- Two fundamental inequalities are to be considered for this purpose.

Kraft's Inequality (1949)

If Σ is an alphabet of size D and W contains N words then a necessary and sufficient condition that there exists an instantaneous code $f : W \to \Sigma^*$ with word lengths $l_1, l_2, \ldots, l_N$ is that

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

McMillan's Inequality

McMillan's Inequality (1956)

If Σ is an alphabet of size D and W contains N words then a necessary and sufficient condition that there exists a uniquely decipherable code with codewords of length $l_1, l_2, \ldots, l_N$ is that $\sum_{i=1}^{N} D^{-l_i} \leq 1$ holds.

Combining Kraft's and McMillan's inequalities we have

Theorem

A uniquely decipherable code with prescribed words lengths exists if and only if an instantaneous code with the same word lengths exists.

We have proved that we can restrict our attention to instantaneous code in our search for uniquely decipherable.

Agenda

Introduction

Let us consider a memoryless source $\mathcal{S}$ which emits words $w_1, \ldots, w_m$ with probabilities $p_1, \ldots, p_m$ respectively.

- Given an alphabet Σ, the problem is to find a uniquely decipherable code whose average word length is as small as possible. Such code is called a *compact code*.
- Heuristic approach : the source $\mathcal{S}$ has entropy

$$H = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- The maximum entropy of an alphabet of D letters is $\log_2(D)$. Hence the number of symbols of the alphabet needed on the average to encode a word of the source should be about $\frac{H}{\log_2(D)}$.

Shannon's First Theorem (memoryless sources)

Shannon's First Theorem (memoryless sources)

If a memoryless source has entropy H, then any uniquely decipherable code for this source into an alphabet of D symbols must have length at least $\frac{H}{\log_2(D)}$. Moreover, there exists such a uniquely decipherable code $\mathcal{C}$ having average word length $l(\mathcal{C})$ less than equal to $1 + \frac{H}{\log_2(D)}$.

- This means that we can always find a uniquely decipherable code for which we have

$$\frac{H}{\log_2(D)} \leq l(\mathcal{C}) \leq 1 + \frac{H}{\log_2(D)}$$

- Statistical interpretation of entropy by means of source extension encoding.

Agenda

Introduction

- We have bounds on code average length $l(\mathcal{C})$ regarding a memoryless source $\mathcal{S}$

$$\frac{H(\mathcal{S})}{\log_2(D)} \leq l(\mathcal{C}) \leq 1 + \frac{H(\mathcal{S})}{\log_2(D)}$$

- The lower bound is satisfied whenever $p_i = (\frac{1}{D})^k$ for some integer k.
- From the Kraft-McMillan inequalities we have

Compact uniquely decipherable and instantaneous codes

If there exists a compact uniquely decipherable code of average length l, then there exists a compact instantaneous code of average length l.

- We may once again restrict attention to instantaneous codes.
- Simple techniques developped by (Huffman, 1952) in the case of binary alphabet.

Properties of Compact Codes

Lemma : compact code with two words

A compact code for a source with just two words w_1 and w_2 is

$$w_1 \to 0 \quad \text{and} \quad w_1 \to 1$$

Lemma : compact instantaneous codes

If $\mathcal{C}$ is instantaneous and compact and $p_i > p_j$, then $l(w_i) < l(w_j)$.

Lemma 2 : compact instantaneous codes

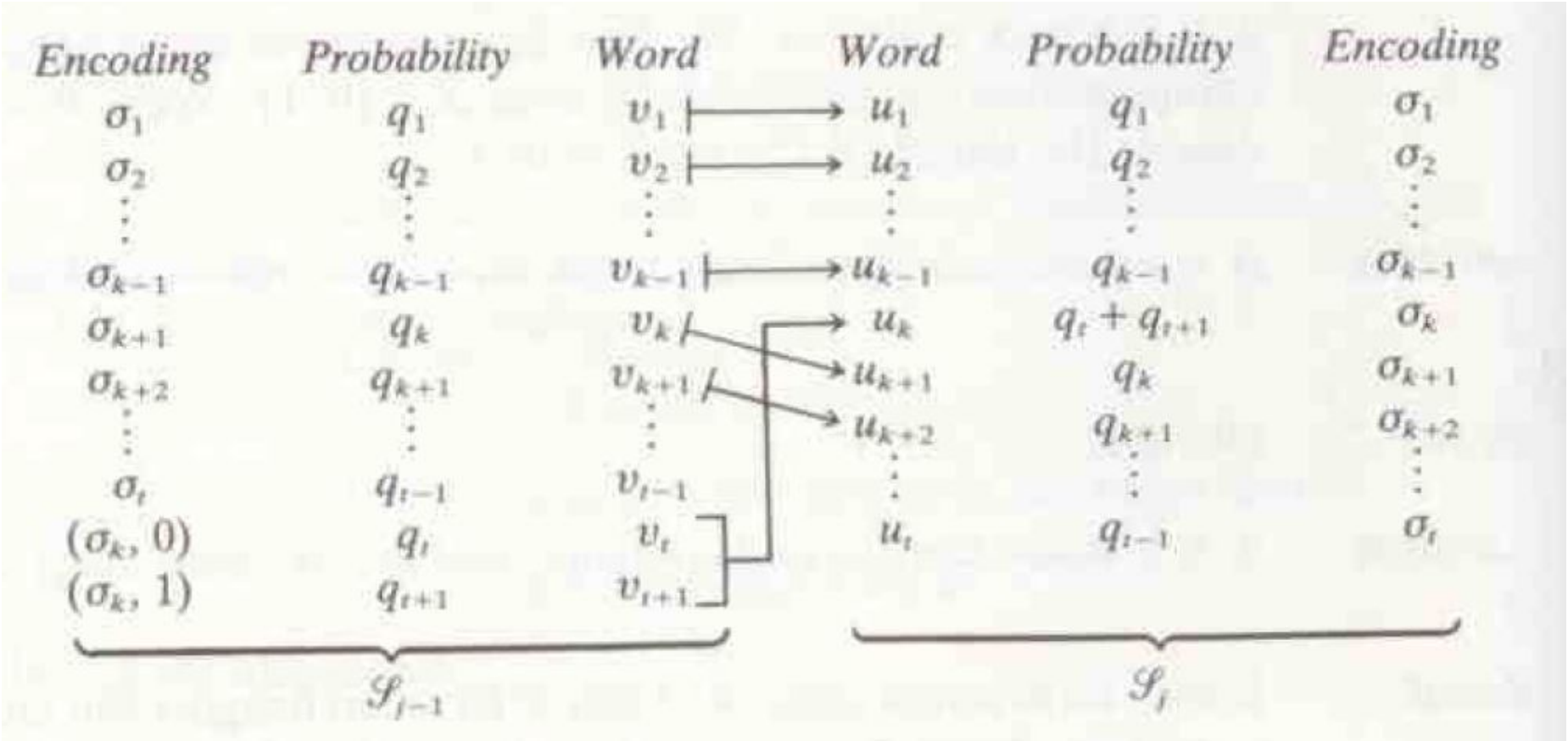
If $\mathcal{C}$ is instantaneous and compact then among the codeswords in $\mathcal{C}$ which have maximum length, there must be at least two agreeing in all but the last digit.

- First proof is obvious while second and third proofs are left as an exercise.

Huffman Algorithm

- We suppose that the source $\mathcal{S}$ has its collection of source words $\{w_1, w_2, \ldots, w_N\}$ ordered so that the probabilities p_i of emitting w_i satisfy
$$p_1 \geq p_2 \geq \ldots \geq p_N$$
- The Huffman procedure consists in building recursively a succession of sources $\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_{N-2}$ such that $\mathcal{S} = \mathcal{S}_0$ and $\mathcal{S}_k$ is obtained from $\mathcal{S}_{k-1}$ by identifying the two least probable symbols of $\mathcal{S}_{k-1}$ with a unique symbol σ in $\mathcal{S}_k$.
- The probability that σ is emitted from $\mathcal{S}$ is the sum of the probabilities of its two constituent symbols in $\mathcal{S}_{k-1}$.
- At each stage of reduction, we have a source with one fewer symbol until $N-2$ reductions we arrive at a source $\mathcal{S}_{N-2}$ with two symbols.

Huffman Coding Algorithm



Huffman Algorithm (2)

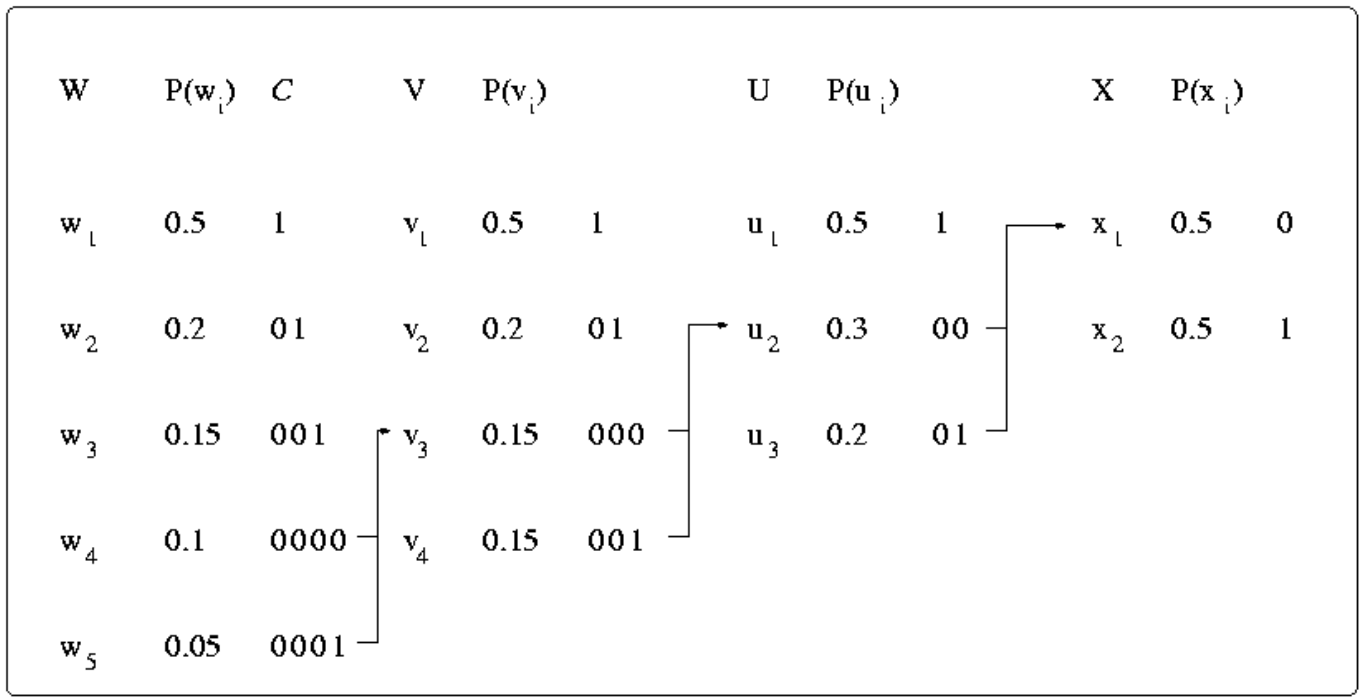
- The Huffman procedure to encode $\mathcal{S}_{j-1}$ is shown in the left-hand column of previous figure.
- Formally we have the following rule

$$v_i \mapsto \sigma_i \quad (1 \leq i \leq k-1) \qquad v_i \mapsto \sigma_{i+1} (k \leq i \leq t-1)$$

$$v_t \mapsto (\sigma_i, 0) \qquad v_{t+1} \mapsto (\sigma_i, 1)$$

- Thus we go backwards to build the codewords according to this procedure.
- Consider the next example : follow the procedure and compute $D, l(\mathcal{S}), H(\mathcal{S})$ and verify that $\mathcal{S}$ satisfies Shannon's First Theorem.

Huffman Coding Algorithm



Huffman Algorithm over Non-binary Alphabets

- We consider now an alphabet $\Sigma = \{0, 1, \ldots r-1\}$ of r symbols.
- The same algorithm basically applies.
- We just have to finish with a source $\mathcal{S}_t$ having r symbols.
- Two key points however
  - As we move from $\mathcal{S}_j$ to $\mathcal{S}_{j+1}$ collect not 2 but r least probable symbols of $\mathcal{S}_j$ into one symbol of $\mathcal{S}_{j+1}$. Thus $\mathcal{S}_{j+1}$ has $r-1$ fewer symbols.
  - Since the final source $\mathcal{S}_t$, we need to start off with a source $\mathcal{S}$ of $r + t(r-1)$ symbols. If not artificially augment $\mathcal{S}_t$ with $r + t(r-1) - |\mathcal{S}|$ dummy words having null probability.

Agenda

Conclusion

- It is possible to optimize the encoding of any source while being close to its amount of information.
- Huffman algorithm provides an optimal procedure for uniquely decipherable compact codes.
- Other algorithms known like (Shannon, Fano & Ellias coding).
- Shannon's first theorem addresses the compression of data as well.
  - The maximum value of the source entropy $\mathcal{S}$ arises when all source words $\{w_1, w_2, \ldots, w_m\}$ are equiprobable.
  - However, if they are not, the entropy $\mathcal{S}$ is strictly less than $\log_2(m)$. So the optimal code does compress the message.
- Go now to the computer room to practice with exercices.

Agenda

Essential Bibliography

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

- Huffman, D. A. (1952). A Method for the Construction of Minimum Redundancy Codes. *Proc. IRE*, 40 (10), pp. 1098–1101.
- Kraft, Leon G. (1949). *A device for Quantizing, Grouping, and Coding Amplitude Modulated Pulses*, Cambridge, MS Thesis, Electrical Engineering Department, Massachusetts Institute of Technology.
- McMillan, Brockway (1956). "Two Inequalities Implied by Unique Decipherability", *IEEE Trans. Information Theory* 2 (4) : 115–116, doi :10.1109/TIT.1956.1056818.
- Shannon, C.E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656.