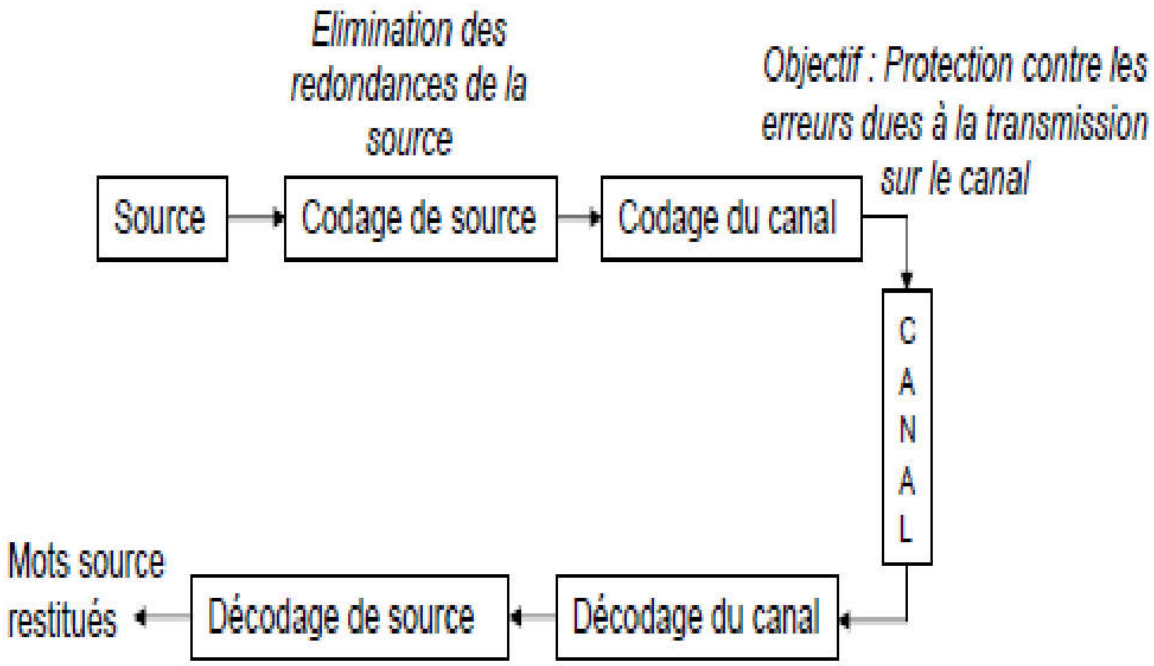
Chapitre3 Codage de source discrète sans mémoire

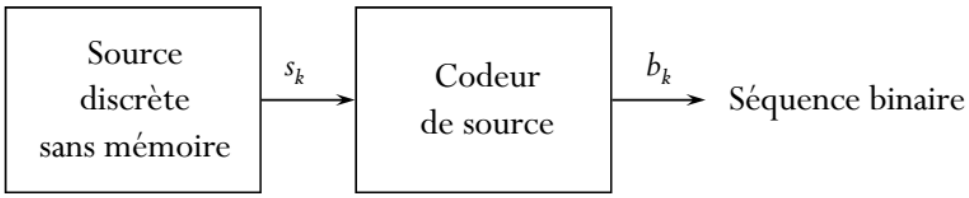
1 .Introduction :

- Le codage de source consiste à éliminer les redondances de la source afin d'en réduire le débit binaire Efficacité
- Le codage de canal a un rôle de protection contre les erreurs (dues à la transmission sur le canal) qui est assuré en ajoutant de la redondance (codes correcteurs d'erreurs). Fiabilité
- Les points de vue codage de source et codage de canal sont donc fondamentalement différents.



2. problématique

- Développer un codeur de source efficace satisfaisant aux contraintes suivantes :
  les mots-code sont sous la forme de données binaires ;
- les mots-code doivent être décodés de manière unique, i.e. la séquence émise par la source doit être parfaitement reconstruite à partir de la séquence des mots-code.



Exemple de MORSE

L'idée générale : coder par des mots de code courts les lettres les plus fréquentes. C'est le cas du codage de Morse En Morse, la lettre 'e' est représentée par le mot '·', alors que la lettre 'q', beaucoup moins fréquente, est codée par le mot '--.-'. De tels code sont dits de longueur variable

A	.-	N	-.	0	-----
B	-...	O	---	1	.----
C	-.-.	P	.--.	2	..---
D	-..	Q	--.-	3	...--
E	.	R	.-.	4	....-
F	..-.	S	...	5	.....
G	--.	T	-	6	-....
H	....	U	..-	7	--...
I	..	V	...-	8	---..
J	.---	W	.--	9	----.
K	-.-	X	-..-	.	.-.-.-
L	.-..	Y	-.--	,	--..--
M	--	Z	--..	?	..--..

3. Code et codage

Soit un alphabet (fini) X = (x1,…, xK) muni d'une loi de probabilité Px.

Définition 1 Un code de X est une application

φ: X➝ {0,1}*

(l'ensemble des mots binaires de longueur arbitraire).

Définition 2 Un mot de code est une séquence (non vide) de symboles pris dans l'alphabet de codage. Un mot de code est un élément de φ(X).

Définition 3 Un codage de X est une application

φ:X*➝ {0,1}*

qui a toute séquence finie de lettres de X associe une séquence binaire. Le codage fait correspondre les symboles-sources et les mots de code.

A tout code φ de X on peut associer le codage (x1, x2,…,xL)⟶ (φ(x1) ǁ φ (x2) ǁ …. ǁ φ (xL))

remarque : La réciproque n'est pas vraie.

Définition 4 Un code (resp. codage) est dit régulier si deux lettres (resp. séquences de lettres) distinctes sont codées par des mots distincts.

Un code non régulier implique une perte d'information. Un code régulier est un code non ambigu.

Définition 5 on dit qu'un code est instantané si et seulement si chaque mot de code dans toute chaine de mots de code peut être décodé des que l'on a atteint sa fin.

Cette définition garantit qu'il n'est ni nécessaire de mémoriser les mots de code reçus ni d'attendre les suivants pour effectuer le décodage. Un tel code permet d'économiser du temps et de l'espace dans le processus de décodage d'un message codé.

Exemple : Considérons la source composée des trois symboles a, b et c. Un message est une séquence de ces symboles. Soit les deux codes suivants :

Code 1 : a→ 0 b → 1 c → 01

Code 2 : a→ 0 b → 11 c → 01.

Le code 1 est ambigu, la séquence de code 0101 peut correspondre à plusieurs messages à savoir abab, abc, cab ou cc. Par contre, le code 2 est régulier.

4. Arbre de codage

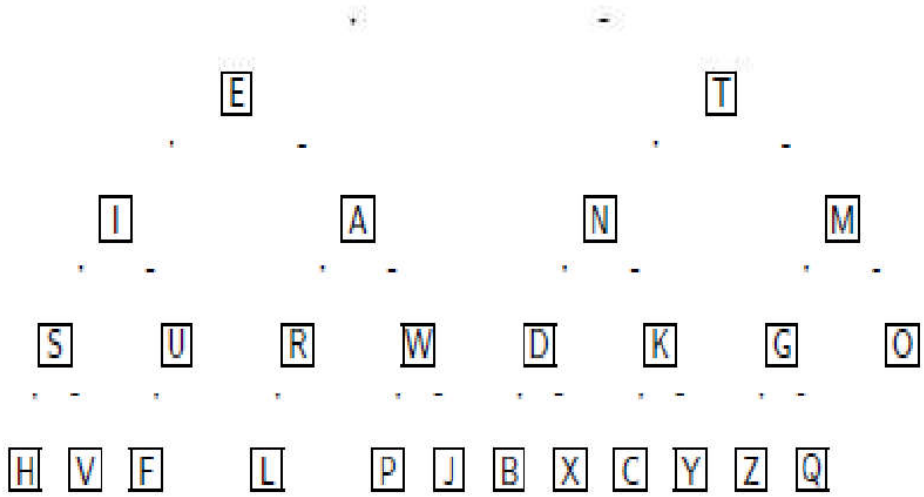
Un arbre est un graphe de nœuds relies par des arcs ; le nœud de départ est dit racine et les nœuds terminaux sont dits feuilles. Un arbre est dit n-aire (n ≥ 1) si chaque nœud intérieur a n fils et il est dit n-aire complet si toutes les branches ont la même profondeur. Un arbre de codage est un arbre n-aire, dont les arcs sont étiquetés par des lettres d'un alphabet donné de taille n, de façon à ce que chaque lettre apparaisse tout au plus une fois à partir d'un nœud donné. Les mots de code définis par un tel arbre correspondent à des séquences d'étiquettes le long des chemins menant de la racine à une feuille. Pour tout code instantané il existe au moins un arbre tel que chaque mot de code corresponde à un unique chemin de la racine à une feuille.

Remarque Il peut arriver que quelques *feuilles de l'arbre soient non utilisées* pour quelque code instantané.

Définition 1 : On dit que le code est un code complet lorsqu'il n'y a pas de feuille vide dans l'arbre de codage n-aire correspondant.

Exemple de MORSE

On peut représenter le codage de Morse à l'aide d'un arbre binaire. Chaque nœud, à l'exception de la racine, est le codage d'une lettre.



5. Code préfixé :

Supposons que nous cherchons à transmettre le message BOF en utilisant le trois codes figurant dans le tableau suivant

Symbole	Probabilité	Code I	Code II	Code III
I	½	1	0	0
B	¼	00	10	01
F	1/8	01	110	011
O	1/8	10	111	111

- Avec le code I, le message envoyé est : 001001
- Avec le code II le message envoyé est: 10111110
- Avec le code III, le message envoyé est 01111011

La question posée est : « le message décodée à la réception est-il identique au message envoyée ?» Pour répondre à cette question il faut étudier comment chaque code décode le message reçue. Avec le code I, le message reçu est : 001001. Un décodage de cette séquence peut être interpréter par BIBI (00 1 00 1) ou par BOF (00 10 01). Un problème est évident : le message n'est pas décodable de manière unique. Ce problème est dû au fait que le code de « I » qui est « 1 » est un aussi la première valeur de la séquence du code de « O » qui est équivalente à « 10 ». Avec le code III, le message reçu est 01111011. Au décodage nous pouvons voir 0 111….c'est à dire IO… ; Mais ce qui suit peut être soit 1, soit 10, soit 101 qui ne sont pas des codes ce qui nous oblige à revenir en arrière pour modifier l'interprétation soit 01 111 011 et retrouver le bon message. On dit alors que le code n'est pas décodable de manière instantanée.

Avec le code II le message reçu est 10111110. Au décodage on récupère le bon message BOF. Le code II possède les deux propriétés souhaitées:
✓ décodable de manière unique et
✓ Décodable de manière instantanée. Le code II est dit code préfixe ; aucun de ses mots de code n'est un préfix d'un autre. On dit qu'une séquence x est un préfixe d'une autre séquence x' si et seulement si les n premiers symboles de x' forment exactement la séquence x.

Définition 1 Code préfixe On dit que le code d'une source discrète est préfixe si et seulement si aucun mot de code n'est le préfixe d'un autre mot de code.

Remarque 1 Un code préfix est décodable de manière unique et instantanée.

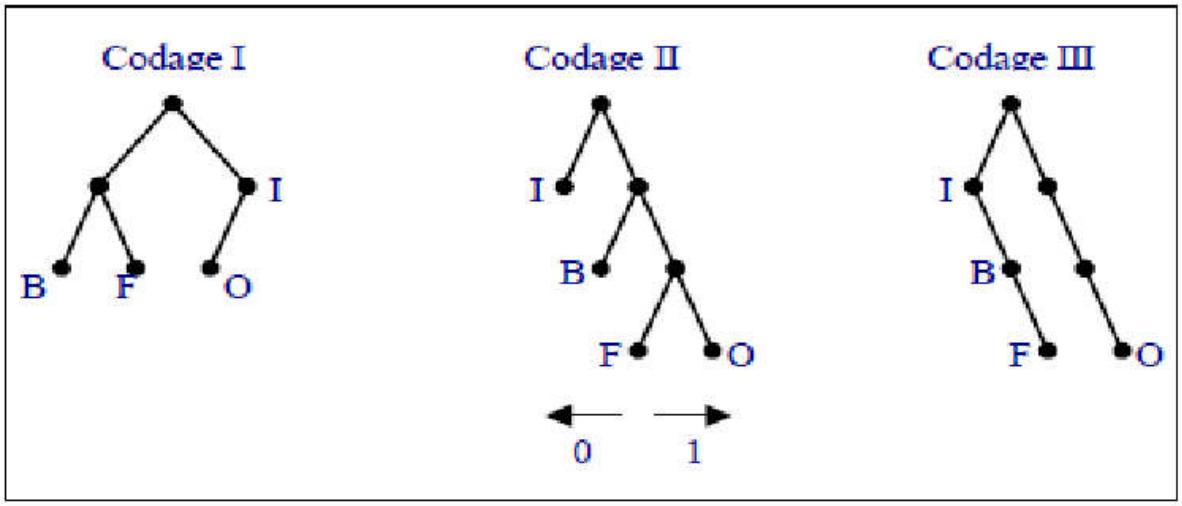


Figure Arbre de chaque code

Remarque 2 Un code préfix sur l'alphabet binaire {0,1} est représenté par un arbre binaire où tous les symboles sont des feuilles.

6. Définition

- Un code source C est un mapping (conversion) entre l'alphabet *X* et l'alphabet *D*. On note C(x) le mot-code associé au symbole x, et *l(x)* la longueur de ce mot-code. Un code source doit posséder des propriétés telles que l'information de base puisse être reconstruite, et doit utiliser au mieux la capacité d'un canal (exemple code morse).
- La longueur moyenne d'un code source C pour une source X est donnée par :

$$L(C) = \sum_{x \in \mathcal{X}} p(x)\ell(x)$$

Prenons l'exemple de codage suivant pour l'alphabet *X = {a, b, c, d}* : Les probabilités d'avoir un certain symbole en provenance de la source sont données par

$$\begin{cases} P(X=a) = \frac{1}{2} \\ P(X=b) = \frac{1}{4} \\ P(X=c) = \frac{1}{8} \\ P(X=d) = \frac{1}{8} \end{cases}$$

On décide de coder chaque symbole en bits. Par exemple, *C*(*a*) est le codage de a. On note donc

$$\begin{cases} C(a) = 0 \\ C(b) = 10 \\ C(c) = 110 \\ C(d) = 111 \end{cases}$$

Ce qui donne les longueurs suivantes :

$$\begin{cases} \ell(a) = 1 \\ \ell(b) = 2 \\ \ell(c) = 3 \\ \ell(d) = 3 \end{cases}$$

A partir de ces informations, on peut donc calculer l'entropie de la source et la longueur moyenne
du codage. L'entropie vaut :

$$H(X) = \frac{1}{2}\underbrace{\log_2 2}_{1} + \frac{1}{4}\underbrace{\log_2 4}_{2} + \frac{2}{8}\underbrace{\log_2 8}_{3} = 1,75\ bits/symbole$$

Et la longueur moyenne vaut :

$$L(C) = \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{3}{8} = \frac{7}{4} = 1,75\ bits/symbole$$

On a dans ce cas-ci un codage optimal, et en plus la longueur moyenne est égale à l'entropie, ce qui sont deux critères qui ne sont pas forcément toujours compatibles. C'est le cas du codage suivant.
On a les probabilités suivantes :

$$\begin{cases} P(X=a) = \frac{1}{3} \\ P(X=b) = \frac{1}{3} \\ P(X=c) = \frac{1}{3} \end{cases}$$

Avec le codage et les longueurs suivantes

$$\begin{cases} C(a) = 0 \\ C(b) = 10 \\ C(c) = 11 \end{cases} \quad \begin{cases} \ell(a) = 1 \\ \ell(b) = 2 \\ \ell(c) = 2 \end{cases}$$

On obtient alors la longueur moyenne

$$L(C) = \frac{1}{3}1 + \frac{2}{3}2 = \frac{5}{3} = 1,667\ bits/symbole$$

Et l'entropie :

$$H(X) = \log_2 3 = 1,58\ bits/symbole$$

On voit ici que L est différent de H. Pourtant, et le codage est optimal.

7. Classes de codes

• **Codes non singuliers**

C est un code non singulier si :

$$\forall x_i, x_j \in X : \{x_i \neq x_j \Rightarrow C(x_i) \neq C(x_j)\}$$

Contrairement aux codes singuliers, ils ne sont pas ambigus.

• **Codes décodables de façon unique (DFU)**

C est un code DFU si son extension, c'est-à-dire la concaténation des mots-code individuels

$C*(x1, ..., xn) = C(x1)C(x2)...C(xn)$, est non singulière.

• **Codes instantanés** : C est un code instantané (ou prex code, ou self-ponctuating code) s'il vérifie la condition de préfixe : "aucun mot-code ne peut être le préfixe d'un autre mot-code".
Exemple
Prenons le codage suivant :

$$\begin{cases} C(a) = 0 \\ C(b) = 10 \\ C(c) = 111 \\ C(d) = 110 \end{cases}$$

Introduire 1110 serait rendre le code non instantané puisque 111 serait son préfixe. Si on reçoit la séquence 01011111010, on peut directement, bit après bit, décoder l'arrivée de nouveaux symboles, et traduire directement en abcdb sans qu'il ait fallu envoyer de bits pour indiquer la séparation entre les symboles, ou sans qu'il ait fallu attendre plusieurs bit excédentaires pour connaître l'identité d'un symbole.

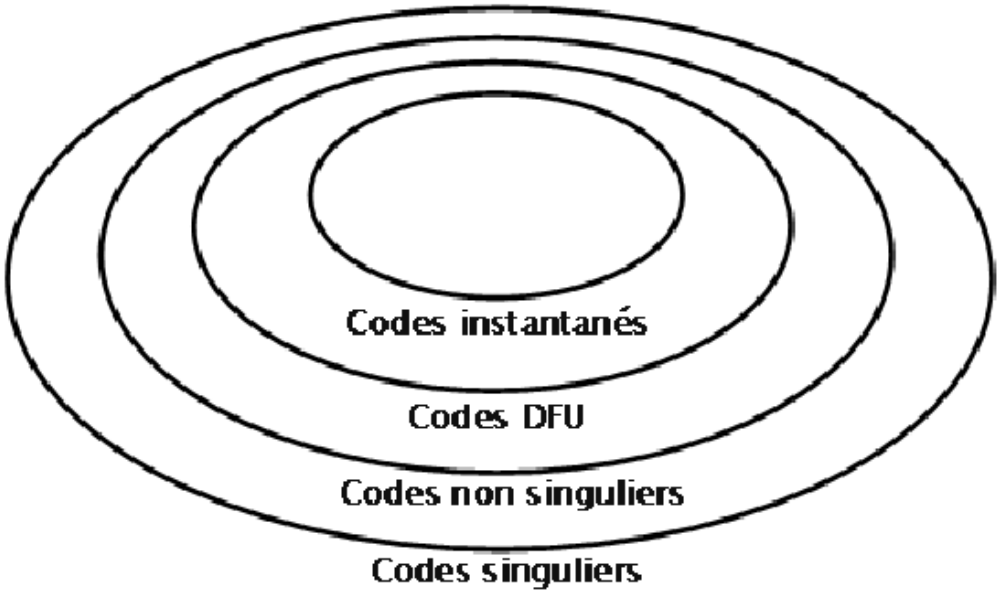


Schéma reprenant les différentes classes de codes

Ci-dessus est représenté un schéma comprenant les différentes classes de codes, groupées en ensembles. D'autres exemples sont donnés dans le tableau ci-dessous

	Non singulier	DFU	instantané
X=a	0	10	0
X=b	010	00	10
X=c	01	11	110
X=d	10	110	111

exemple de code pour chaque classe.

8. Inégalité de Kraft-MacMillan

Il existe un code instantané de n mots de code et dont les longueurs des mots de code sont les entiers positifs L1, L2, . . . , Ln si et seulement si :

$$\sum_{i=1}^{n} 2^{-L_i} \leq 1.$$

Exemple 1: (0, 01, 011)

$$2^{-1} + 2^{-2} + 2^{-3} = 7/8<1$$

Uniquement décodable (satisfait l'égalité de Kraft–McMillan) Mais pas instantané (non prefix)

Exemple 2: (0, 01, 001)

$$2^{-1} + 2^{-2} + 2^{-3} = 7/8<1.$$

Uniquement décodable (satisfait l'égalité de Kraft–McMillan) Instantané (prefix)

9. Efficacité d'un code

Définition 1 :

Une source X est dite sans mémoire si sa loi de probabilité Px ne varie pas au cours du temps. Son entropie est égale :

$$H(X) = \sum_{x \in X} - P_X(x) \log_2 P_X(x)$$

Définition 2 :

La longueur moyenne d'un code d'une source discrète sans mémoire est défini par:

$$\bar{n}(\varphi) = \sum_{x \in X} P_X(x) |\varphi(x)$$

(|φ(x) est la longueur de φ(x))

Définition 3 :

L'efficacité d'un code φ d'une source discrète sans mémoire X est définie par

$$E(\varphi) = \frac{H(X)}{\bar{n}(\varphi)}$$