Nicolas Sendrier

Majeure d'informatique

Introduction à la théorie de l'information

Cours n°2

Codage des sources discrètes sans mémoire

Codage de source

L'idée générale : coder par des mots de code courts les lettres les plus fréquentes. C'est le cas du code Morse

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	--..	?	..--..

Il s'agit en fait d'un code ternaire, puisqu'il faut un symbole supplémentaire pour séparer les lettres.

Impossible sinon distinguer, par exemple,

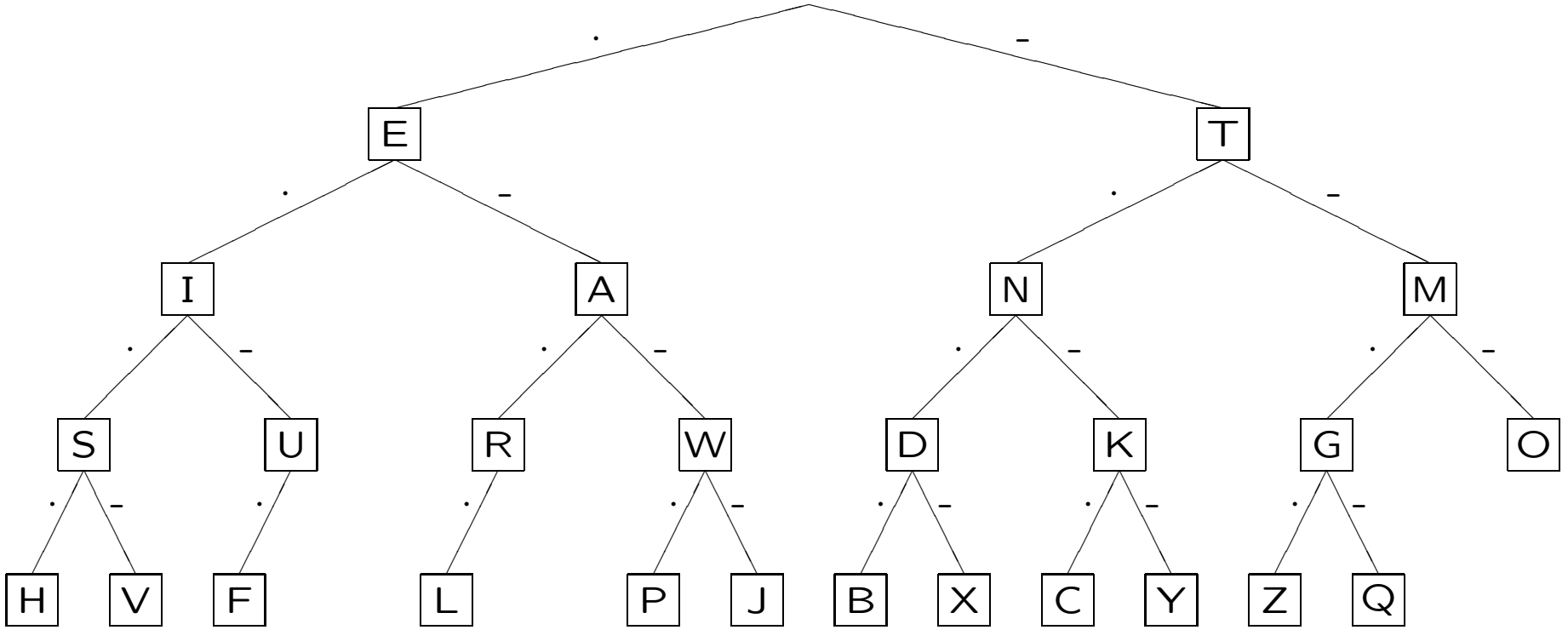
– "BAM" $\rightarrow$ "-....---"

– "NIJ" $\rightarrow$ "-....---"

Adapté à un opérateur humain, mais pas aux moyens de communication modernes (synchrones).

Code Morse

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



Code et codage

Soit un alphabet (fini) $\mathcal{X}$.

Définition Un *code* de $\mathcal{X}$ est une application $\varphi : \mathcal{X} \to \{0,1\}^*$ (l'ensemble des mots binaires de longueur arbitraire).

Définition Un *mot de code* est un élément de $\varphi(\mathcal{X})$.

Définition Un *codage* de $\mathcal{X}$ est une application $\psi : \mathcal{X}^* \to \{0,1\}^*$, qui à toute séquence finie de lettres de $\mathcal{X}$ associe une séquence binaire.

À tout code φ de X on peut associer le codage

$$(x_1, x_2, \ldots, x_L) \to (\varphi(x_1) \parallel \varphi(x_2) \parallel \ldots \parallel \varphi(x_L))$$

(la réciproque n'est pas vraie)

Définition 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.

Source sans mémoire – Efficacité

Une source discrète $X = (\mathcal{X}, p)$ est un alphabet $\mathcal{X}$ muni d'une loi de probabilité p.

Définition Une source $X = (\mathcal{X}, p)$ est dite *sans mémoire* si sa loi de probabilité ne varie pas au cours du temps. Son *entropie* est égale à

$$H(X) = \sum_{x \in \mathcal{X}} -p(x) \log_2 p(x).$$

Définition La *longueur moyenne* d'un code φ d'une source discrète sans mémoire $X = (\mathcal{X}, p)$ est définie par

$$|\varphi| = \sum_{x \in \mathcal{X}} p(x)|\varphi(x)|$$

($|\varphi(x)|$ est la longueur de $\varphi(x)$)

Définition L'*efficacité* d'un code φ d'une source discrète sans mémoire $X = (\mathcal{X}, p)$ est définie par

$$E(\varphi) = \frac{H(X)}{|\varphi|}.$$

Efficacité d'un codage

Soit $(x_1, \ldots, x_n)$ une séquence finie de lettres de $\mathcal{X}$, nous noterons

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$$

sa probabilité. La *longueur moyenne par lettre* du codage ψ sera définie par la limite suivante (si elle existe)

$$\mathcal{L}(\psi) = \lim_{n \to \infty} \frac{1}{n} \sum_{x_1, \ldots, x_n} p(x_1, \ldots, x_n) |\psi(x_1, \ldots, x_n)|$$

Définition Soit X une source discrète sans mémoire et ψ un codage de X dont la longueur moyenne par lettre est définie. L'*efficacité* de ψ est égale à

$$E(\psi) = \frac{H(X)}{\mathcal{L}(\psi)}.$$

Codes de longeur fixe

Proposition Pour tout code régulier de longueur n d'une source X de cardinal K, nous avons

$$\log_2 K \leq n$$

L'efficacité d'un tel code est donc limitée par $H(X)/\log_2 K$ (qui vaut 1 si la loi de X est uniforme).

Proposition Pour toute source X de cardinal K, il existe un code régulier de longueur n telle que

$$\log_2 K \leq n < 1 + \log_2 K$$

Corollaire Il existe un codage régulier de X dont l'efficacité est arbitrairement proche de $H(X)/\log_2 K$.

Codes de longeur variable

Définition Un code est dit *à décodage unique* si son codage associé est injectif.

Autrement dit, une séquence binaire finie donnée correspond au plus à un séquence de lettres de la source.

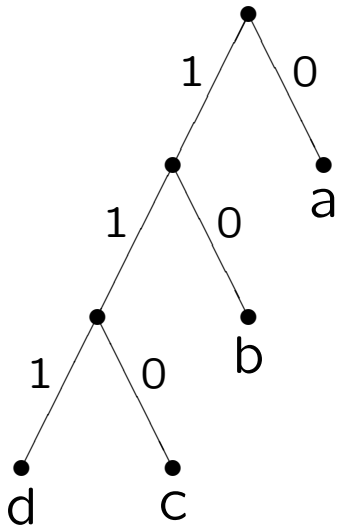
Condition du préfixe
Aucun mot de code n'est le début d'un autre

Définition Un code est dit *préfixe* s'il vérifie la condition du préfixe. Nous parlerons aussi de code *instantané*.

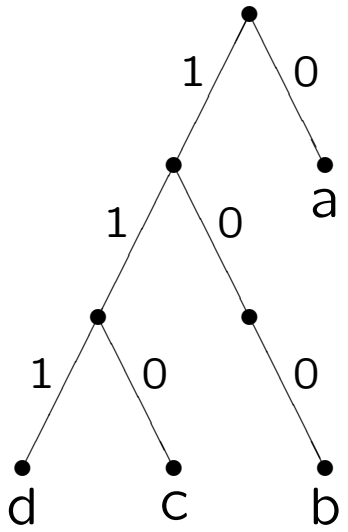
Proposition Tout code préfixe est à décodage unique.

Arbre associé à un code préfixe

Pour tout code préfixe, il existe un arbre dont les mots de codes sont les feuilles (condition nécessaire et suffisante).



$a \rightarrow 0$
$b \rightarrow 10$
$c \rightarrow 110$
$d \rightarrow 111$



$a \rightarrow 00$
$b \rightarrow 100$
$c \rightarrow 110$
$d \rightarrow 111$

Inegalité de Kraft – Théorème de Mac Millan

Théorème (Kraft) Il existe un code préfixe dont les K mots ont pour longueur $n_1, n_2, \ldots, n_K$ si et seulement si

$$\sum_{k=1}^{K} \frac{1}{2^{n_k}} \leq 1.$$

Théorème (Mac Millan) Il existe un code à décodage unique dont les K mots ont pour longueur $n_1, n_2, \ldots, n_K$ si et seulement si

$$\sum_{k=1}^{K} \frac{1}{2^{n_k}} \leq 1.$$

Premier théorème de Shannon

Proposition

1. Pour toute source d'entropie H codée au moyen d'un code à décodage unique de longueur moyenne $\bar{n}$, on a $\bar{n} \geq H$.
2. Pour toute source d'entropie H, il existe un code préfixe de longueur moyenne $\bar{n}$ telle que $H \leq \bar{n} < H + 1$.

Théorème (Shannon) Pour toute source discrète sans mémoire, il existe un codage régulier dont l'efficacité est arbitrairement proche de 1.