Nicolas Sendrier

Majeure d'informatique

Introduction à la théorie de l'information

Cours n°2

Codage des sources discrètes

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.

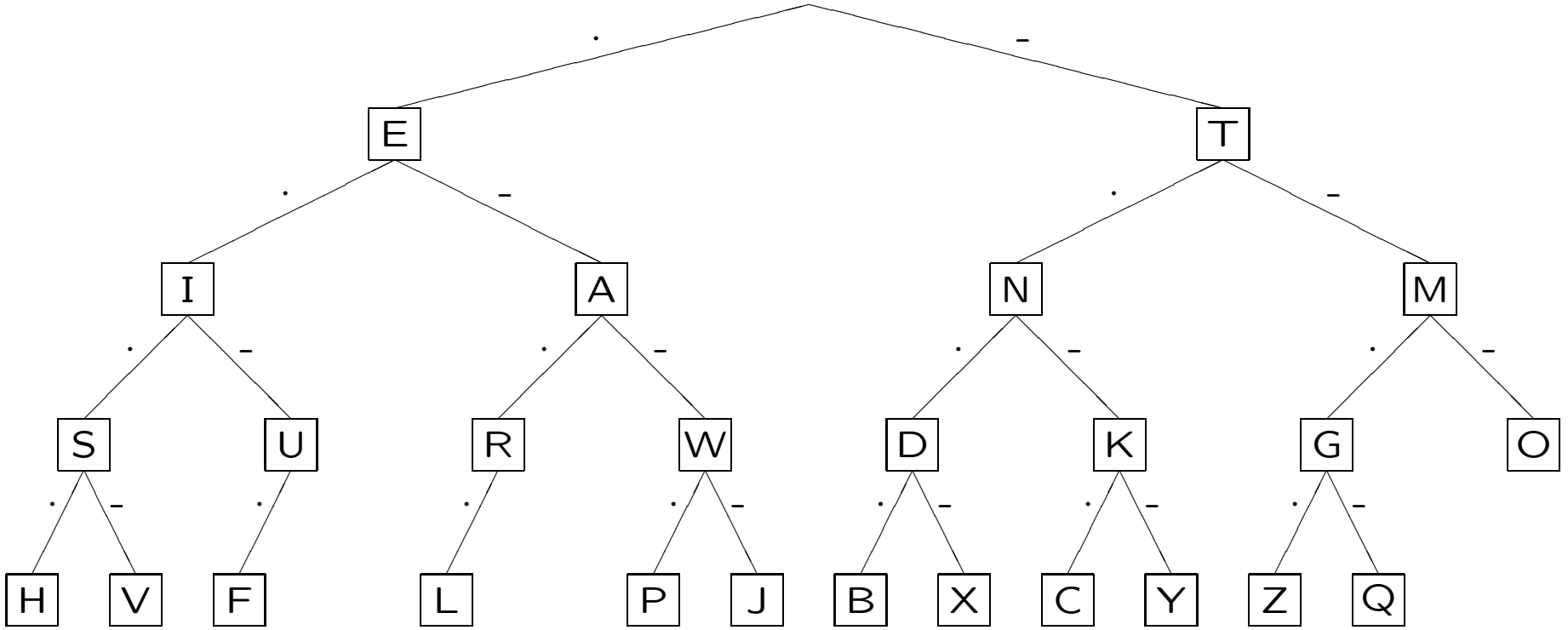
Impossible sinon distinguer, par exemple,

- "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

Une source discrète X est un alphabet (fini) $\mathcal{X} = \{a_1, \ldots, a_K\}$ muni d'une loi de probabilité P_X.

Définition Un *code* de 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 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

Définition Une source X est dite *sans mémoire* si sa loi de probabilité P_X ne varie pas au cours du temps. Son *entropie* est égale à

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

Définition La *longueur moyenne* d'un code φ d'une source discrète sans mémoire est défini par

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

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

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

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

Efficacité d'un codage

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

$$P_{X^L}(x_1, \ldots, x_L) = \prod_{i=1}^{L} P_X(x_i)$$

sa probabilité. La *longueur moyenne par lettre* des séquences de longueur L est définie par

$$\bar{n}_L(\psi) = \frac{1}{L} \sum_{(x_1,\ldots,x_L)\in\mathcal{X}^L} P_{X^L}(x_1, \ldots, x_L)|\psi(x_1, \ldots, x_L)|$$

Définition L'*efficacité* d'un codage ψ d'une source discrète sans mémoire X est définie, lorsque la limite ci-dessous existe, par

$$E(\psi) = \lim_{L\to\infty} \frac{H(X)}{\bar{n}_L(\psi)}.$$

Codes de longueur 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

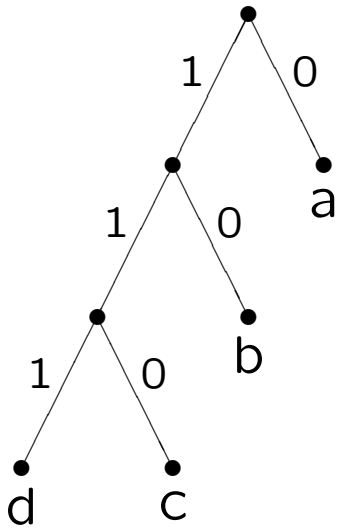
Aucun mot de code n'est le début d'un autre

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

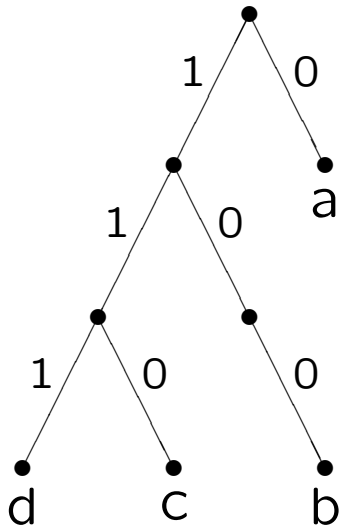
Proposition Tout code préfixe est à décodage unique.

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.

Théorème de Shannon pour une source stationnaire

Une source discrète stationnaire est un processus stochastique, soient $X_1, X_2, \ldots, X_L, \ldots$ les variables aléatoires à valeur dans l'alphabet $\mathcal{X}$ associées à ce processus (X_i est la lettre émise par la source à la i-ème unité de temps).

L'entropie par lettre est définie par

$$H(\mathcal{X}) = \lim_{L\to\infty} \frac{1}{L} H(X_1, \ldots, X_L) = \lim_{L\to\infty} H(X_L \mid X_{L-1}, \ldots, X_1)$$

et la longueur moyenne et l'efficacité d'un codage ψ par

$$\bar{n}_L(\psi) = \frac{1}{L} \sum_{x_1,\ldots,x_L} P(x_1, \ldots, x_L)|\psi(x_1, \ldots, x_L)| \text{ et } E(\psi) = \frac{H(\mathcal{X})}{\bar{n}_L(\psi)}$$

Théorème (Shannon) Pour toute source stationnaire, il existe un codage régulier dont l'efficacité est arbitrairement proche de 1 (sans pouvoir dépasser ce nombre).