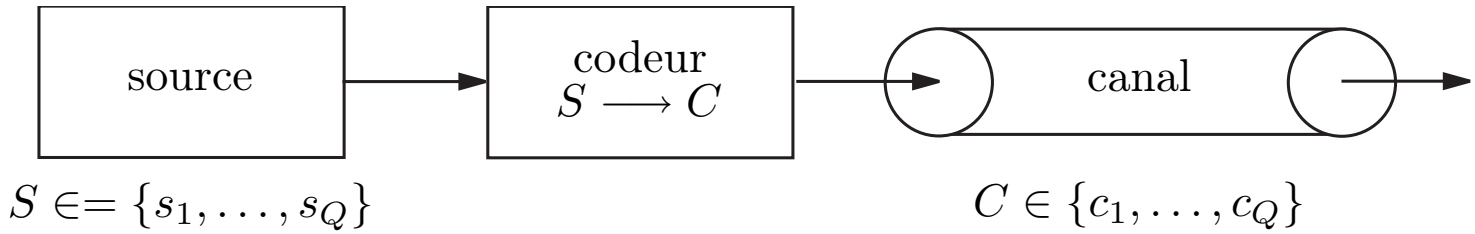
Théorie et codage de l'information

Codage de source discrète

- Chapitre 3 -

Codage de source discrète

On associe à chacun des Q états s_i de la source un mot approprié, c'est-à-dire une suite de n_i symboles d'un alphabet q-aire. Ceux-ci constituent un code source que l'on note $\mathcal{C} = \{c_1, \ldots, c_q\}$.



Exemple. Le code Morse

- ▷ code quaternaire (point, trait, espace long, espace court)
- ▷ code de longueur variable
- ▷ la séquence la plus courte associée à "E"

Problématique

Adaptation d'une source à un canal non-bruité

Soit S une source caractérisée par un débit D_s (symbole Q-aire/seconde). Soit un canal non-bruité de débit maximal D_c (symbole q-aire/seconde). On définit

- taux d'émission de la source : $T \triangleq D_s\, H(S)$
- capacité du canal : $C \triangleq D_c \log q$

Si T > C : le canal ne peut écouler l'information

Si T $\leq$ C : le canal peut en théorie écouler l'information

Si on dispose d'un code q-aire dont la longueur moyenne $\overline{n}$ des mots est telle $\overline{n}\, D_s \leq D_c$, alors celui-ci peut être utilisé pour la transmission.

Dans le cas contraire, comment coder les états de la source pour rendre leur transmission possible puisque rien ne s'y oppose en théorie ?

Le codage de source vise à éliminer la redondance d'information SANS PERTE ! ! !

Source discrète en temps discret

Modèle général

Une source discrète est définie par un alphabet $\mathcal{A} = \{s_1, \ldots, s_Q\}$ et un mécanisme d'émission. Il s'agit d'un processus aléatoire en temps discret

$$S_1, \ldots, S_{i-1}, S_i, S_{i+1}, \ldots$$

caractérisé par les lois conjointes :

$$P(S_1, \ldots, S_n), \forall n \in \mathbb{N}^*$$

▷ **modèle trop général pour donner lieu à des développements simples**

Source discrète en temps discret

Hypothèses complémentaires

Par simplification, on fait des hypothèses sur le modèle de source.

Propriété 1 (Processus stationnaire)**.** *Un processus aléatoire S_i est dit stationnaire si les lois de probabilité qui le régissent sont indépendantes de l'origine des temps, c'est-à-dire*

$$P(S_1 = s_{i_1}, \ldots, S_n = s_{i_n}) = P(S_{n_0+1} = s_{i_1}, \ldots, S_{n_0+n} = s_{i_n}),$$

pour tous n_0 et n positifs.

Exemple. Une *source sans mémoire* est caractérisée par des S_i indépendants et identiquement distribués. Il s'agit d'un processus stationnaire.

$$P(S_1 = s_{i_1}, \ldots, S_n = s_{i_n}) = P(S = s_{i_1}) \ldots P(S = s_{i_n}).$$

Source discrète en temps discret

Hypothèses complémentaires

Toujours par souci de simplification, on fait l'hypothèse d'ergodicité.

Propriété 2 (Processus ergodique)**.** *On dit qu'un processus aléatoire stationnaire S_i est ergodique si, pour tout $k = 1, 2, \ldots$, pour toute suite d'indices $i_1, \ldots, i_k$ et pour toute fonction bornée $f(\cdot)$ de $\mathcal{A}^k$ dans $\mathbb{R}$, on a*

$$\frac{1}{n}\sum_{k=1}^{n} f(S_{i_1}, \ldots, S_{i_k}) \xrightarrow{p.s.} E\{f(S_{i_1}, \ldots, S_{i_k})\}.$$

Intérêt. Le processus considéré peut être étudié en observant une trajectoire quelconque mais suffisamment longue de celui-ci.

Source discrète en temps discret

Sources de Markov

Une source quelconque émet un symbole selon une loi qui peut dépendre des symboles qui l'ont précédé.

Définition 1 (Source markovienne)**.** *Une source S est dite markovienne si elle décrit une chaîne de Markov, soit*

$$P(S_{n+1} = s_{i_{n+1}} | S_n = s_{i_n}, \ldots, S_1 = s_{i_1}) = P(S_{n+1} = s_{i_{n+1}} | S_n = s_{i_n})$$

pour tous symboles $s_{i_1}, \ldots, s_{i_{n+1}}$ issus de $\mathcal{A}$.

Il en résulte directement que

$$P(S_1, \ldots, S_n) = P(S_1)\, P(S_2|S_1) \ \ldots \ P(S_n|S_{n-1})$$

Source discrète en temps discret

Sources de Markov

Définition 2 (Invariance dans le temps)**.** *Une source markovienne S est dite invariante dans le temps si, pour tout $n \in \{1, 2, \ldots\}$, on a*

$$P(S_{n+1}|S_n) = P(S_2|S_1)$$

Une telle source est entièrement définie par un vecteur $p|_{t=0}$ de probabilités initiales et la matrice de transition Π dont les éléments sont

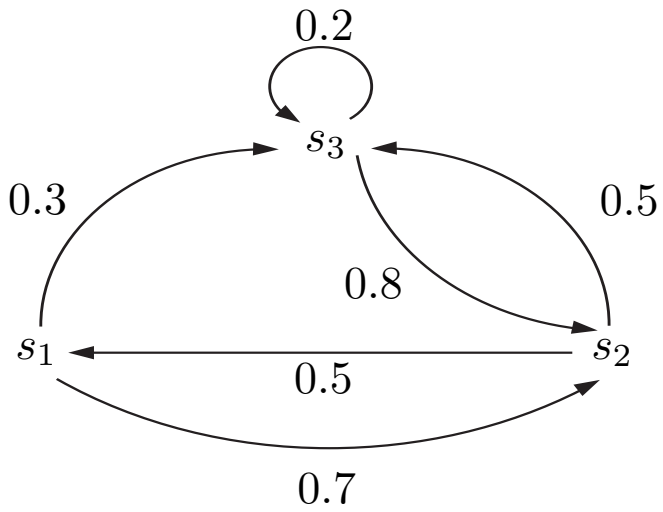
$$\Pi(i, j) = P(S_2 = s_j|S_1 = s_i)$$

Évidemment, on a $\sum_{j=1}^{q} \Pi(i, j) = 1$ et $\Pi(i, j) \geq 0$.

Source discrète en temps discret

Exemple de source de Markov

On considère la source markovienne suivante :



La matrice de transition de celle-ci s'écrit ainsi :

$$\Pi = \begin{pmatrix} 0 & 0.7 & 0.3 \\ 0.5 & 0 & 0.5 \\ 0 & 0.8 & 0.2 \end{pmatrix}$$

Source discrète en temps discret

Source de Markov en régime permanent

Définition 3 (régime permanent - version 1)**.** *Une source markovienne S atteint un régime permanent si*

$$\lim_{n\to\infty} P(S_n = s_i)$$

existe pour tout $i \in \{1, \ldots, Q\}$.

On note $p|_{t\to\infty}$ la distribution limite si elle existe. Sachant que $p|_{t=n} = p|_{t=n-1}\,\Pi$, on a nécessairement

$$p|_{t\to\infty} = p|_{t\to\infty}\,\Pi$$

On dit que $p|_{t\to\infty}$ est une distribution stationnaire puisque l'initialisation de la chaîne de Markov avec celle-ci la rend stationnaire.

Inconvénient. Le régime permanent ainsi défini dépend de la distribution $p|_{t=0}$ initiale. D'autres définitions existent.

Source discrète en temps discret

Source de Markov en régime permanent

Définition 4 (régime permanent - version 2)**.** *Une source markovienne S atteint un régime permanent si*

$$\lim_{n\to\infty} P(S_n = s_i | S_1 = j)$$

existe pour tous $i, j \in \{1, \ldots, Q\}$.

Avantage. Le comportement asymptotique de la source est indépendant de la distribution initiale.

Source discrète en temps discret

Source de Markov d'ordre m

Une source de Markov est caractérisée par une mémoire de taille $m = 1$. Ceci peut être généralisé à des mémoires de taille $m > 1$.

Définition 5 (Source markovienne de taille m)**.** *Une source S est dite markovienne de taille m si elle satisfait à*

$$P(S_{n+1} = s_{i_{n+1}} | S_n = s_{i_n}, \ldots, S_1 = s_{i_1}) = P(S_{n+1} = s_{i_{n+1}} | S_{n-m} = s_{i_{n-m}}, \ldots, S_n = s_{i_n})$$

pour tous symboles $s_{i_1}, \ldots, s_{i_{n+1}}$ de $\mathcal{A}$.

Remarque. Une source markovienne de taille m peut être ramenée à une source markovienne de taille 1 en considérant *une extension d'ordre m* ou plus de celle-ci.

Modèles de source discrète

Entropie d'une source stationnaire

Une source quelconque émet un symbole selon une loi qui peut dépendre des symboles qui l'ont précédé. La définition de l'entropie doit en tenir compte.

Définition 6 (Entropie d'une source stationnaire - version 1)**.** *L'entropie d'une source S stationnaire est définie par :*

$$H_0 \triangleq \lim_{n \to +\infty} H(S_n | S_1, \ldots, S_{n-1}).$$

Validité. Cette définition n'a de sens que si la limite existe.

Modèles de source discrète

Entropie d'une source stationnaire

Validation de la définition. Il convient de s'assurer de l'existence de la limite

$$\lim_{n \to +\infty} H(S_n | S_1, \ldots, S_{n-1})$$

Le conditionnement d'une variable aléatoire diminuant son entropie, on a :

$$0 \leq H(S_n | S_1, S_2, \ldots, S_{n-1}) \leq H(S_n | S_2, \ldots, S_{n-1}) \leq \ldots \leq H(S_n).$$

Puisque la source considérée est stationnaire, on peut écrire :

$$H(S_n) = H(S_1) \qquad H(S_n | S_{n-1}) = H(S_2 | S_1) \qquad \ldots$$

L'inégalité peut donc être remplacée par :

$$0 \leq H(S_n | S_1, \ldots, S_{n-1}) \leq H(S_{n-1} | S_1, \ldots, S_{n-2}) \leq \ldots \leq H(S_1).$$

La suite $\{H(S_n | S_1, \ldots, S_{n-1})\}_{n \geq 1}$ est décroissante et minorée. Elle est donc convergente, assurant la validité de la définition dans le cas stationnaire.

Modèles de source discrète

Entropie d'une source stationnaire

Définition 7 (Entropie d'une source quelconque - version alternative)**.** *L'entropie d'une source S stationnaire est définie par :*

$$H_0 \triangleq \lim_{n \to +\infty} \frac{H(S_1, \ldots, S_n)}{n}.$$

Les deux définitions proposées sont équivalentes dans le cas stationnaire. En effet, il résulte de l'égalité suivante

$$H(S_1, \ldots, S_n) = H(S_1) + H(S_2|S_1) + \ldots + H(S_n|S_1, \ldots, S_{n-1})$$

que $H(S_1, \ldots, S_n)/n$ est la moyenne arithmétique des n premiers termes de la suite $H(S_1)$, $H(S_2|S_1)$, $\ldots$, $H(S_n|S_1, \ldots, S_{n-1})$. Le théorème présenté ci-dessous conduit directement au résultat.

Théorème de Cesaro. Si $a_n \underset{n \to \infty}{\longrightarrow} a$, alors $\frac{1}{n} \sum_{k=1}^{n} a_k \underset{n \to \infty}{\longrightarrow} a$

Modèles de source discrète

Exemples d'entropies de sources stationnaires

Exemple 1. Dans le cas d'une source sans mémoire, caractérisée par des S_i indépendants et distribués selon une même loi, on a :

$$H_0 = H(S_1).$$

Exemple 2. Si S désigne une source markovienne invariante dans le temps, l'entropie de celle-ci est donnée par :

$$H_0 = H(S_2|S_1).$$

Caractérisation d'un codage

Vocabulaire par l'exemple

Le codage de source consiste à associer à chaque symbole s_i d'une source une séquence d'éléments de l'alphabet q-aire de destination, appelée *mot du code*.

Exemple 1. Codes ASCII (7 bits) et ASCII étendu (8 bits), code Morse, etc.

Exemple 2.

	code A	code B	code C	code D	code E	code F	code G
s_1	1	0	00	0	0	0	0
s_2	1	10	11	10	01	10	10
s_3	0	01	10	11	011	110	110
s_4	0	11	01	110	0111	1110	111

Caractérisation d'un codage

Vocabulaire

Régularité. Un code est dit régulier, ou encore non-singulier si tous les mots de code sont distincts.

Déchiffrabilité. Un code régulier est dit déchiffrable, ou encore à décodage unique, si toute suite de mots de code ne peut être interprétée que de manière unique.

> **Longueur fixe.** Avec des mots de longueur fixe, on peut décoder tout message sans ambiguïté.
>
> **Séparateur.** On consacre un symbole de l'alphabet de destination comme séparateur de mot.
>
> **Sans préfixe.** On évite qu'un mot du code soit identique au début d'un autre mot. Un tel code est qualifié de *code instantané*.

Exercice. Caractériser les codes A à G.

Vers le premier théorème de Shannon

Inégalité de Kraft

On se propose de construire des codes déchiffrables, et plus particulièrement instantanés, aussi économiques que possible. L'inégalité de Kraft fournit une condition nécessaire et suffisante d'existence de codes instantanés.

Théorème 1 (Inégalité de Kraft)**.** *On note $n_1, \ldots, n_Q$ les longueurs des mots candidats pour coder les Q états d'une source dans un alphabet q-aire. Une condition nécessaire et suffisante d'existence d'un code instantané ayant ces longueurs de mots est donnée par :*

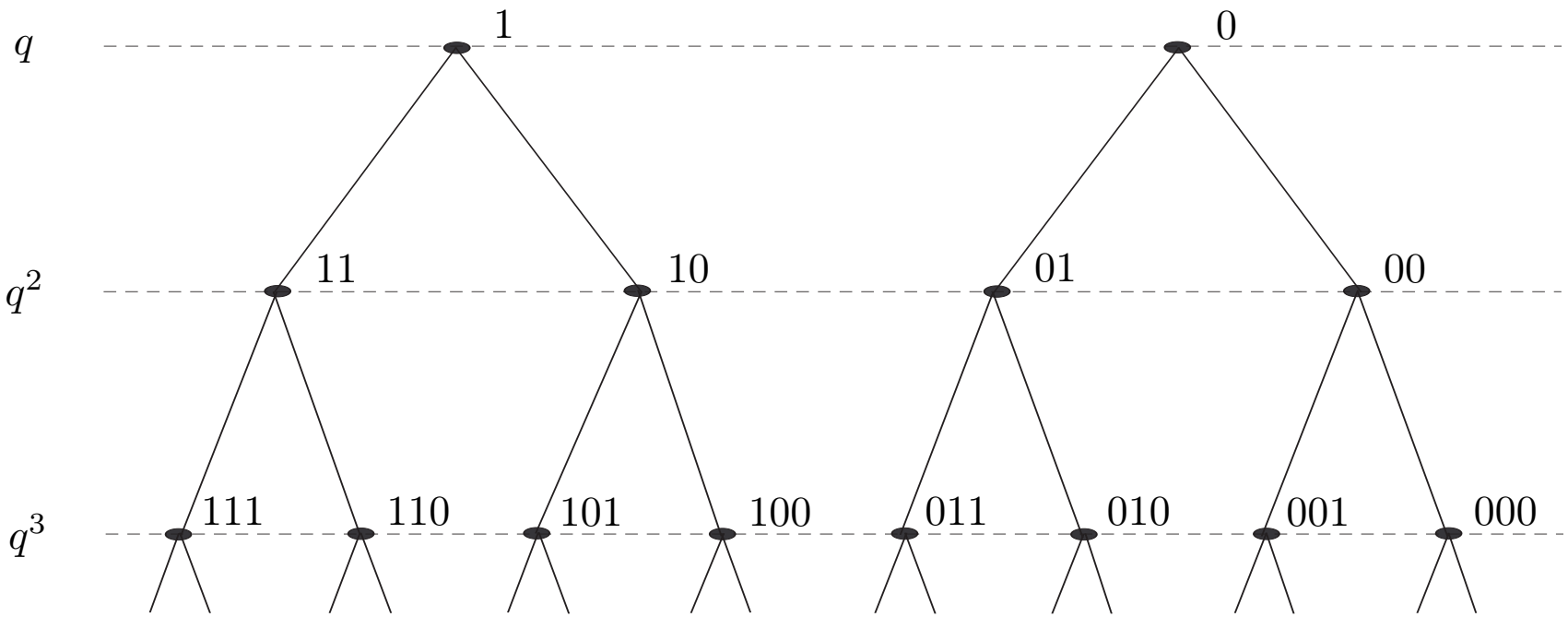
$$\sum_{i=1}^{Q} q^{-n_i} \leq 1.$$

Remarque. La même condition nécessaire et suffisante a été établie par McMillan pour les codes déchiffrables, antérieurement à l'inégalité de Kraft.

Vers le premier théorème de Shannon

Inégalité de Kraft

Preuve. La représentation graphique suivante, dans le cas d'un code binaire, rend la démonstration plus aisée.



Vers le premier théorème de Shannon

Inégalité de Kraft

On pose $n_1 \leq \ldots \leq n_Q$ et on considère un arbre q-aire de profondeur n_Q, comportant donc q^{n_Q} sommets terminaux.

Condition nécessaire. La condition du préfixe impose qu'un mot de longueur n_i exclut $q^{n_Q - n_i}$ sommets terminaux. Le nombre total de sommets exclus vaut donc :

$$\sum_{i=1}^{Q} q^{n_Q - n_i} \leq q^{n_Q}.$$

Condition suffisante. On sélectionne d'abord un nœud à la profondeur n_1, ce qui exclut $q^{n_Q - n_1}$ sommets terminaux. Il en existe toutefois encore car l'inégalité de Kraft entraîne $q^{n_Q - n_1} < q^{n_Q}$. Sur le trajet menant à l'un des sommets terminaux non-exclus, on sélectionne un nœud à la profondeur n_2...

Vers le premier théorème de Shannon

Inégalité de McMillan

L'inégalité de Kraft induit le caractère suffisant de l'inégalité de McMillan puisque tout code à préfixe est déchiffrable.

Condition nécessaire de l'inégalité de McMillan. On développe l'expression suivante selon

$$\left(\sum_{k=1}^{s} r_k\, q^{-k}\right)^m = \sum_{n=m}^{ms} \nu(n)\, q^{-n}$$

où $\nu(n) = \sum_{i_1+\dots i_m=n} r_{i_1} \dots r_{i_m}$. En interprétant r_k comme le nombre de mots de longueur k du code, $\nu(n)$ désigne le nombre de texte de longueur n. La condition de déchiffrabilité implique que $\nu(n) \leq q^n$. On a donc

$$\sum_{k=1}^{s} r_k\, q^{-k} \leq (m\, s)^{\frac{1}{m}},$$

ce qui conduit au résultat en passant à la limite.

Vers le premier théorème de Shannon

Inégalité de McMillan

Définition 8 (Code complet)**.** *Un code est dit complet s'il vérifie la relation*

$$\sum_{i=1}^{Q} q^{-n_i} = 1.$$

Vers le premier théorème de Shannon

Inégalité de McMillan

A titre d'exemple, on applique l'inégalité de McMillan à différents codes.

	code A	code B	code C
s_1	00	0	0
s_2	01	100	10
s_3	10	110	110
s_4	11	111	11
$\sum_{i=1}^{4} 2^{-n_i}$	1	7/8	9/8

Les codes A et B sont déchiffrables, le premier étant complet. Le code C n'est pas déchiffrable.

Vers le premier théorème de Shannon

Conséquences de l'inégalité de McMillan

Soit S une source sans mémoire à Q états. Soit p_i la probabilité d'apparition de s_i, auquel est associé un mot de code déchiffrable q-aire de longueur n_i. En posant

$$q_i = \frac{q^{-n_i}}{\sum_{j=1}^{Q} q^{-n_j}},$$

puis en appliquant l'inégalité de Gibbs à p_i et q_i, on obtient alors

$$\sum_{i=1}^{Q} p_i \log \frac{1}{p_i} + \sum_{i=1}^{Q} p_i \log q^{-n_i} \leq \log \sum_{i=1}^{Q} q^{-n_i}.$$

En appliquant le théorème de McMillan au dernier membre de l'inégalité, il en résulte finalement

$$H(S) - \overline{n} \log q \leq \log \sum_{i=1}^{Q} q^{-n_i} \leq 0,$$

où $\overline{n} = \sum_{i=1}^{Q} p_i \, n_i$ représente la longueur moyenne des mots du code.

Vers le premier théorème de Shannon

Conséquences de l'inégalité de McMillan

Théorème 2. *La longueur moyenne $\overline{n}$ des mots de tout code déchiffrable est bornée inférieurement selon*

$$\frac{H(S)}{\log q} \leq \overline{n}.$$

Condition d'égalité. L'inégalité ci-dessus se transforme en égalité à condition que $\sum_{i=1}^{Q} q^{-n_i} = 1$, c'est-à-dire si $p_i = q^{-n_i}$. Ceci signifie que

$$n_i = \frac{\log \frac{1}{p_i}}{\log q}.$$

Définition 9. *Un code dont la longueur de chaque mot est telle que $n_i = \frac{\log \frac{1}{p_i}}{\log q}$ est dit absolument optimum.*

Vers le premier théorème de Shannon

Conséquences de l'inégalité de McMillan

La condition d'égalité précédente n'est généralement pas vérifiée. Il est cependant possible de constituer un code tel que

$$\frac{\log \frac{1}{p_i}}{\log q} \leq n_i < \frac{\log \frac{1}{p_i}}{\log q} + 1.$$

En multipliant par p_i et en sommant sur i, ceci signifie que

$$\frac{H(S)}{\log q} \leq \overline{n} < \frac{H(S)}{\log q} + 1.$$

Définition 10 (Codes compact et de Shannon)**.** *Un code dont la longueur moyenne des mots vérifie la double inégalité présentée ci-dessus est dit compact. Plus particulièrement, on parle de code de Shannon lorsque*

$$n_i = \left\lceil \frac{\log \frac{1}{p_i}}{\log q} \right\rceil .$$

Premier théorème de Shannon

Énoncé et démonstration

Les bornes qui viennent d'être établies vont nous permettre de démontrer le premier théorème de Shannon, qui s'énonce ainsi :

Théorème 3. *Pour toute source stationnaire, il existe un procédé de codage déchiffrable où la longueur moyenne des mots est aussi voisine que l'on veut de sa borne inférieure.*

Preuve pour une source sans mémoire. On considère la $k^{\text{ème}}$ extension de la source S. Dans le cas d'une source sans mémoire

$$\frac{kH(S)}{\log q} \leq \overline{n}_k < \frac{kH(S)}{\log q} + 1.$$

Dans cette expression, $\overline{n}_k$ désigne la longueur moyenne des mots de code utilisés dans le cadre de la $k^{\text{ème}}$ extension de S. On divise par k et on passe à la limite.

Premier théorème de Shannon

Énoncé et démonstration

Preuve pour une source stationnaire. On considère la $k^{\text{ème}}$ extension de la source S. Dans le cas d'une source sans mémoire

$$\frac{H(S_1,\ldots,S_k)}{k\,\log q} \leq \frac{\overline{n}_k}{k} < \frac{H(S_1,\ldots,S_k)}{k\,\log q} + \frac{1}{k}.$$

Dans cette expression, $\overline{n}_k$ désigne la longueur moyenne des mots de code utilisés dans le cadre de la $k^{\text{ème}}$ extension de S.

Dans le cas d'une source stationnaire, on sait que $\lim_{k\to\infty} H(S_1,\ldots,S_k)$ existe. En reprenant la notation conventionnelle H_0 de cette limite, on aboutit à

$$\lim_{k\to\infty} \frac{\overline{n}_k}{k} = \frac{H_0}{\log q}.$$

Remarque. D'un point de vue pratique, l'intérêt du Premier Théorème de Shannon est limité.

Techniques de codage binaire

Méthode directe

Le premier théorème de Shannon exprime une propriété asymptotique du langage, mais ne fournit aucune méthode pratique pour y parvenir.

Une technique de codage directe consiste à associer à chaque état de la source un nombre de symboles n_i tel que

$$n_i = \left\lceil \frac{\log \frac{1}{p_i}}{\log q} \right\rceil .$$

Remarque. Le code obtenu est un code de Shannon.

Techniques de codage binaire

Méthode directe

On considère un système à 5 états $\{s_1, \ldots, s_5\}$ définis par les probabilités :

$$\begin{array}{lll} p_1 = 0.35 & -\log_2 p_1 = 1.51 \longrightarrow & n_1 = 2 \\ p_2 = 0.22 & -\log_2 p_2 = 2.18 \longrightarrow & n_2 = 3 \\ p_3 = 0.18 & -\log_2 p_3 = 2.47 \longrightarrow & n_3 = 3 \\ p_4 = 0.15 & -\log_2 p_4 = 2.73 \longrightarrow & n_4 = 3 \\ p_5 = 0.10 & -\log_2 p_5 = 3.32 \longrightarrow & n_5 = 4. \end{array}$$

Il est aisé d'obtenir un code instantané vérifiant la condition précédente sur les n_i à l'aide d'un arbre. On obtient par exemple :

$$s_1 : 00 \qquad s_2 : 010 \qquad s_3 : 011 \qquad s_4 : 100 \qquad s_5 : 1010.$$

On aboutit à $\overline{n} = 2.75$, à comparer à $H(S) = 2.19$ Sh/symb.

Techniques de codage binaire

Code de Shannon-Fano

Le code de Shannon-Fano est le premier code à avoir exploité la redondance d'une source. On en expose à présent le principe.

1. Ranger les états du système par probabilités décroissantes.
2. Subdiviser les états du système en 2 groupes G_0 et G_1 de probabilités voisines, *sans modifier l'ordre* dans lequel ils ont été rangés en 1.
3. Chaque groupe G_i est subdivisé en 2 sous-groupes G_{i0} et G_{i1} de probabilités aussi voisines que possibles, une fois encore *sans modifier l'ordre* des états.
4. La procédure s'arrête lorsque chaque sous-groupe est constitué d'un unique élément. L'indice du groupe donne le mot de code.

Techniques de codage binaire

Code de Shannon-Fano

Pour élaborer un code de Shannon-Fano, on procède ainsi :

état	p_i	étape 1	étape 2	étape 3	code
s_1	0.35	0	0		00
s_2	0.22	0	1		01
s_3	0.18	1	0		10
s_4	0.15	1	1	0	110
s_5	0.10	1	1	1	111

On aboutit à $\overline{n} = 2.25$, à comparer à $H(S) = 2.19$ Sh/symb.

Techniques de codage binaire

Code de Huffman

La méthode de Huffman fournit un code instantané compact de longueur moyenne minimale. Pour y parvenir, elle exploite la propriété suivante.

Lemme 1. *Pour toute source, il existe un code instantané de longueur moyenne minimale satisfaisant les propriétés suivantes.*

1. *Si* $P(S = s_i) > P(S = s_j)$, *alors* $n_i \leq n_j$.
2. *Les deux mots les plus longs, donc associés aux états les moins probables, ont même longueur et ne diffèrent que d'un bit.*

La méthode de Huffman consiste à regrouper les deux états les moins probables, puis à les traiter comme un seul en sommant leur probabilité. Cette technique est alors réitérée sur les états restants, jusqu'à ce qu'il n'en reste que deux.

Techniques de codage binaire

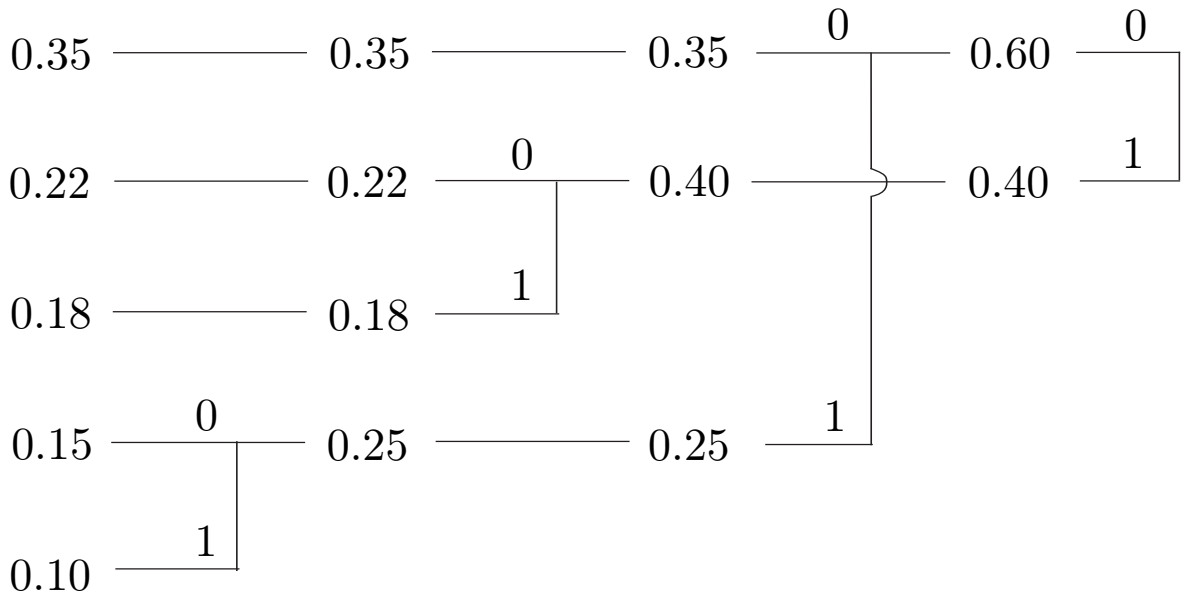
Code de Huffman

On construit un arbre en partant des feuilles les plus profondes, qui représentent les états de la source.

1. A chaque étape, on fusionne les feuilles les moins probables en une seule.
2. La procédure s'arrête lorsque on aboutit à une feuille unique constituée de tous les symboles.
3. Le parcours inverse de l'arbre fournit les mots du code.

Techniques de codage binaire

Code de Huffman



Finalement, le parcours inverse de l'arbre fournit le résultat suivant :

$$s_1 : 00 \qquad s_2 : 10 \qquad s_3 : 11 \qquad s_4 : 010 \qquad s_5 : 011.$$

On aboutit à $\overline{n} = 2.25$, à comparer à $H(S) = 2.19$ Sh/symb.