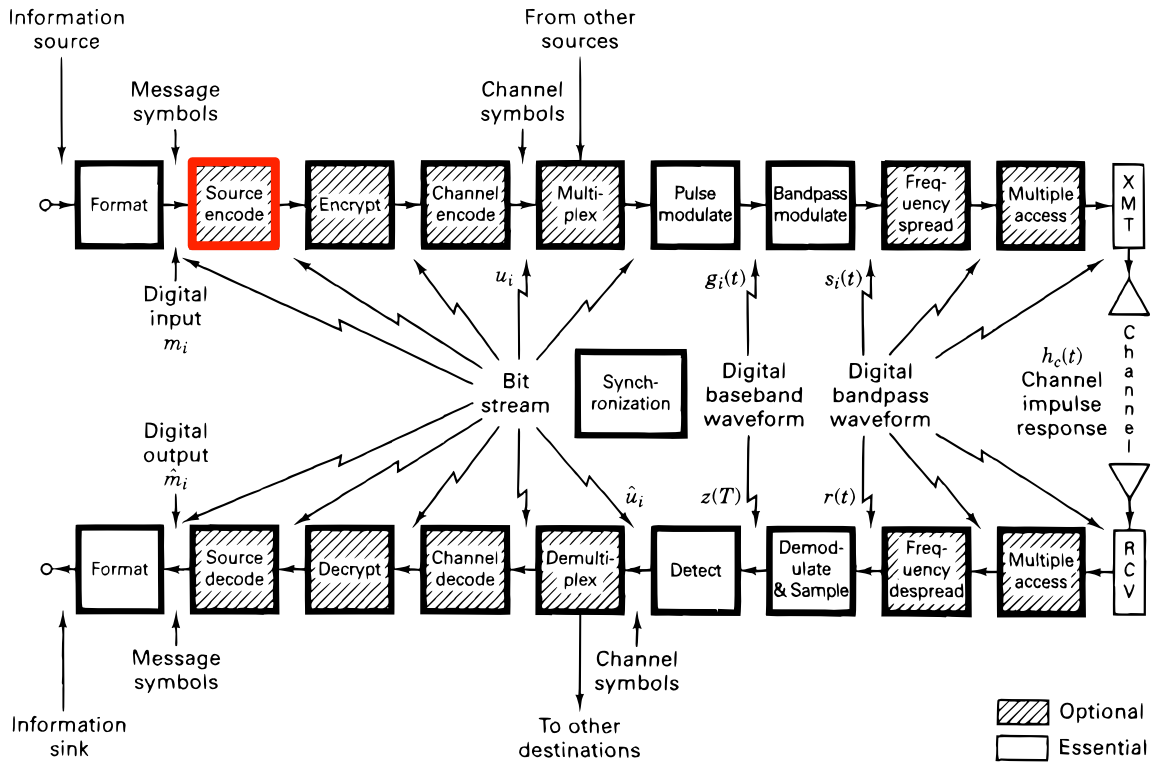
Codage de Source

Transmissions Numériques

Cédric RICHARD
Université Nice Sophia Antipolis

Codage de source



Digital Communications, B. Sklar, Prentice Hall

Codage de source

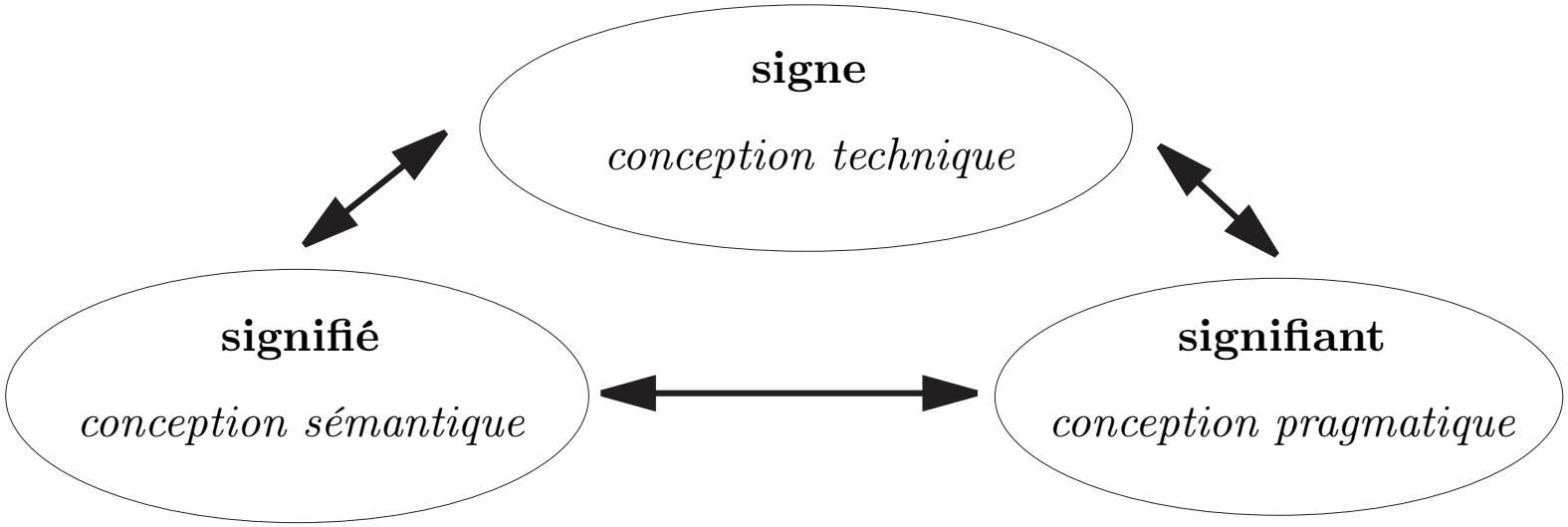
Plusieurs conceptions de l'information

La notion d'information diffère selon qu'on se place du côté de la machine ou de l'individu. La conception analytique en rend compte.

Conception analytique de l'information.

Codage de source

Plusieurs conceptions de l'information



Conception systémique de l'information.

Codage de source

Objectifs de la théorie de l'information

Priorités du système informatique

Importance du signe prépondérante dans le traitement, le stockage et la transmission.

Priorités du système d'information

Aspects sémantiques et pragmatiques privilégiés.

Exemple : la facturation électronique

Remplace ou accompagne la facturation classique ?
Nombre de signes échangés, flux de données ?

La théorie de l'information s'intéresse au signe.

Codage de source

Les origines de la théorie de l'information (1928 – ...)

Travaux de H. Nyquist pour la théorie des communications

- ▷ Liens entre bande passante et vitesse d'émission.
- ▷ Etude des distorsions inter-symboles.

Travaux de R.V. Hartley

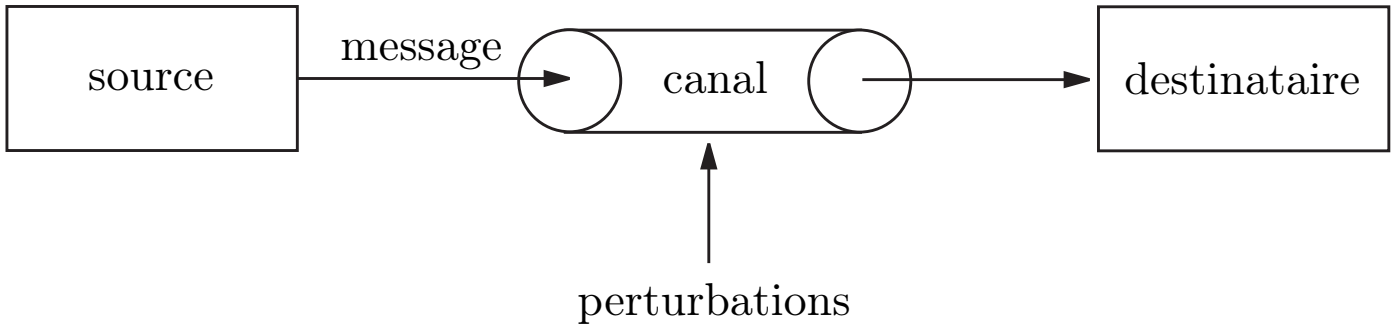
- ▷ Une définition de la notion d'information.

Oeuvre de C.E. Shannon

- ▷ Performances limites en présence de perturbations.
- ▷ Notions de source d'information et de canal de transmission.

Codage de source

Modèle de communication : le paradigme de Shannon



▷ source : générateur de *message*.

▷ message : suite de *symboles* d'un *alphabet* donné.

▷ canal : vecteur de l'information entre *source* et *destinataire*.

▷ perturbations : stochastiques par nature.

Codage de source

Quantité d'information propre d'un événement

Soit A un événement de probabilité $P(A)$ non-nulle.

L'information $h(A)$ apportée par la réalisation de A est d'autant plus grande qu'elle est improbable. Elle peut s'exprimer ainsi :

$$h(A) = f\left(\frac{1}{P(A)}\right).$$

La fonction $f(\cdot)$ vérifie les contraintes suivantes :

- ▷ $f(\cdot)$ est croissante
- ▷ info. apportée par 1 événement sûr est nulle : $\lim_{p\to 1} f(p) = 0$
- ▷ info. apportée par 2 événements indépendants : $f(p_1 \cdot p_2) = f(p_1) + f(p_2)$

Ceci nous conduit à utiliser la fonction logarithmique pour $f(\cdot)$

Codage de source

Quantité d'information propre d'un événement

Lemme 1. *La fonction* $f(p) = -\log_b p$ *est la seule qui soit à la fois positive, continue sur* $]0,1]$*, et qui vérifie* $f(p1 \cdot p2) = f(p1) + f(p2)$.

Preuve. La démonstration comporte les étapes suivantes :

1. $f\left(p^n\right) = n\, f(p)$
2. $f\left(p^{1/n}\right) = \frac{1}{n}\, f(p)$ après avoir remplacer p par $p^{1/n}$
3. $f\left(p^{m/n}\right) = \frac{m}{n}\, f(p)$ en combinant le deux égalités précédentes
4. $f(p^q) = q\, f(p)$ où q désigne un nombre rationnel positif quelconque
5. $f\left(p^r\right) = \lim_{n\to+\infty} f\left(p^{q_n}\right) = \lim_{n\to+\infty} q_n\, f(p) = r\, f(p)$

Soient p et q appartenant à $]0,1[$. On peut écrire $p = q^{\log_q p}$, ce qui entraîne

$$f(p) = f\left(q^{\log_q p}\right) = f(q)\, \log_q p.$$

On aboutit finalement au résultat escompté, soit

$$\boldsymbol{f(p) = -\log_b p}$$

Codage de source

Quantité d'information propre d'un événement

Définition 1. *Soit A un événement de probabilité $P(A)$ non-nulle. On associe à la réalisation de A la quantité d'information propre :*

$$h(A) = -\log P(A).$$

Unités. L'unité dépend de la base choisie pour le logarithme.

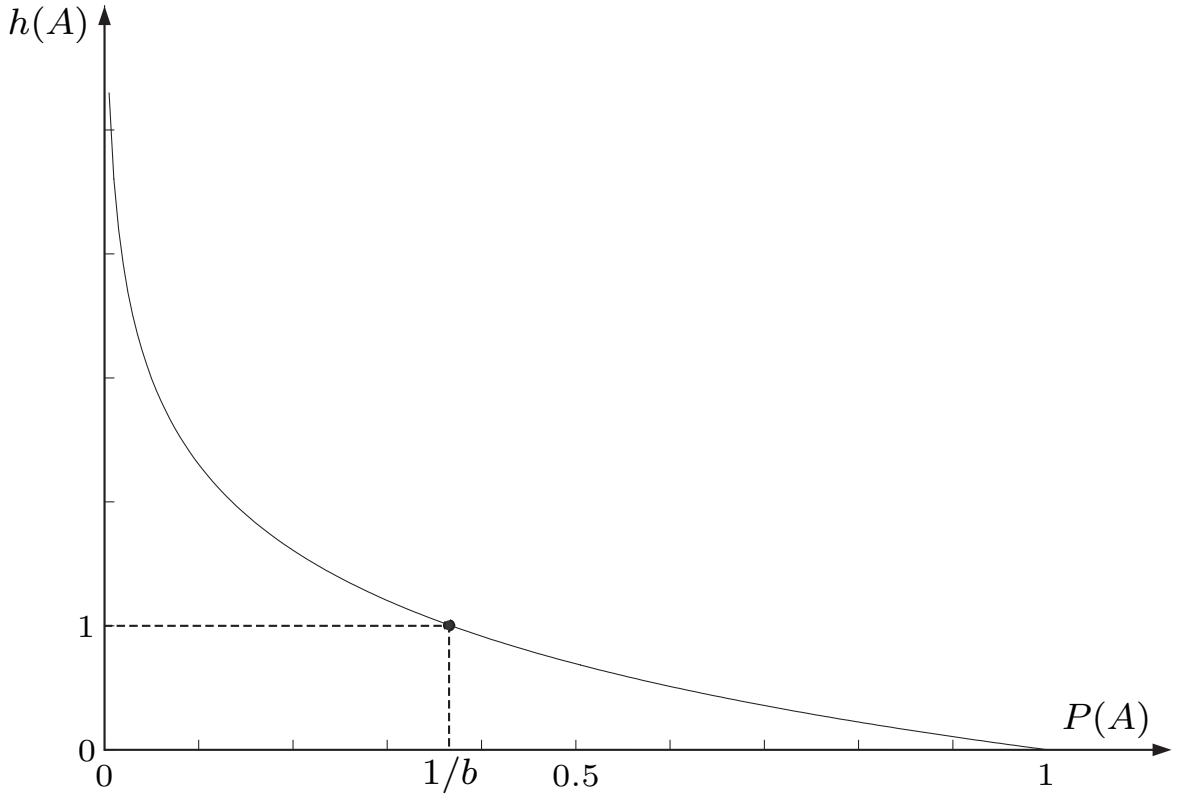
- ▷ $\log_2$: Shannon, bit (binary unit)
- ▷ $\log_e$: logon, nat (natural unit)
- ▷ $\log_{10}$: Hartley, decit (decimal unit)

Vocabulaire. $h(\cdot)$ est désigné par *incertitude* ou encore *quantité d'information*.

Codage de source

Quantité d'information propre d'un événement

Quantité d'information propre ou incertitude : $h(A) = -\log_b P(A)$



Codage de source

Entropie d'une source

Soit une source S d'information sans mémoire sélectionnant aléatoirement un symbole parmi les n éléments d'un alphabet $\mathcal{S} = \{s_1, \ldots, s_n\}$. Soit p_i la probabilité d'apparition de s_i. La quantité d'information moyenne associée à l'apparition de chaque symbole possible est donnée par :

$$H(S) = E\{h(s)\} = -\sum_{i=1}^{n} p_i \log p_i.$$

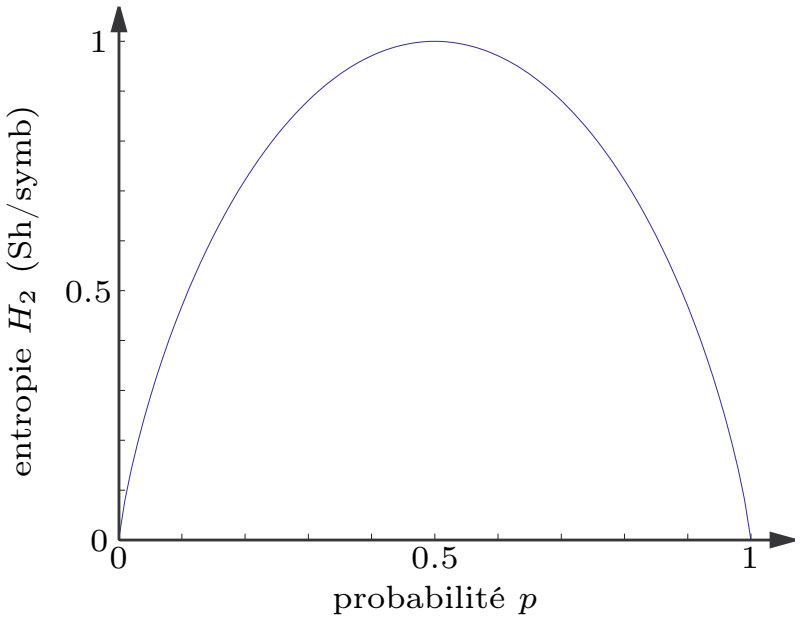
L'entropie est une quantité d'information moyenne.

Codage de source

Exemple : entropie d'une source binaire

L'entropie d'une source binaire est donnée par :

$$H(S) = -p \log p - (1-p) \log(1-p) \triangleq H_2(p).$$



Lemme 2 (Inégalité de Gibbs)**.** *Etant donné 2 distributions de probabilité discrètes* $(p_1, \ldots, p_n)$ *et* $(q_1, \ldots, q_n)$ *sur un même univers fini, on a :*

$$\sum_{i=1}^{n} p_i \log \frac{q_i}{p_i} \leq 0,$$

l'égalité étant obtenue lorsque $\forall i : p_i = q_i$*.*

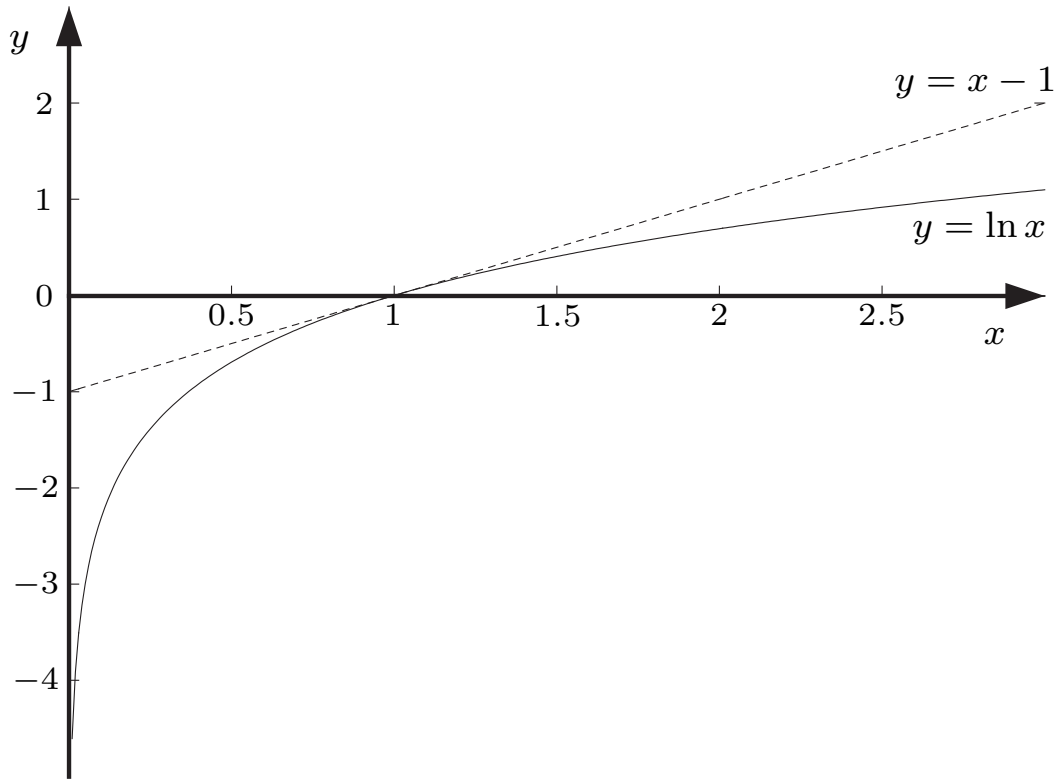
Preuve. On effectue la démonstration dans le cas du logarithme népérien et on note que $\ln x \leq x - 1$, l'égalité étant obtenue pour $x = 1$. On pose $x = \frac{q_i}{p_i}$ et on a

$$\sum_{i=1}^{n} p_i \ln \frac{q_i}{p_i} \leq \sum_{i=1}^{n} p_i (\frac{q_i}{p_i} - 1) = 1 - 1 = 0.$$

Codage de source

Entropie : notation et propriété préalables

Vérification graphique de l'inégalité $\ln x \leq x - 1$



Codage de source

Quelques propriétés de l'entropie

Propriété 1. *L'entropie vérifie l'inégalité suivante*

$$H_n(p_1, \ldots, p_n) \leq \log n,$$

l'égalité étant réalisée dans le cas d'une loi uniforme, c'est-à-dire $\forall i : p_i = \frac{1}{n}$.

Preuve. A partir de l'inégalité de Gibbs, on pose $q_i = \frac{1}{n}$. L'incertitude sur le résultat d'une expérience est d'autant plus grande que tous les résultats possibles sont équiprobables.

Codage de source

Entropie conjointe

Définition 2. *Soient S et T deux sources d'alphabets $\{s_1, \ldots, s_n\}$ et $\{t_1, \ldots, t_m\}$. L'entropie conjointe de S et T est donnée :*

$$H(S,T) \triangleq -\sum_{i=1}^{n}\sum_{j=1}^{m} P(S=s_i, T=t_j)\, \log P(S=s_i, T=t_j).$$

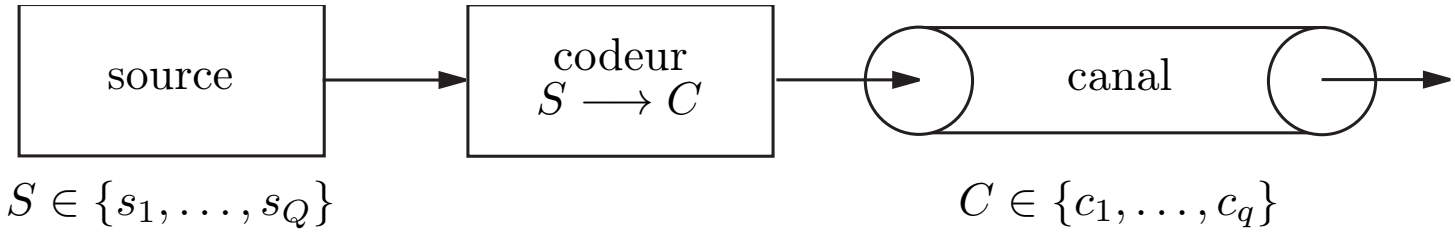
▷ l'entropie conjointe est une grandeur symétrique : $H(S,T) = H(T,S)$

Exemple. Cas de sources indépendantes.

Codage de source

Adaptation d'une source à un canal non-bruité

On associe à chacun des Q états s_i d'une source $S \in \{s_i, \ldots, s_Q\}$ 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 désigne par $C \in \{c_1, \ldots, c_q\}$.



Exemple. Le code Morse

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

Codage de source

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 ≤ 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 ! ! !

Codage de source

Modèle général de source

Une source discrète est définie par un alphabet $\mathcal{S} = \{s_1, \ldots, s_Q\}$ 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**

Codage de source

Hypothèses complémentaires

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

Propriété 2 (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.

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

Codage de source

Source markovienne

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

Définition 3 (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 $\mathcal{S}$.

Il en résulte directement que

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

Codage de source

Source markovienne

Définition 4 (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

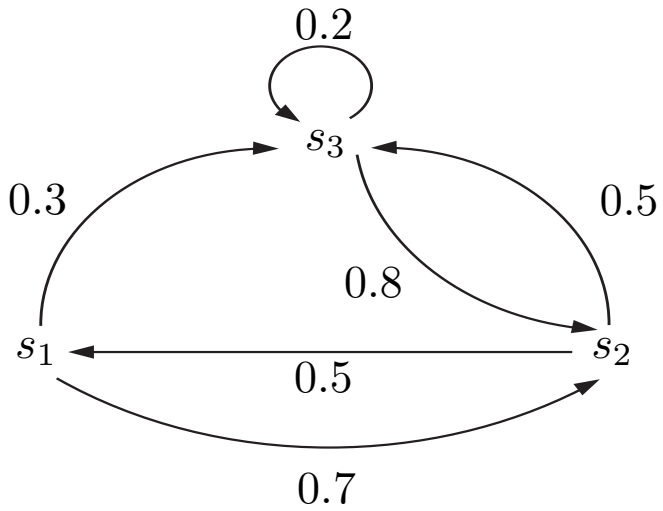
$$\Pi(i, j) = P(S_2 = s_j|S_1 = s_i)$$

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

Codage de source

Exemple de source markovienne

On considère la source markovienne suivante :



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

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

Codage de source

Entropie d'une source stationnaire

Définition 5. *L'entropie d'une source S stationnaire est définie par :*

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

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 :

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

Codage de source

Caractérisation d'un code : 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

Codage de source

Caractérisation d'un code

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.

Codage de source

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.

Codage de source

Code complet

Définition 6 (Code complet)**.** *Un code est dit complet s'il vérifie la relation*

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

Codage de source

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.

Codage de source

Vers le premier théorème de Shannon

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.

Codage de source

Vers le premier théorème de Shannon

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 7. *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.*

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 8 (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*

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

Codage de source

Premier théorème de Shannon

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.

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.

Codage de source

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.

Codage de source

Code de Shannon

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

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

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.

Codage de source

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.

Codage de source

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.

Codage de source

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 3. *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.

Codage de source

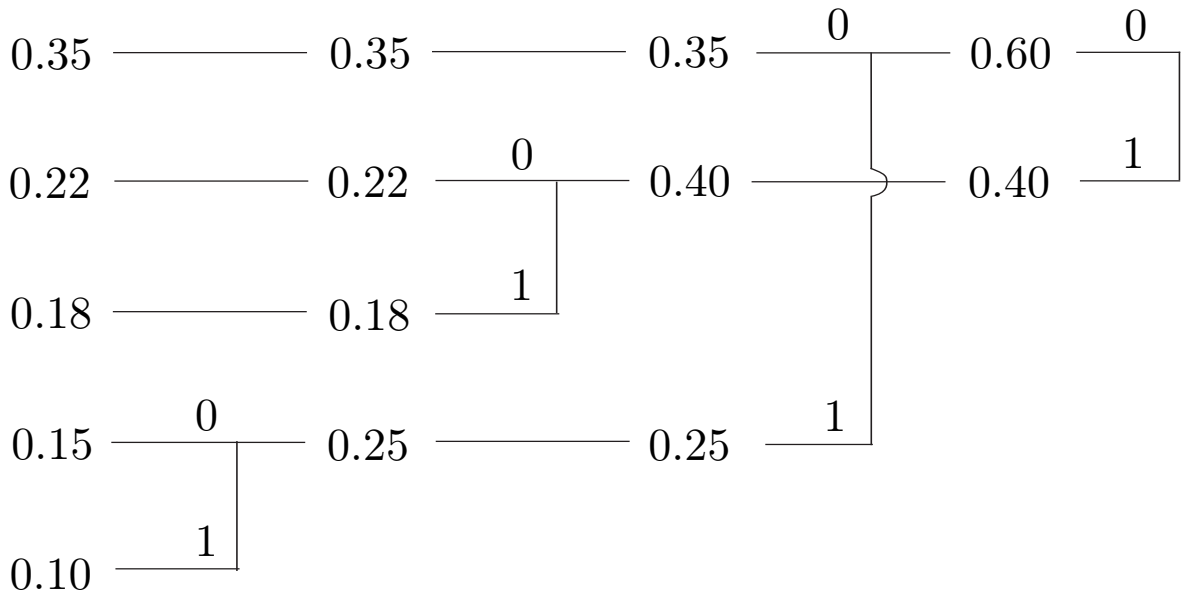
Code de Huffman

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.

Codage de source

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.