Théorie de l'Information
Notes de Cours (part 4)
2006-2007

SIC-SICOM

Maria-João Rendas

October 27, 2006

Contents

Chapter 4

Taux d'entropie (Entropy rate)

Dans le Chapitre 3 nous avons étudié le Théorème du Codage Source pour des sources blanches, c'est à dire, dont les symboles produits par la source sont statistiquement indépendants et identiquement distribuées. Dans ce Chapitre nous allons généraliser ce résultat pour des **sources avec mémoire**, en faisant appel à la notion de *taux d'entropie*. Nous introduisons également la notion de *codes universels*, et nous présentons l'algorithme de Lempel-Ziv, comme exemple de code universel qui exploite la structure de corrélation de la source.

4.1 Taux d'entropie et Codeur Universel

Définition 1 *Taux d'entropie (*Entropy rate*)*
Soit X_n une source stationnaire (dont la distribution est invariante par rapport à des translations (*shifts*) dans le temps). Son taux d'entropie (*entropy rate*), $\overline{H}(X)$ est, par définition

$$\overline{H}(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1^n), \tag{4.1}$$

où nous utilisons la notation $X_1^n = \{X_1, X_2, \ldots, X_n\}$. △

Propriété 1 Pour une source stationnaire, le limite dans la définition (4.1) existe et est égal à

$$H'(X) = \lim_{n\to\infty} H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1) \tag{4.2}$$

△

Démonstration
Nous allons démontrer dans un premier temps que le limite H' dans l'équation (4.2) existe.

$$0 \quad \leq \quad H(X_n|X_1^{n-1}) = H(X_n|X_{n-1}, \ldots, X_2, X_1)$$

$$\begin{aligned} &\overset{(a)}{\leq} && H(X_n|X_{n-1},\ldots,X_2) \\ &\overset{(b)}{=} && H(X_{n-1}|X_{n-2},\ldots,X_1) = H(X_{n-1}|X_1^{n-2}), \end{aligned}$$

où (a) est justifiée car le conditionnement diminue l'entropie, et (b) par la stationnarité de X_n.
Nous voyons donc que

$$\alpha_n = H(X_n|X_1^{n-1}) \tag{4.3}$$

est une séquence non-croissante ($\alpha_n \leq \alpha_{n-1}$) de nombres non-négatifs ($\alpha_n \geq 0$). Elle doit nécessairement avoir une limite:

$$\lim_{n\to\infty} \alpha_n = H' \,.$$

Nous énonçons maintenant un Lemme qui sera utilisé par la suite pour établir l'égalité entre (4.1) et (4.2).

Lemme 1 *Moyenne de Cesáro*
Soit a_n une séquence et a sa limite :

$$a_n \to a.$$

Soit b_n la séquence :

$$b_n = \frac{1}{n}\sum_{i=1}^{n} a_i.$$

Alors

$$b_n \to a \Leftrightarrow \lim_{n\to\infty} b_n = \lim_{i\to\infty} a_i.$$

△

Démonstration

$$\lim_{n\to\infty} b_n = b \Leftrightarrow \forall\delta > 0 \;\; \exists N(\delta) \;:\; \forall n > N(\delta) \qquad |b_n - b| < \delta.$$

Comme $a_n \to a$:

$$\forall\epsilon > 0, \;\; \exists n(\epsilon) \;: \forall n > n(\epsilon) \qquad |a_n - a| < \epsilon.$$

Soit $n(\epsilon)$ l'ordre dont l'existance est garantie par la convergence de la série a_n. Alors, pour $n \gg n(\epsilon)$,

$$\begin{aligned} |b_n - a| &= \left|\frac{1}{n}\sum_{i=1}^{n} a_i - a\right| \\ &= \left|\sum_{i=1}^{n}\left(\frac{1}{n}a_i - \frac{a}{n}\right)\right| \end{aligned}$$

$$
\begin{aligned}
&= \left|\frac{1}{n}\sum_{i=1}^{n}(a_i - a)\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^{n(\epsilon)}(a_i - a)\right| + \left|\frac{1}{n}\sum_{i=n(\epsilon)+1}^{n}(a_i - a)\right| \qquad (4.4)
\end{aligned}
$$

car $|a+b| \leq |a| + |b|$.
Le premier terme :

$$
\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^{n(\epsilon)}(a_i - a)\right| &\leq \frac{1}{n}\sum_{i=1}^{n(\epsilon)}|a_i - a| \\
&\leq \frac{1}{n}\sum_{i=1}^{n(\epsilon)}\max_{i\leq n(\epsilon)}|a_i - a| \\
&= \max_{i\leq n(\epsilon)}|a_i - a|\,\frac{n(\epsilon)}{n}
\end{aligned}
$$

Pour

$$
n > N^{\star}(\epsilon) = \frac{n(\epsilon)}{\epsilon \max_{i\leq n(\epsilon)}|a_i - a|} \Rightarrow \left|\frac{1}{n}\sum_{i=1}^{n(\epsilon)}(a_i - a)\right| < \epsilon\ .
$$

Pour le deuxième terme dans l'équation (4.4) :

$$
\begin{aligned}
\left|\frac{1}{n}\sum_{i=n(\epsilon)+1}^{n}(a_i - a)\right| &\leq \frac{1}{n}\sum_{i=n(\epsilon)+1}^{n}|a_i - a| \\
&\leq \frac{1}{n}\sum_{i=n(\epsilon)+1}^{n}\max_{i>n(\epsilon)}|a_i - a| \\
&= \max_{i>n(\epsilon)}|a_i - a|\,\frac{n - n(\epsilon)}{n} \\
&\leq \max_{i>n(\epsilon)}|a_i - a| \leq \epsilon
\end{aligned}
$$

par la définition de $n(\epsilon)$.
On peut donc affirmer que $\forall n > N^{\star}(\epsilon)$

$$
|b_n - a| < 2\epsilon.
$$

Si nous prenons $\delta = 2\epsilon$, alors,

$$
\forall \delta > 0\ \ \forall n > N^{\star}(\delta/2) = \frac{n(\delta/2)}{\delta/2\max_{i\leq n(\delta/2)}} \qquad |b_n - a| < \delta
$$

c'est à dire, la série b_n converge, et elle a la même limite que a_n. $\triangle$

Nous pouvons maintenant finir la démonstration de l'égalité entre (4.1) :

$$\overline{H}(X) \overset{(a)}{=} \lim_{n\to\infty} \frac{1}{n} H(X_1^n) \overset{(b)}{=} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i|X_1^{i-1}) \overset{(c)}{=} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \alpha_i,$$

– où nous considérons la définition de α_i de l'équation (4.3) – et le limite intervenant dans l'équation (4.2). Dans l'équation précédante nous avons utilisé : *(a)* la définition de $\overline{H}$; *(b)* la régle de la chaîne pour l'entropie conjointe; *(c)* la définition de la série α_i. L'application du Lemme qui vient d'être énoncé à cette expression conduit à

$$\overline{H}(X) = \lim_{i\to\infty} \alpha_i = H'(X),$$

ce qui complète la démonstration.

Propriété 2 $H(X_1^n)$ satisfait les inégalités suivantes:

$$\frac{1}{n} H(X_1^n) \geq H(X_n|X_{n-1}, \ldots, X_1) \tag{4.5}$$

et

$$\frac{1}{n} H(X_1^n) \leq \frac{1}{n-1} H(X_1, \ldots, X_{n-1}). \tag{4.6}$$

Ces deux inégalités nous disent que la séquence $b_n = \frac{1}{n} H(X_1^n)$ *est décroissante et bornée inférieurement par* $\alpha_n = H(X_n|X_1^{n-1})$.

△

L'inégalité (4.5) peut être démontrée de la façon suivante:

$$\begin{aligned} H(X_1^n) &= \sum_{i=1}^{n} H(X_i|X_1^{i-1}) \\ &\overset{(a)}{=} \sum_{i=1}^{n} H(X_n|X_{n-i-1}^{n-1}) \\ &\overset{(b)}{\geq} \sum_{i=1}^{n} H(X_n|X_1^{n-1}) \\ &= nH(X_n|X_1^{n-1}), \end{aligned}$$

où (a) est justifiée par la stationnarité de la séquence, et (b) par le fait que le conditionnement diminue l'entropie.
L'inégalité (4.6) est facilement obtenue en utilisant (4.5) dans l'expansion de $H(X_1^n)$:

$$\begin{aligned} \frac{1}{n} H(X_1^n) &= \frac{1}{n} \left[H(X_1^{n-1}) + H(X_n|X_1^{n-1}) \right] \\ &\leq \frac{1}{n} \left[H(X_1^{n-1}) + \frac{1}{n} H(X_1^n) \right] \\ \Leftrightarrow (n-1) H(X_1^n) &\leq nH(X_1^{n-1}) \\ \Rightarrow \frac{1}{n} H(X_1^n) &\leq \frac{1}{n-1} H(X_n|X_1^{n-1}). \end{aligned}$$

Théorème 1 *(Codage Source)*
Soit $L_n^\star(X)$ la longueur moyenne (par symbole) d'un code optimal sans pertes pour des séquences de taille n : $X^n = \{X_1, \ldots, X_n\}$. Alors

$$L_n^\star(X) \stackrel{n\to\infty}{\longrightarrow} \overline{H}(X).$$

△

Ce théorème affirme que le taux d'entropie $\overline{H}(X)$ est asymptotiquement (dans le limite de blocs de grande taille ($n \to \infty$)) le *nombre minimal de bits par symbole source* pour coder sans pertes les séquences de la source X_n. Le taux de compression optimal est donc le taux d'entropie de la source.

Définition 2 *Code universel*
Soit C_n un code sans pertes pour des séquences de n symboles source, et ℓ_n la fonction qui décrit la taille des mots de ce code. C_n est un code *universel* si

$$\lim_{n\to\infty} E\frac{1}{n}\ell_n(X^n) = \overline{H}(X),$$

pour *toutes les sources X stationnaires* . △

Nous verrons dans une Section ultérieure qu'il existent effectivement des codes universels, et que le code de Lempel-Ziv est un exemple bien connu de ce type de codes.

La version que nous avons présentée (dans le Chapitre 3) du Théorème du Codage Source, pour des sources i.i.d., est basée dans la Propriété d'équi-répartition asymptotique, qui affirme que si X_n sont des variables i.i.d., alors

$$-\frac{1}{n}\log p(x^n) \stackrel{n\to\infty}{\longrightarrow} H(X),$$

où la convergence est en probabilité, et $H(X)$ est l'entropie de Shanon des variables aléatoires i.i.d. X_n.

Nous avons alors vu que pour n suffisament grand, l'ensemble ϵ-typique défini par

$$A_n^\epsilon = \left\{ x^n : \left| -\frac{1}{n}\log p(x^n) - H(X) \right| < \epsilon \right\}$$

satisfait les relations suivantes:

$$\Pr\{A_n^\epsilon\} \geq 1-\epsilon, \qquad 2^{n(H(X)-\epsilon)} \leq |A_n^\epsilon| \leq 2^{n(H(X)+\epsilon)}.$$

Ces équations affirment qu'il existent $\simeq 2^{nH(X)}$ séquences qui concentrent presque toute la masse de probabilité. Nous pouvons alors concentrer nos efforts de codage sur cet ensemble, en utilisant un code dont la longeur est près de l'entropie de la source, avec une probabilité d'erreur aussi petite que l'on souhaite.

Cependant, la Loi Forte des Grands Nombres, dans laquelle est basée la Propriété d'équi-répartition asymptotique, est valable pour des processus plus généraux que les sources i.i.d.: les processus *ergodiques*, dont nous donnons maintenant la définition.

Définition 3 *Source ergodique*
Soit $X_n = \{\ldots, X_{-1}, X_0, X_1, \ldots\}$, $X_n \in \mathcal{X}$, une source stationnaire, et représentons par $T(X)$ l'opérateur de translation (*shift*):

$$Y = T(\{\ldots, X_{-1}, X_0, X_1, \ldots\}) = \{\ldots, X_0, X_1, X_2, , \ldots\}, \qquad \Rightarrow Y_n = X_{n-1}.$$

Soit $T^k(X)$ la translation de X par k unités de temps:

$$Y = T^k(X) \Rightarrow Y_n = X_{n-k}.$$

La source X_n est *ergodique* si pour toute fonction mesurable $f : \mathcal{X} \to \Re$ avec $E[f(X)] < \infty$

$$\frac{1}{n}\sum_{i=1}^{n} f\left(T^k(X)\right) \overset{n\to\infty}{\longrightarrow} E\left(f(X)\right).$$

△

D'une façon informelle, nous pouvons dire qu'une source est ergodique si sa caractérisation statistique peut être déduite à partir de l'observation d'*une* de ses réalisations (un seul *sample path*).

Définition 4 *Code ponctuellement universel (*pointwise universal code*)*
Un code $C_n(X)$ est ponctuellement (*pointwise*) universel si sa longueur ℓ_n satisfait

$$\lim_{n\to\infty} \frac{1}{n}\ell_n(X^n) \to \overline{H}(X) \qquad \text{w.p. 1,}$$

pour toute source X *stationnaire et ergodique*. △

Remarquez que cette notion de codeur universel implique l'optimalité (asymptotique) du code *pour toute séquence* de la source.

4.2 Taux d'entropie de Chaînes de Markov

Définition 5 *Processus de Markov*
Une série aléatoire (un processus aléatoire à temps discret) X_n est un *processus de Markov* (avec état $X_n \in \mathcal{X}$) si

$$p(X_{n+1} = x_{n+1}|X_n = x_n, \ldots, X_1 = x_1) = p(X_{n+1} = x_{n+1}|X_n = x_n),$$

où $x_n \in \mathcal{X}, n = 1, \ldots, n+1$.
Nous désignons les noyaux $p(X_{n+1} = x_{n+1}|X_n = x_n)$ – distributions conditionnelles de X_{n+1} sachant la valeur de X_n – par probabilités de transition du processus de Markov. △

Ceci est l'expression mathématique de la notion intuitive de "processus sans mémoire." Formulé autrement, nous dirons que le passé ($X_i, i < n$) et le futur ($X_i, i > n$) sont statistiquement indépendants sachant le présent (X_n).

Pour un processus de Markov, il est imédiat de vérifier (application répétée de la loi de Bayes pour la probabilité conditionelle) que la probabilité d'une séquence x^n factorise de la forme suivante:

$$p(x^n) = p(x_1)p(x_2|x_1)\cdots p(x_n|x_{n-1}).$$

Définition 6 *Processus invariant dans le temps*
Un processus de Markov est *invariant dans le temps* si sa probabilité de transition ne dépend pas de n (elle est indépendante de l'origine du temps):

$$p(X_n = x_n|X_{n-1} = x_{n-1}) = p(X_{n+k} = x_n|X_{n+k-1} = x_{n-1}), \forall k\ .$$

△

Définition 7 *Chaîne de Markov*
Un processus de Markov, où l'état X_n prend des valeurs dans un ensemble *fini* $\mathcal{X}, |\mathcal{X}| = m < \infty$, est appelé *Chaîne de Markov*. △

Pour une Chaîne de Markov, $X_n \in \mathcal{X} = \{x_1, \ldots, x_m\}$, avec $|\mathcal{X}| = m$, les probabilités de transition sont spécifiées par des *matrices de transition* P_n, de dimension $m \times m$, qui ont comme élément générique $[P_n]_{ij}$

$$[P_n]_{i,j} = p(X_n = x_i|x_{n-1} = x_j), \qquad x_i \in \mathcal{X}, i, j = 1, \ldots, m. \tag{4.7}$$

La matrice P est une *matrice stochastique*: la somme des éléments de toutes ses colonnes doit être égale à 1. Ces matrices possèdent plusieurs propriétés algébriques intéréssantes, comme nous le verrons par la suite.

La loi de probabilité pour l'état de la Chaîne à l'instant n, X_n, est décrite par un vecteur p_n de dimension m

$$p_n = \begin{bmatrix} p(X_n = x_1) \\ \vdots \\ p(X_n = x_m) \end{bmatrix},$$

qui doit vérifier les conditions suivantes:

$$\sum_{i=1}^{m} [p_n]_i = 1, \qquad [p_n]_i \in [0, 1] \qquad \forall i = 1, \ldots, m,$$

et qui satisfait l'équation de récurrence suivante (loi de la probabilité totale)

$$p_{n+1} = P_{n+1}p_n,$$

où P_{n+1} est la matrice de transition, définie dans (4.7).
Pour une Chaîne de Markov invariante dans le temps, les matrices P_n ne dépendent pas de n: $P_n = P, \forall n$. Dans ce cas, on obtient facilement

$$p_n = P^n p_0, \tag{4.8}$$

où p_0 est la probabilité de l'état initial.

Exemple 1 *Stepping stone model*
Nous présentons dans cet exemple, un modèle de Chaîne de Markov qui a été utilisé en études de génétique. Il modélise l'état d'un tableau de n-par-n carrés, où chaque carré peut prendre une d'entre K couleurs possibles, $X_k \in \{1, \ldots, K\}^{n \times n}$. L'état initial de chaque carré, $X_0(i,j)$, est choisi aléatoirement (égale probabilité de prendre une des K couleurs), indépendamment des autres carrés

$$X_0(i,j) \sim u = \left[\frac{1}{K} \cdots \frac{1}{K}\right], \qquad i,j = 1, \ldots, n.$$

À chaque pas, l'état de chaque carré est modifié en fonction de son *voisinage* $V_{i,j}$. Nous précisons cette notion de voisinage :

$$V_{i,j} = \{(p,q), p \in \{(i-1)_n, (i+1)_n\} \cup \{(p,q), q \in \{(j-1)_n, (j+1)_n\}\}, i = j, i = 1, \ldots, n.$$

où

$$(a)_n = a, \text{ si } a \in \{1, \ldots, n\}, \qquad (0)_n = n, \qquad (n+1)_n = 1.$$

Ceci défini une géométrie de "doughnut" dans le carré (comme si on construisait un cilyndre en collant son coté inférieur à son coté supérieur, et après un "doughnut" en collant les deux frontières circulaires ensemble).
Avec cette définition, l'état de chaque carré est déterminé de la façon suivante. Pour chaque site (i,j) on choisit (avec égale probabilité) un élément $(p,q) \in V_{i,j}$ de son voisinage, et le carré $X_{k+1}(i,j)$ prend la couleur $X_k(p,q)$:

$$X_{k+1}(i,j) = X_k(p,q), \qquad i,j = 1, \ldots, n.$$

Ce modèle est assez facile à simuler, mais l'analyse de sa matrice de transition est difficil (Essayez de spécifier cette matrice, même pour le cas simple de $K = 2$ ("image" binaire) et $n = 3$. Notez que dans ce cas la dimension de l'espace d'états est $K^{n^2} = 2^9 = 512$!). La Figure 4.1 illustre la configuration initiale X_1 et les configurations d'une réalisation de cette Chaîne pour $n = 1, 30, 31$ and 80.

Vous pouvez constater que cette Chaîne tends vers un des k états où tout le tableau a la même couleur (états *absorbants* de la Chaîne). △

Définition 8 *États absorbants et transitoires*
Un état $x_i \in \mathcal{X}$ d'une Chaîne de Markov est *absorbant* si $P_{ii} = 1$, c'est à dire, qu'une fois que la Chîne passe par cet état elle ne peut plus le quitter.

Un état $x \in \mathcal{X}$ est *transitoire* s'il n'est pas absorbant. △

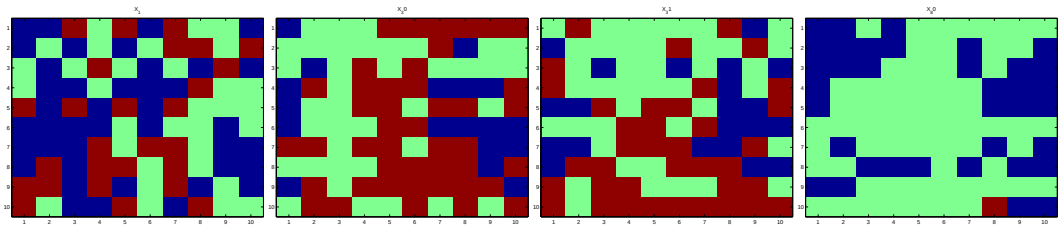


Figure 4.1: Évolution de l'état du modèle de l'exemple 1, pour $k = 1, 30, 31, 80$.

Nottez que $P_{ii} = 1 \Rightarrow P_{ji} = 0, j \neq i$.

Pour la Chaîne de l'exemple 1, il existent K états absorvents :

$$X(i,j) = k, \forall (i,j), k \in \{1, \ldots, K\}$$

(les tableaux de couleur constante). Tous les autres états sont transitoires.

Définition 9 *Chaîne asorbante*
Une Chaîne de Markov est *absorbante* si elle possède *au moins un état absorvent*, et s'il est possible de transiter à partir de tous états transitoires vers un état absorbant de la Chaîne. △

La Chaîne de l'exemple 1 est absorbante.

Définition 10 *Forme canonique de la matrice de transition*
Considérez la re-numérotation des éléments de l'espace d'états $\mathcal{X}$, $|\mathcal{X}| = m$, d'une Chaîne de Markov absorvente avec r états absorbants, de façon que les premiers $m - r$ états soient les états transitoires de la Chaîne. Soit $\mathcal{P} = \{\mathcal{X}_a, \mathcal{X}_t\}$ la partition suivante de $\mathcal{X}$:

$$\mathcal{X} = \mathcal{X}_a \cup \mathcal{X}_t, \qquad \mathcal{X}_t = \{1, \ldots, m - r\}, \qquad \mathcal{X}_a = \{m - r + 1, \ldots, m\}.$$

($\mathcal{X}_t$ regroupe les $m - r$ états transitoires, et $\mathcal{X}_a$ les r états absorbants.)
Nous pouvons alors écrire la matrice de transition de la Chaîne de la façon suivante

$$P = \begin{bmatrix} Q & 0 \\ R & I_r \end{bmatrix}, \tag{4.9}$$

où Q est une matrice de dimension $(m - r) \times (m - r)$ et I_r est la matrice identité de dimension r. △

Le diagramme de la Figure 4.2 illustre cette partition des états, et le sens des matrices Q et R.

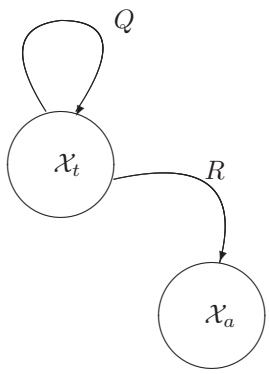


Figure 4.2: Structure d'une Chaîne de Markov absorbante.

Il est aisé de constater qu'avec cette écriture de la matrice de transition, la matrice P^n (voir eq. (4.8)) est de la forme

$$P^n = \begin{bmatrix} Q^n & 0 \\ \star & I_r \end{bmatrix}.$$

où nous n'explicitons pas la matrice de transition entre les états transitoires $\mathcal{X}_t$ et les états absorbants $\mathcal{X}_a$, qui est simplement indiquée par $\star$.

Définition 11 *Distribution stationnaire*
Soit μ une loi de probabilité (éléments positifs dont la somme est unitaire) telle que

$$\mu = P\mu,$$

où P est la matrice de transition d'une Chaîne de Markov X invariante dans le temps. L'équation précédante affirme que μ est un vecteur propre de P avec valeur propre unitaire. Alors, μ est une *loi stationnaire* de la Chaîne X. △

Définition 12 *Chaîne stationnaire*
Si $\forall n \geq 1$ $p_n = \mu$, alors la Chaîne de Markov est un processus stationnaire (sa *distribution* est invariante par rapport à des translations temporelles). △

Définition 13 *Chaîne irréductible (ou ergodique)*
Si $\forall i, j \in \{1, \ldots, m\}$, il existe un k tel que

$$\left[P^k\right]_{ij} > 0,$$

la Chaîne de Markov est *irréductible* (connectée). △

Ceci veut dire que la Chaîne peut transiter, avec probabilité non-nulle, de $X_n = x_j$ vers $X_{n+k} = x_i$, pour n'importe quel pair d'états $(x_i, x_j) \in \mathcal{X}^2$. Nottez que le nombre de pas k peut dépendre du pair d'éléments $(x_i, x_j) \in \mathcal{X}^2$ de départ et d'arrivé considérés. Ceci exclue, bien évidement, la possibilité de l'existance d'états absorvents.

Définition 14 *Chaîne fortement connectée (ou régulière)*
S'il existe un k tel que

$$\left[P^k\right]_{ij} > 0, \qquad \forall i, j \in \{1, \ldots, m\},$$

la Chaîne de Markov est fortement connectée (irréductible et apériodique). △

Dans cette dernière définition, la valeur de k est la même pour tous les pairs $(x_i, x_j) \in \mathcal{X}^2$. Notez que si une chaîne est régulière alors elle est nécessairement ergodique :

$$\text{régularité} \Rightarrow \text{ergodicité}$$

mais l'inverse n'est pas vrai. Toute matrice de transition qui ne contient pas de zéros définie une chaîne régulière (et donc ergodique).

Exemple 2 *Chaîne binaire*
Considérons une Chaîne de Markov invariante dans le temps, avec matrice de transition

$$P = \begin{bmatrix} 1-\alpha & \beta \\ \alpha & 1-\beta \end{bmatrix}.$$

Si $\alpha =$ ou $\beta = 0$, la Chaîne n'est pas connectée (il existe un état qui ne peut pas être atteint à partir de l'autre état).
Pour $\alpha \neq 0$ et $\beta \neq 0$, la Chaîne possède la distribution stationnaire suivante:

$$\mu = \begin{bmatrix} \frac{\beta}{\alpha+\beta} \\ \frac{\alpha}{\alpha+\beta} \end{bmatrix}.$$

Si $\alpha = \beta = 1$, la Chaîne est connectée, mais elle n'est pas fortement connectée (en effect, elle est périodique, de période égale à 2) :

$$P^{2k+1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad P^{2k} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

△

Théorème 2 *Probabilité d'absortion*
Soit X_n une Chaîne de Markov **absorbante** *($r \geq 1$), et P sa matrice de transition, exprimée dans la forme canonique (4.9). Alors, la probabilité que la chaîne soit absorbée est 1, c'est à dire,*

$$\lim_{n\to\infty} Q^n = 0. \tag{4.10}$$

△

Démonstration
Soit m_j le nombre minimal de pas pour que la chaîne puisse passer à un état absorbant $x \in \mathcal{X}_a$ à partir de l'etat j:

$$m_j = \min_{k \geq 1} \ : \ \exists i \in \mathcal{X}_a \ \ : \ P_{ij}^k > 0.$$

Soit p_j la probabilité pour que la chaîne ne soit pas absorbée en m_j pas à partir de l'état j:

$$p_j = 1 - \sum_{i \in \mathcal{X}_a} P_{ij}^{m_j} < 1,$$

où l'inégalité est stricte à cause de la définition de m_j. Soient

$$m = \max_{j \in \mathcal{X}_t} m_j, \qquad p = \max_{j \in \mathcal{X}_t} p_j.$$

La probabilité de ne pas être absorbée en m pas est

$$\Pr\{ \text{ pas absorbée en } m \text{ pas } \} \leq p.$$

De la même manière,

$$\Pr\{ \text{ pas absorbée en } im \text{ pas } \} \leq p^i,$$

et donc

$$\lim_{i \to \infty} \Pr\{ \text{ pas absorbée en } im \text{ pas } \} \to 0.$$

Comme

$$\Pr\{ \text{ pas absorbée en } m+1 \text{ pas } \} \leq \Pr\{ \text{ pas absorbée en } m \text{ pas } \},$$

est une séquence monotonne en n, nous pouvons conclure que

$$\Pr\{ \text{ pas absorbée en } n \text{ pas } \} \to 0,$$

et donc

$$\lim_{n \to \infty} Q^n = 0.$$

△

Théorème 3 *Matrice fondamentale*
Soit X_n une Chaîne de Markov **absorbante**, *et P sa matrice de transition, exprimée dans sa forme canonique (4.9).*
Alors $I - Q$ possède une inverse, $N = (I - Q)^{-1}$ appelée matrice fondamentale, de la forme

$$N = (I - Q)^{-1} = I + Q + Q^2 + \cdots . \tag{4.11}$$

L'entrée N_{ij} de la matrice N est égale à l'espérance statistique du nombre de fois que la chaîne passe par l'état i, sachant qu'elle a été initiée dans l'état j. △

Démonstration
Nous constatons d'abord que $I - Q$ est une matrice inversible, en montrant que son espace nul ne contient que le vectuer zéro :

$$(I - Q)x = 0 \Rightarrow x = 0. \tag{4.12}$$

Soit x une solution de $(I - Q)x = 0$:

$$x = Qx \Rightarrow x = Q^n x.$$

Le Théorème 2 nous permet d'affirmer

$$\lim_{n\to\infty} Q^n x = 0 \Rightarrow x = 0.$$

La seule solution de (4.12) est donc le vecteur nul, et la matrice $I - Q$ possède une inverse, que nous désignons par $N = (I - Q)^{-1}$.

Nous dérivons ensuite l'expression de l'inverse de $I - Q$ donnée dans le théorème 3. Il est facile de vérifier que

$$(I - Q)\left(I + Q + Q^2 + \cdots + Q^n\right) = I - Q^{n+1}.$$

et donc

$$I + Q + Q^2 + \cdots + Q^n = N(I - Q^{n+1}),$$

ce qui implique

$$\lim_{n\to\infty} \sum_{i=0}^{n} Q^i = \lim_{n\to\infty} N(I - Q^{n+1}) = N, \tag{4.13}$$

et nous avons ainsi obtenu l'expression (4.11).

Pour établir l'interprétation des éléments de N comme l'espérance de l'"occupation" des différents états à partir des états transitoires, nous définissons les variables binaires suivantes:

$$\ell_{ij}(k) = \begin{cases} 0, & \text{si } X_k = i, X_0 = j \\ 1, & \text{si } X_k \neq i, X_0 = j \end{cases}$$

$\ell_{ij}(k)$ est donc égal à 1 pour tous les instants où la Chaîne passe par l'état x_i. Le nombre de fois que la Chaîne est passée par x_i jusqu'à l'instant k est

$$N_{ij}(k) = \sum_{p\leq k} \ell_{ij}(p).$$

La distribution statistique des variables $\ell_{ij}(k)$ est déduite de la distribution de l'état de la chaîne à l'instant k:

$$\Pr\{\ell_{ij}(k) = \ell\} = \begin{cases} \Pr\{X_k = i | X_0 = j\}, & \ell = 1 \\ \Pr\{X_k \neq i | X_0 = j\}, & \ell = 0 \end{cases}$$

c'est à dire,

$$\Pr\{\ell_{ij}(k) = \ell\} = \begin{cases} Q_{ij}^k, & \ell = 1 \\ 1 - Q_{ij}^k, & \ell = 0 \end{cases}$$

Comme ils s'agit de variables aléatoires binaires dans $\{0, 1\}$, il résulte imédiatement

$$\mathrm{E}\{\ell_{ij}(k)\} = Q_{ij}^k.$$

L'espérance du nombre de fois que la chaîne passe par l'état x_i jusqu'à l'instant k à partir de l'état x_j à l'instant zéro est donc

$$\mathrm{E}\{N_{ij}(k)\} = \mathrm{E}\left\{\sum_{p\leq k} \ell_{ij}(p)\right\} = \sum_{p\leq k} \mathrm{E}\{\ell_{ij}(p)\} = \sum_{p\leq k} Q_{ij}^p$$

En prennant la limite $k \to \infty$:

$$\lim_{k\to\infty} \mathrm{E}\{N_{ij}(k)\} = \sum_{p=0}^{\infty} Q_{ij}^p = N_{ij}.$$

où nous avons utilisé (4.13). △

Nous venons de présenter quelques résultats concernant le comportement asymptotique de chaînes *absorbantes*. Nous allons maintenant énoncer des théorèmes similaires pour des chaînes *régulières*. Le Lemme suivant sera utilisé.

Lemme 2 *"Contraction"*
Soit P une matrice stochastique de dimension$m \times m$ avec toutes les composantes différentes de zéro (et donc correspondante à une chaîne fortement connectée, ou régulière). Soit d sa plus petite composante:

$$d = \min_{i,j} P_{ij}.$$

Soit y un vecteur (ligne) de dimension m, et $z = yP$. Soient

$$M_y = \max_i y_i, \qquad m_y = \min_i y_i; \qquad M_z = \max_i z_i, \qquad m_z = \min_i z_i.$$

Alors

$$M_z - m_z \leq (1 - 2d)(M_y - m_y) \tag{4.14}$$

Ce Lemme affirme que les éléments de $z = yP$ sont plus "proches" les uns des autres que ceux de y.

Démonstration

Comme P est une matrice stochastique, les éléments du vecteur $z = yP$

$$z_i = \sum_j y_j P_{ji}$$

sont des moyennes des composantes du vecteur y, avec des poids ($\{P_{ji}\}_{j=1}^{m}$) qui sont données par les colonnes de P. Nous allons déterminer des bornes (inférieure et supérieure) pour ces moyennes.

La moyenne *la plus grande*, M_z, est obtenue pour un vecteur y qui a toutes ses composantes égales à la valeur maximale (M_y) et une seule composante (la k-ième) égale à la valeur minimale (m_y), et quand cette dernière composante est multipliée par la plus petite entrée de P. Nous avons donc, dans ces conditions :

$$M_z \leq dm_y + \sum_{i \neq k} M_y P_{ij} = dm_y + M_y \sum_i P_{ij} = dm_y + (1-d)M_y \leq M_y$$

où nous avons utilisé le fait que la somme de tous les éléments d'une colonne sauf le k-ième est égale à $1-d$, et la dernière inégalité découle du fait que $m_y \leq M_y$.

La valeur *la plus petite* possible, m_z, est obtenue dans la situation inverse : tous les les éléments sauf un sont égales à la valeur la plus petite (m_y) et la valeur la plus grande (M_y) est multipliée par d:

$$m_z \geq dM_y + (1-d)m_y \geq m_y$$

où la dernière inégalité est justifiée par le fait que $M_y \geq m_y$ (la moyenne d'un ensemble (m_z) est nécessairement supérieure ou égale à la valeur la plus petite dans l'ensemble). De ces deux inégalités nous pouvons déduire la relation (4.14)

$$M_z - m_z \leq dm_y + (1-d)M_y - dM_y - (1-d)m_y \Rightarrow M_z - m_z \leq (1-2d)(M_y - m_y).$$

△

Nous pouvons maintenant énnoncer un théorème qui concerne la structure algébrique asymptotique de P^n.

Théorème 4 *Forme dyadique de P^n*
Soit P la matrice de transition d'une chaîne régulière *dans l'alphabet $\mathcal{X}$, avec $|\mathcal{X}| = m$. Alors,*

$$\lim_{n \to \infty} P^n = w\mathbf{1}^T,$$

où $w \in [0,1]^m$ est une loi de probabilité :

$$w_i \in [0,1], \qquad \forall i = 1, \ldots, m; \qquad \sum_{i=1}^{m} w_i = 1,$$

et $\mathbf{1}$ est le vecteur de dimension m avec toutes ses composantes égales à 1. △

Démonstration (pour $P_{ij} \neq 0, \forall i, j$)
Nous nottons dans un premier temps que les vecteurs w et $\mathbf{1}$ sont des *vecteur propres* (à droite et à gauche, respectivement) de P avec *valeurs propres unitaires*.

Comme P est une matrice stochastique, la somme de toutes ses colonnes est égale à 1:

$$\mathbf{1}^T P = \mathbf{1},$$

ce qui montre que $\mathbf{1}$ est effectivement un vecteur propre (à gauche) de P avec valeur propre unitaire.
La décomposition de $\lim_{n\to\infty} P^n$ donnée dans le théorème implique

$$P \cdot \lim_{n\to\infty} P^n = Pw\mathbf{1}^T = \lim_{n\to\infty} P^{n+1} = w\mathbf{1}^T \Rightarrow Pw\mathbf{1}^T = w\mathbf{1}^T.$$

si nous multiplions cette équation à droite par $\mathbf{1}$, nous obtenons

$$Pwm = wm \Rightarrow Pw = w,$$

ce qui montre que : *(i)* w est un vecteur propre (à droite) de P ; *(ii)* w est la distribution stationnaire associée à la matrice de transition P.

Nous avons donné un sens plus précis aux vecteurs w et $\mathbf{1}$ qui interviennent dans l'énnoncé de ce théorème. Nous allons maintenant le démontrer.

Soit y un vecteur (ligne) de dimension m, et, comme dans le Lemme 2, pour $z^{(n)} = yP^n$, soient

$$M_z^n = \max_i z_i^{(n)}, \qquad m_z^n = \min_i z_i^{(n)}.$$

Le même argument que nous avons utilisé pour démontrer le Lemme 2 nous permet d'affirmer

$$M_z^1 \geq M_z^2 \geq \cdots \qquad \leq m_z^1 \leq m_z^2 \leq \cdots.$$

Ces séquences monotonnes sont encadrées par les valeurs minimales et maximales de y :

$$M_y \geq M_z^n \geq m_z^n \geq m_y$$

et donc elles possèdent une limite quand $n \to \infty$. Soient

$$m = \lim_{n\to\infty} m_z^n, \qquad M = \lim_{n\to\infty} M_z^n.$$

Nous allons démontrer que $M - m = 0$.
Soit, comme dans le Lemme 2, d la plus petite valeur de P:

$$d = \min_{ij} P_{ij} > 0.$$

Le Lemme 2 affirme que

$$M_z^n - m_n \leq (1 - 2d)(M_z^{n-1} - m_z^{n-1})$$

ce qui implique

$$M_z^n - m_z^n \leq (1 - 2d)^n (M_y - m_y).$$

Pour $m \geq 2$ (pour $m = 1$ le théorème est trivial), nous avons nécessairement $d \leq 1/2$, et donc $1 - 2d \leq 1$, ce qui implique

$$\lim_{n\to\infty} M_n - m_n = 0.$$

Ceci veut dire que les composantes de yP^n tendent toutes vers la même valeur, égale à $m = M$.
Prenons maintenant

$$y = e_i,$$

le vecteur (ligne) de dimension m avec toutes les composantes égales à zéro sauf la i-ème. La limte $\lim_{n\to\infty} yP^n$ est dans ce cas égale à la ligne i de $P^\infty = \lim_{n\to\infty} P^n$. Nous venons donc de démontrer que cette ligne tend vers une valeur constante P_i^∞. Comme ceci doit être vrai pour toutes les lignes de la matrice, nous pouvons conclure que P^n est effectivement une matrice de rang unitaire.
Comme les éléments de $\lim_{n\to\infty} P^n$ sont des probabilités, P_i^∞ est nécessairement positive. Comme $\lim_{n\to\infty} P^n$ est une matrice stochastique, elle possède une valeur propre (à gauche) égale à $\mathbf{1}$, et elle doit donc pouvoir s'écrire dans la forme dyadique

$$\lim_{n\to\infty} P^n = w\mathbf{1}^T$$

ce qui démontre le théorème. △

(Nous venons de démontrer le théorème pour des matrices de transition positives : $d > 0$. Il peut être démontré pour des chaînes régulières: *(i)* Si P est régulière, alors $\exists k : [P^k]_{ij} > 0 \forall (i,j)$. Alors on peut montrer que $M_{nk} - m_{nk} \to_{n\to\infty} 0$. *(ii)* Il peut être démontré que la différence $M_n - m_n$ est non-croissante. *(i)* et *(ii)* impliquent que la séquence entière tend vers 0.)

Ce théorème conduit directement au résultat suivant.

Théorème 5 *Théorème fondamental des Chaînes de Markov*
Soit P la matrice de transition d'une chaîne **régulière***, et μ la distribution asymptotique associée à P par le théorème précédant, de façon que*

$$\lim_{n\to\infty} P^n = \mu\mathbf{1}^T.$$

Alors, indépendement de sa distribution initiale p_0,

$$\lim_{n\to\infty} p_n = \mu.$$

△

Démonstration(1)
Soit p_0 la distribution initiale de de la Chaîne. Sa distribution à l'instant n est

$$p_n = P^n p_0.$$

L'application du Théorème précédant conduit directement au résultat prétendu:

$$\lim_{n\to\infty} p_n = \lim_{n\to\infty} P^n p_0 = \mu \mathbf{1}^T p_0 = \mu.$$

△

Cette démonstration est basée dans la structure de P^n établie dans le Théorème 4. Nous allons maintenant présenter une démonstration alternative.

Démonstration(2)
Soit $X_n \in \mathcal{X}$ une chaîne de Markov avec matrice de transition P et distribution initiale p_0, et $Y_n \in \mathcal{X}$ une autre chaîne, avec la même matrice de transition, mais initialisée avec la distribution stationnaire μ. Formons la chaîne $Z_n \in \mathcal{X} \times \mathcal{X}$:

$$Z_n = \begin{bmatrix} X_n \\ Y_n \end{bmatrix}.$$

Les deux chaînes sont évoluées de manière indépendante, de façon que les éléments de la matrice de transition de Z_n (de dimension $m \times m$, sont des produits des entrées de P. La régularité de P implique donc la régularité de Z_n, et donc la nouvelle chaîne peut atteindre n'importe quel état $z \in \mathcal{X}' = \mathcal{X} \times \mathcal{X}$ dans un nombre fini de pas. Soit $T^\star$ l'instant de premier passage de Z_n sur la "diagonale" de $\mathcal{X}'$, c'est à dire, un état de la forme $(i,i), i \in \{1,\ldots,m\}$. Il peut être démontré que

$$\lim_{n\to\infty} \Pr\{T^\star > n\} = 0.$$

(application de l'inégalité de Chebychev $\Pr\{T^\star > n\} \leq \mathrm{E}[T^\star]/n$, avec le fait que le temps moyen pour aller dans un état, $\mathrm{E}[T^\star]$, est fini).
Pour $n > T^\star$,

$$p(X_n = j | n \geq T^\star) = p(Y_n = j | n \geq T^\star).$$

Comme

$$p(X_n = j) = p(X_n = j | n \geq T^\star)\Pr\{T^\star \geq n\} + p(X_n = j | n < T^\star)\Pr\{n < T^\star\}$$

nous obtenons

$$\begin{aligned} \lim_{n\to\infty} p(X_n = j) &= \lim_{n\to\infty} p(X_n = j | n \geq T^\star)\Pr\{T^\star \geq n\} \\ &= \lim_{n\to\infty} p(Y_n = j | n \geq T^\star)\Pr\{T^\star \geq n\} = \mu_j, \end{aligned}$$

car la chaîne Y_n suit, pour tout n, la distribution stationnaire μ, ce qui complète la démonstration. △

Théorème 6 *Taux d'entropie pour une Chaîne de Markov stationnaire*
Le taux d'entropie d'une Chaîne de Markov *stationnaire* est

$$\overline{H}(X) = -\sum_{i,j} \mu_i P_{ji} \log P_{ji}. \tag{4.15}$$

△

La démonstration fait appel à la stationnarité de la Chaîne, et à la propriété de Markovianité :

$$\begin{aligned}
\overline{H}(X) &= H'(X) = \lim_{n\to\infty} H(X_n|X_1^{n-1}) \\
&= \lim_{n\to\infty} H(X_n|X_{n-1}) \quad \text{(Markov)} \\
&= H(X_n|X_{n-1}) \quad \text{(stationnarité)} \qquad (4.16) \\
&= \sum_i p(X_{n-1}=x_i) \left[-\sum_j p(X_n=x_j|X_{n-1}=x_i) \log p(X_n=x_j|X_{n-1}=x_i) \right].
\end{aligned}$$

où la dernière équaiton découle de la définition d'entropie conditionnelle.
L'équation (4.15) est obtenue en identifiant $p(X_{n-1}=x_i)=\mu_i$ et $p(X_n=x_j|X_{n-1}=x_i)=P_{ji}$.

Ce théorème peut être généralisé à des chaînes régulières:

Théorème 7 *Taux d'entropie pour une Chaîne de Markov irréductible et apériodique*
Le taux d'entropie d'une Chaîne de Markov *invariante dans le temps, irréductible et apériodique* est

$$\overline{H}(X) = -\sum_{i,j} \mu_i P_{ji} \log P_{ji},$$

où μ est la distribution stationnaire de la Chaîne:

$$\mu = P\mu.$$

△

Nottez que ce résultat est valable même dans le cas où la Chaîne n'est pas stationnaire (elle n'est pas initialisée avec la distribution stationnaire).

Dans les conditions plus générales du Théorème 7,

$$\begin{aligned}
\overline{H}(X) = H'(X) &= \lim_{n\to\infty} H(X_n|X_1^{n-1}) \\
&= \lim_{n\to\infty} H(X_n|X_{n-1}) \quad \text{(Markov)} \\
&= \lim_{n\to\infty} \sum_i p(X_{n-1}=x_i) \left[-\sum_j P_{ji} \log P_{ji} \right]. \quad \text{(invariance temporelle)}
\end{aligned}$$

L'application du Théorème 5 complète la démonstration.

Nous présentons ici une inégalité qui est souvent utilisée pour démontrer des inégalités en Théorie de l'Information.

Lemme 3 *(Inégalité du logarithme de la somme).*
Soient $a_i, b_i \geq 0$. Alors

$$\left(\sum_i a_i\right) \log \frac{\sum_i a_i}{\sum_i b_i} \leq \sum_i a_i \log \frac{a_i}{b_i}, \tag{4.17}$$

avec égalité si et seulement si $a_i = b_i, \forall i$. △

Démonstration
Basée sur le fait que $t \log t$ est une fonction convexe, et donc, par l'inégalité de Jensen:

$$\mathrm{E}\left[t \log t\right] \geq \mathrm{E}[t] \log \mathrm{E}[t],$$

avec égalité si et seulement si t est une constante.

Dans notre cas,

$$\sum_i \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} \geq \left(\sum_i \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i}\right) \log \left(\sum_i \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i}\right).$$

△

Dans cette section nous avons caractérisé le comportement asymptotique des Chaînes de Markov régulières, démontrant, en particulier, l'existence d'une distribution stationnaire asymptotique (Théorème 5) , et établissant une expression pour leur taux d'entropie $\overline{H}$ (Théorème 7), qui indique la longueur des codes optimaux. La section suivante présente le l'algorithme de Lempel-Ziv, qui est un exemple de codeur universel, atteignant la longueur de code optimale sans connaissance du modèle probabiliste de la source.

4.3 L'algorithme de Lempel-Ziv

Nous allons maintenant présenter l'algorithme de Lempel-Ziv, qui est un exemple d'un codeur universel (pointwise).

L'algorihme de Lempel-Ziv est basé sur la notion de *parsing*. La séquence d'entré est divisée en phrases, chaque phrase étant la séquence de symboles source la plus petite qui n'a pas encore été trouvée. Par exemple, la séquence $x^n = 1011010100010$ donne origine aux phrases

$$1 \;\; 0 \;\; 11 \;\; 01 \;\; 010 \;\; 00 \;\; 10.$$

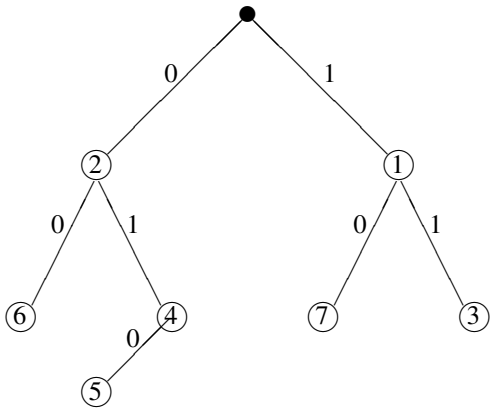
Chaque nouvelle phrase est de la forme $w\,b$, où w est une phrase trouvée précédemment, et b un bit $b \in \{0, 1\}$. Nous pouvons alors la décrire par le pair (i, b), où i est l'index de w (ou *pointer*):

$$w\,b \leftrightarrow (i, b). \tag{4.18}$$

Pour la séquence de l'exemple précédent, nous obtenons

$$(0,1)\ \ (0,0)\ \ (1,1)\ \ (2,1)\ \ (4,0)\ \ (2,0)\ \ (1,0).$$

L'algorithme de Lempel-Ziv construit donc, de manière incrémentale, un *dictionnaire*, formé par toutes les phrases distinctes dans lesquelles la séquence d'origine peut être décomposée. Ce dictionnaire peut être représenté par un arbre, où chaque noeud représente un mot du dictionnaire, et les branches descendantes correspondent aux nouveaux symboles qui sont ajoutés aux mots déjà existants pour former les nouveaux mots. Dans le cas de notre exemple, l'arbre qui représente le dicitonnaire est



Soit $c(x^n)$ le nombre de phrases dans la séquence x^n (le nombre de noeuds dans l'arbre qui représente le dictionnaire). La description de chaque phrase y_i requiert un nombre de bits

$$\ell_i \leq 1 + \lceil \log c(x^n) \rceil < 1 + (1 + \log c(x^n)) = 2 + \log c(x^n) \text{ bits},$$

qui correspondent au nouveau bit b, et à l'utilisation d'un code de longueur constante pour spécifier l'index i (voire eq. (4.18)). Si nous ajoutons un nombre $\log n$ de bits pour coder le nombre de bits avec lequel les indexes sont codés, nous obtenons un nombre total de bits par symbole source

$$\frac{\ell(x^n)}{n} \leq \frac{c(x^n)\,(2 + \log c(x^n)) + \log n}{n},$$

où $\ell(x^n)$ est la taille total du message codé.

Le Lemme suivant établit une borne supérieure pour le nombre de phrases $c(x^n)$.

Lemme 4 *Nombre maximal de phrases*
Le nombre de phrases distinctes dans une séquence de longueur n satisfait

$$c(x^n) \leq \frac{n}{(1-\epsilon_n)\log n}, \tag{4.19}$$

où $\epsilon_n \to 0$ quand $n \to \infty$. △

Ce Lemme affirme que le nombre de phrases croît sous-linéairement:

$$\frac{1}{n}c(x^n) \leq \frac{1}{(1-\epsilon_n)\log n} \longrightarrow_{n\to\infty} 0.$$

Démonstration
Soit n_k la somme des longueurs de toutes les séquences distinctes de longueur $\leq k$:

$$n_k = \sum_{i=1}^{k} i2^i, \tag{4.20}$$

car il y a 2^i séquences distinctes de longueur i. Il est facile de vérifier que

$$n_k = (k-1)2^{k+1} + 2 > (k-1)2^{k+1}, \tag{4.21}$$

en constatant que les deux expressions satisfont l'équation de récurrence suivante, avec la même condition initiale :

$$n_k = n_{k-1} + k\,2^k, n_1 = 2.$$

Le nombre de phrases distinctes de longueur $\leq k$ dans un séquence binaire de longueur n, $c_n(k)$, doit donc vérifier

$$c_n(k) = \sum_{i=1}^{k} 2^k = 2^{k+1} - 2 < 2^{k+1} = \frac{(k-1)2^{k+1}}{k-1} \leq \frac{n_k}{k-1}, \tag{4.22}$$

où nous avons utilisé la borne inférieure (4.21) pour n_k.
Soit $k(n)$ la valeur de k tel que

$$n_{k(n)} \leq n \leq n_{k(n)+1},$$

de façon que l'on peut écrire

$$n = n_{k(n)} + \Delta.$$

Le cas où le nombre de phrases distinctes dans la séquence x^n (de longueur n), $c(x^n)$, est *le plus grand possible* est quand $n_{k(n)}$ bits de x^n contiennent toutes les séquences distinctes de longueur $\leq k(n)$ et les Δ bits restant définissent des nouvelles phrases de longueur $k(n)+1$. Alors

$$c(x^n) \quad \leq \quad c_n(k(n)) + \frac{\Delta}{k(n)+1}$$

$$\begin{aligned} &\leq \frac{n_{k(n)}}{k(n)-1} + \frac{\Delta}{k(n)+1} \\ &\leq \frac{n_{k(n)}+\Delta}{k(n)-1} \\ &= \frac{n}{k(n)-1}, \end{aligned}$$

où nous avons utilisé (4.22).
Comme

$$n \geq n_{k(n)} = (k(n)-1)2^{k(n)+1} + 2 \geq 2^{k(n)} \qquad \Rightarrow k(n) \leq \log n.$$

Et nous avons aussi,

$$n \leq n_{k(n)+1} = k(n)2^{k(n)+2} + 2 \leq (k(n)+2)2^{k(n)+2} \leq (\log n + 2)2^{k(n)+2}.$$

Allors,

$$\begin{aligned} k(n)+2 \geq \log \frac{n}{\log n + 2} \Rightarrow k(n)-1 &\geq \log n - \log(\log n + 2) - 3 \\ &= \left(1 - \frac{\log(\log n + 2) - 3}{\log n}\right) \log n \\ \text{(pour } n \geq 4) \qquad &\geq \left(1 - \frac{\log(2\log n) + 3}{\log n}\right) \log n \\ &= \left(1 - \frac{\log\log n + 4}{\log n}\right) \log n = (1-\epsilon_n)\log n \end{aligned}$$

où

$$\epsilon_n = \min\left(1, \frac{\log\log(n) + 4}{\log n}\right),$$

ce qui, utilisé en (4.23) complète la démonstration du Lemme 4.

Soit $\{X_n\}$ une source stationnaire et ergodique, de fonction de distribution d'ordre n $P(x_1, \ldots, x_n)$. Pour un k fixé soit $P(\cdot|\cdot)$ la distribution de X_j sachant $X_{j-k}^{j-1} = X_{j-k}, \ldots, X_{j-1}$, et soit Q_k l'*approximation de Markov d'ordre* k de P:

$$Q_k(x_{-k+1}, \ldots, x_{-1}, x_0, x_1 \ldots, x_n) = P(x_{-k+1}^0) \prod_{j=1}^{n} P(x_j | x_{j-k}^{j-1}).$$

Par la Loi des grands nombres,

$$\begin{aligned} -\frac{1}{n}\log Q_k(x_1, x_2, \ldots, x_n | x_{-k+1}^0) &= \frac{1}{n}\sum_{j=1}^{n} \log P(x_j | x_{j-k}^{j-1}) \\ &\to -E\left[\log P(x_j | x_{j-k}^{j-1})\right] \\ &= H(X_j | X_{j-k}^{j-1}) \end{aligned} \tag{4.23}$$

Nous allons borner le taux du code de Lempel-Ziv par le taux d'entropie de l'approximation de Markov d'ordre k de la loi de la source, pour toutes valeurs de k. Comme celle-ci converge vers le taux d'entropie de la source ergodique quand $k \to \infty$, nous pouvons ainsi démontrer l'*optimalité de l'algorithme de Lempel-Ziv*.

Admettons que la séquence x^n est divisé en c phrases distinctes $y_1, \ldots, y_c$, et soit ν_i le premier bit de la phrase i (voir diagramme) :

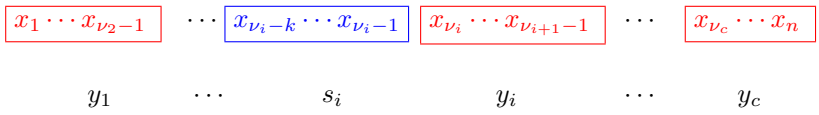
$$y_i = x_{\nu_i}^{\nu_{i+1}-1}.$$

Pour $i = 1, \ldots, c$, soient

$$s_i = x_{\nu_i-k}^{\nu_i-1},$$

l'ensemble de bits qui détermine x_{ν_i} selon le modèle de Markov Q_k: les k bits qui précèdent y_i (voir diagramme). Soit $c_{\ell s}$ le nombre de phrases y_i de taille ℓ qui sont précédées par la séquence s, pour $\ell = 1, 2, \ldots$ et $s \in \mathcal{X}^k$. Alors

$$\sum_{\ell,s} c_{\ell s} = c \text{ (le nombre total de phrases)}, \qquad \sum_{\ell,s} \ell c_{\ell s} = n \text{ (le nombre total de bits)}.$$



Le Lemme suivant donne une borne pour *la probabilité d'une séquence* basée son *découpage* en phrases.

Lemme 5 Pour tout découpage (dans des phrases distinctes) de $x^n = y_i \cdots y_c$,

$$\log Q_k(x_1^n | s_1) \leq -\sum_{\ell,s} c_{\ell s} \log c_{\ell s}.$$

△

Démonstration

$$\begin{aligned}
\log Q_k(x_1^n | x_{-k+1}^0) &= \sum_{i=1}^{c} \log P(y_i | s_i) \\
&= \sum_{\ell s} \sum_{i\,:\,|y_i|=\ell, s_i=s} \log P(y_i | s_i) \\
&= \sum_{\ell s} c_{\ell s} \sum_{i\,:\,|y_i|=\ell, s_i=s} \frac{1}{c_{\ell s}} \log P(y_i | s_i)
\end{aligned}$$

$$\leq \sum_{\ell s} c_{\ell s} \log \left(\sum_{i\,:\,|y_i|=\ell, s_i=s} \frac{1}{c_{\ell s}} P(y_i|s_i) \right)$$

où la dernière inégalité découle de l'inégalité de Jensen et de la concavité du logarithme. Comme les phrases y_i sont distinctes,

$$\sum_{i\,:\,|y_i|=\ell, s_i=s} P(y_i|s_i) \leq 1,$$

ce qui utilisé dans l'équation précédante implique le Lemme 5.

Lemme 6 *Entropie de la loi géométrique*
Soit z une variable aléatoire avec des valeurs dans les entiers positifs, avec moyenne μ. Alors

$$H(Z) \leq (\mu+1)\log(\mu+1) - \mu \log \mu.$$

△

La démonstration de cette inégalité découle du fait que la distribution d'entropie maximale dans les entiers positifs avec une moyenen donnée μ est la distribution géométrique (dont l'entropie est égale au membre droit de cette inégalité). (Vérifiez cette affirmation.)

Lemme 7
Pour toute séquence binaire $x \in \{0,1\}^\infty$,

$$\frac{c(x^n)\log c(x^n)}{n} \leq -\frac{1}{n} \log \max_{P \in \mathcal{P}_k} Q_k(x_1^n | x_{-k+1}^0) + \epsilon_k(n),$$

où $\epsilon_k(n) \to 0$ quand $n \to \infty$ (uniformément en $x \in \{0,1\}^\infty$). △

Démonstration
Pour simplifier la notation, nous utilisons dans cette démonstration c pour désigner $c(x^n)$.
De l'inégalité du Lemme 5

$$\begin{aligned} \log Q_k(x_1^n|s_1) &\leq -\sum_{\ell,s} c_{\ell s} \log \frac{c_{\ell s} c}{c} \\ &= -c\log c - c\sum_{\ell s} \frac{c_{\ell s}}{c} \log \frac{c_{\ell s}}{c} \end{aligned} \tag{4.24}$$

Soient

$$\pi_{\ell,s} = \frac{c_{\ell s}}{c}, \qquad \sum_{\ell,s} \pi_{\ell,s} = 1, \qquad \sum_{\ell,s} \ell \pi_{\ell,s} = \frac{n}{c}.$$

Soient U et V des variables aléatoires telles que

$$\Pr\{U = \ell, V = s\} = \pi_{\ell,s}$$

Alors,

$$E_\pi\{U\} = \frac{n}{c}.$$

et de (4.24)

$$-\frac{1}{n}\log Q_k(x_1^n|s_1) \geq \frac{c}{n}\log c - \frac{c}{n}H(U,V), \tag{4.25}$$

ou encore

$$\frac{c}{n}\log c \leq -\frac{1}{n}\log Q_k(x_1^n|s_1) + \frac{c}{n}H(U,V), \tag{4.26}$$

Par le Lemme 6,

$$\begin{aligned} H(U) &\leq (EU+1)\log(EU+1) + EU\log EU \\ &= \left(\frac{n}{c}+1\right)\log\left(\frac{n}{c}+1\right) - \frac{n}{c}\log\frac{n}{c} \\ &= \log\frac{n}{c} + \left(\frac{n}{c}+1\right)\log\left(\frac{c}{n}+1\right). \end{aligned}$$

Nous avons également, car le nombre de phrases distinctes de longueur k est borné par $\mathcal{X}^k$,

$$H(V) \leq \log|\mathcal{X}|^k = k.$$

Comme l'entropie conjointe $H(U,V)$ est inférieure ou égale à la somme des entropies

$$\begin{aligned} \frac{c}{n}H(U,V) &\leq \frac{c}{n}(H(U)+H(V)) \\ &\leq \frac{c}{n}\log\frac{n}{c} + \left(\frac{c}{n}+1\right)\log\left(\frac{c}{n}+1\right) + \frac{c}{n}k \\ &\leq \epsilon_k(n) \end{aligned} \tag{4.27}$$

où la dernière équation découle de (4.19) avec la définition

$$\begin{aligned} \epsilon_k(n) = &-\frac{1}{(1-\epsilon_n)\log n}\log\frac{1}{(1-\epsilon_n)\log n} \\ &+\left(\frac{1}{(1-\epsilon_n)\log n}+1\right)\log\left(\frac{1}{(1-\epsilon_n)\log n}+1\right) \\ &+\frac{k}{(1-\epsilon_n)\log n}. \end{aligned} \tag{4.28}$$

Nottons que

$$\epsilon_k(n) = O\left(\frac{\log\log n}{\log n}\right) \to 0 \qquad \text{(quand } n \to \infty\text{)}$$

indépendamment de x_1^n et de $P \in \mathcal{P}_k$. L'utilisation de ce résultat en (4.26) implique

$$\frac{c}{n}\log c \leq -\frac{1}{n}\log Q_k(x_1^n|s_1) + \epsilon_k(n),$$

pour tout $P \in \mathcal{P}_k$, et donc, en particulier

$$\frac{c}{n} \log c \leq -\frac{1}{n} \log \max_{P \in \mathcal{P}_k} Q_k(x_1^n | s_1) + \epsilon_k(n),$$

qui est l'énoncé du Lemme.

Théorème 8
Soit $\ell(x^n)$ la taille du code produit par l'algorithme de Lempel-Ziv pour une source stationnaire et ergodique x^n. Alors, pour tout $x^n \in \{0,1\}^n$,

$$\limsup_{n\to\infty} \frac{1}{n} \ell(x^n) \leq \lim_{k\to\infty} \limsup_{n\to\infty} \left[-\frac{1}{n} \log \max_{P \in \mathcal{P}_k} Q_k(x_1^n | s_1) \right]$$

△

Démonstration
Conséquence imédiate du fait que

$$\ell(x^n) \leq c(x^n)(\log c(x^n) + 2),$$

et donc

$$\limsup_{n\to\infty} \frac{1}{n} \ell(x^n) \leq \limsup_{n\to\infty} \left(\frac{c(x^n) \log c(x^n)}{n} + 2\frac{c(x^n)}{n} \right)$$

et que, par le Lemme 4

$$\limsup_{n\to\infty} \frac{c(x^n)}{n} = 0,$$

donc :

$$\limsup_{n\to\infty} \frac{1}{n} \ell(x^n) \leq \limsup_{n\to\infty} \frac{c(x^n) \log c(x^n)}{n}.$$

Par le Lemme 7,

$$\limsup_{n\to\infty} \frac{c(x^n) \log c(x^n)}{n} \leq -\limsup_{n\to\infty} \frac{1}{n} \log \max_{P \in \mathcal{P}_k} Q_k(x_1^n | x_{-k+1}^0)$$

Comme ce résultat est valable pour tout k:

$$\limsup_{n\to\infty} \frac{1}{n} \ell(x^n) \leq \limsup_{n\to\infty} \frac{c(x^n) \log c(x^n)}{n} \leq -\lim_{k\to\infty} \limsup_{n\to\infty} \frac{1}{n} \log \max_{P \in \mathcal{P}_k} Q_k(x_1^n | x_{-k+1}^0)$$

qui est l'énnoncé du théorème.

Ce théorème implique le Corolaire suivant:
Corolaire Optimalité du code de Lempel-Ziv
Soit $X = \{X_i\}$ une source stationnaire ergodique avec taux d'entropie $\overline{H}(X)$. Alors le code de Lempel-Ziv satisfait

$$\limsup_{n\to\infty} \frac{1}{n} \ell(x^n) \leq \overline{H}(X), \text{ avec probabilité 1.}$$

Démonstration
Nous avons vu que pour des sources ergodiques (eq. (4.23))

$$\lim_{n\to\infty} -\frac{1}{n}\log Q_k(x_1^n|x_{-k+1}^{-1}) = H(X_j|X_{j-k}^{j-1})$$

et, pour des sources stationnaires,

$$\lim_{k\to\infty} H(X_j|X_{j-1},\ldots,X_{j-k}) = \overline{H}(X).$$

Nous avons montré que le nombre de bits par symbole source utilisé par le code de Lempel-Ziv ne dépasse pas (assymptotiquement) le taux d'entropie de la source. Le code de Lempel-Ziv est un exemple simple de codeur universel, qui atteint un comportement (asymptotiquement) optimal sans avoir besoin de connaître la distribution statistique de la source.

Références

1. J. Ziv, A. Lempel, *A universal algorithm for sequential data compression*, IEEE Trans. Inf. Th., Vol IT-23, pp 337:343, May 1977,

2. J. Ziv, A. Lempel, *Compression of individual sequences via variable rate coding*, IEEE Trans. Inf. Th., Vil IT-24, pp 530:536, Sept. 1978.

3. A. Lempel, J. Ziv, *On the complexity of finite sequences*, IEEE Trans. Inf. Th., Vol IT-22, pp 75:81, Jan. 1976.