Taux d'entropie

Taux d'entropie (*Entropy rate*)

X_n : source. Son taux d'entropie (*entropy rate*), $\overline{H}(X)$ est, par définition

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

où $X_1^n = \{X_1, X_2, \ldots, X_n\}$.

Pour une source **stationnaire**, le limite $\overline{H}(X)$ existe et est égal à

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

Démonstration

- Le limite H' existe

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

(a) : conditionnement diminue l'entropie
(b) par la stationnarité de X_n.

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

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

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

Moyenne Césaro

Moyenne de Cesáro

Soit a_n une séquence et a sa limite :

$$a_n \longrightarrow a.$$

Soit b_n la séquence :

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

Alors

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

$\overline{H}(X) = H'(X)$

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

(a) la définition de $\overline{H}$
(b) la régle de la chaîne pour l'entropie conjointe
(c) la définition de α_i
L'application de la moyenne de Césaro $\Rightarrow$

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

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

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

et

$$\frac{1}{n} H(X_1^n) \leq \frac{1}{n-1} H(X_1^{n-1}).$$

La séquence $b_n = \frac{1}{n} H(X_1^n)$ est décroissante et bornée inférieurement par $\alpha_n = H(X_n | X_1^{n-1})$.

Démonstration

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

(a) est justifiée par la stationnarité de la séquence
(b) par le fait que le conditionnement diminue l'entropie

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

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

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

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

Code universel

C_n : code sans pertes pour des séquences de n symboles source
$\ell_n(\cdot)$ taille des mots de C_n.
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* Le code de Lempel-Ziv est un exemple bien connu de code universel

Source ergodique

$X_n = \{\ldots, X_{-1}, X_0, X_1, \ldots\}$, $X_n \in \mathcal{X}$: source stationnaire
$T(X)$ l'opérateur de translation (*shift*):

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

$T^k(X)$: 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 $E[f(X)] < \infty$

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

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*)

Code ponctuellement universel (*pointwise universal code*)

$C_n(X)$ est ponctuellement (*pointwise*) universel si sa longueur $\ell_n(\cdot)$ 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*.
Cette notion de codage universel implique l'optimalité (asymptotique) du code *pour toute séquence* de la source

Processus de Markov

X_n est un *processus de Markov* d'ordre k si

$$p(X_{n+1}|X_1^n) = p(X_{n+1}|X_{n-k+1}^n),$$

$p(X_{n+1} = x_{n+1}|X_{n-k+1}^n)$ – distributions conditionnelles de X_{n+1} sachant les valeurs de X_{n-k+1}^n : les probabilités de transition du processus de Markov Expression mathématique de la notion intuitive de "processus sans mémoire"
Formulé autrement :
le passé ($X_i, i < n$) et le futur ($X_i, i > n$) sont statistiquement indépendants sachant le présent (X_n)

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

Processus invariant dans le temps

U

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

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

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*.

Chaîne de Markov

matrices de transition P_n, de dimension $m \times m$

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

P_n est une *matrice stochastique*: la somme des éléments de toutes ses colonnes doit être égale à 1.
Loi de probabilité à l'instant n, décrite par un **vecteur** p_n de dimension m

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

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

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

P_{n+1} : matrice de transition

Chaîne de Markov invariante dans le temps

P_n ne dépendent pas de n:

$$P_n = P, \forall n$$

$\Rightarrow$

$$p_n = P^n p_0, \tag{1}$$

p_0 : probabilité de l'état initial.

États absorbants et transitoires

Un état $x_i \in \mathcal{X}$ est *absorbant* si $P_{ii} = 1$: 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.

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

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.

Forme canonique de la matrice de transition

$r = \#$ états absorbants

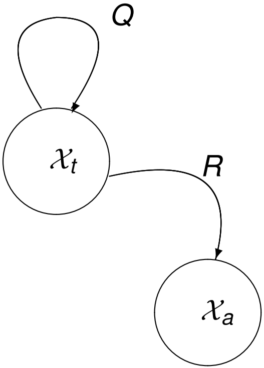
$\{\mathcal{X}_a, \mathcal{X}_t\}$ la partition de $\mathcal{X}$:

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

$\mathcal{X}_t \leftrightarrow m - r$ états transitoires $\qquad \mathcal{X}_a \leftrightarrow r$ états absorbants.

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

$Q\ :\ (m-r) \times (m-r)$ et I_r matrice identité de dimension r.


Q
$\mathcal{X}_t$
R
$\mathcal{X}_a$

Distribution stationnaire

Soit μ une loi de probabilité telle que

$$\mu = P\mu,$$

où P est la matrice de transition d'une Chaîne de Markov X invariante dans le temps. (μ est un vecteur propre de P avec valeur propre unitaire.) Alors, μ est une *loi stationnaire* de la Chaîne X.

Chaîne stationnaire

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

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).

La Chaîne peut transiter de $X_n = x_j$ vers $X_{n+k} = x_i$, pour n'importe quel pair d'états $(x_i, x_j) \in \mathcal{X}^2$.
Le nombre de pas k peut dépendre de $(x_i, x_j) \in \mathcal{X}^2$

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).
La valeur de k est la même pour tous les pairs $(x_i, x_j) \in \mathcal{X}^2$.

Si une chaîne est régulière alors elle est nécessairement ergodique :

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

mais l'inverse n'est pas vrai.

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,

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

Taux d'entropie pour une Chaîne de Markov stationnaire

Le taux d'entropie d'une Chaîne de Markov *stationnaire* est

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

Démonstration

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

Taux d'entropie pour une Chaîne de Markov irréductible et apériodique

Taux d'entropie d'une Chaîne de Markov *invariante dans le temps, irréductible et apériodique*

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

où

$$\mu = P\mu.$$

Démonstration

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

Lempel-Ziv

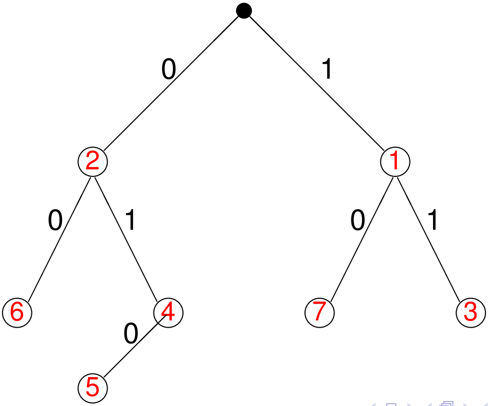
Basé sur la notion de *parsing*.
La séquence est divisée en phrases :
$\rightarrow$ séquence de symboles source la *plus petite* qui n'a *pas encore été trouvée*

Exemple : $x^n = 1011010100010$

1 0 11 01 010 00 10.

Dictionnaire

1 0 11 01 010 00 10.



Chaque phrase y_i : w_i b_i

$w_i = y_j, j < i$ est une phrase trouvée précédemment

$b_i \in \{0, 1\}$: nouveau bit

$w_i\ b_i \leftrightarrow (j, b)$ j est l'index de $w_i = y_j$ (ou *pointer*)

Pour la séquence de l'exemple précédent, nous obtenons

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

Si $c(x^n)$ est le **nombre total de phrases**, alors

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

$\log n$ bits : coder le nombre de bits avec lequel les indexes sont codés

Borne pour la longueur du message en termes du nombre de phrases $c(x^n)$

Nombre total de bits *par symbole source*

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

$\ell(x^n)$: *taille totale* du message codé.

Nous allons voir que pour des sources *stationnaires et ergodiques*

$$\underset{(n\to\infty)}{\frac{\ell(x^n)}{n}} \leq \underset{\substack{\downarrow \\ 0 \\ (A)}}{\frac{2c(x^n)}{n}} + \underset{\substack{\downarrow \\ \overline{H}(X) \\ (B)}}{\frac{c(x^n)\log c(x^n)}{n}} + \underset{\substack{\downarrow \\ 0}}{\frac{\log n}{n}}$$

(A) $\frac{c(x^n)}{n} \to 0$

Le nombre de phrases distinctes dans une séquence de longueur n satisfait

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

où $\epsilon_n \to 0$ quand $n \to \infty$. Le nombre de phrases croît sous-linéairement:

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

$\max c(x^n) \Leftrightarrow \min \ell(y_i)$ $\Leftrightarrow$ toutes les phrases avec $\ell_=$ 1, 2, 3, ..,k

$$n_k = \sum_{j=1}^{k} j2^j = (k-1)2^{k+1} + 2 \qquad (\star)$$

$$\Rightarrow c(n_k) \leq \sum_{j=1}^{k} 2^j \leq 2^{k+1} = \frac{n_k - 2}{k-1} \leq \frac{n_k}{k-1} \rightarrow \frac{c(n)}{n} \leq \frac{1}{k-1}$$

Pour $n_k \leq n < n_{k+1} \Rightarrow n = n_k + \Delta$

$$c(n) \leq c(n_k) + \frac{\Delta}{k+1} \leq \frac{n_k + \Delta}{k-1} = \frac{n}{k-1}$$

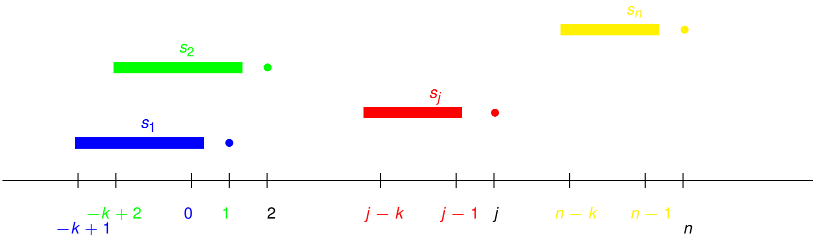
$$n_k \overset{(\star)}{\leq} 2^k \Rightarrow k \leq \log n \qquad (\star\star)$$

$$n \leq n_{k+1} \leq (k+1)2^{k+2} \overset{(\star\star)}{\leq} (\log n + 2)2^{k+1}$$

$$\Rightarrow k + 2 \geq \log \frac{n}{\log n + 2} \Rightarrow k - 1 \geq (1 - \epsilon_n) \log n$$

$\{X_n\}$: source stationnaire et ergodique
Fonction de distribution d'ordre n : $P(x_1, \ldots, x_n)$.
$\forall k$, define

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



Par la Loi des grands nombres,

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

On peut 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, $-\log Q_k/n \longrightarrow_{k\to\infty} \overline{H}$
$\frac{c\log c}{n} \leq -\frac{\log Q_k}{n} \Rightarrow$ optimalité de l'algorithme de Lempel-Ziv

ν_i : *premier bit* de la phrase i

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

$$s_i = x_{\nu_i-k}^{\nu_i-1}, \qquad i = 1, \dots, c \qquad \text{“mémoire” du modèle de Markov}$$

$$c_{\ell s} = \{i \; : \; \ell(y_i) = \ell, \nu_i = s\}$$

$$\sum_{\ell,s} c_{\ell s} = c \quad (\sharp phrases) \qquad \sum_{\ell,s} \ell c_{\ell s} = n \quad (\sharp bits)$$

$$\pi_{\ell s} = \frac{c_{\ell s}}{c} \qquad \text{loi de probabilité}$$

$$\boxed{x_1 \cdots x_{\nu_2-1}} \cdots \boxed{x_{\nu_i-k} \cdots x_{\nu_i-1}} \boxed{x_{\nu_i} \cdots x_{\nu_{i+1}-1}} \cdots \boxed{x_{\nu_c} \cdots x_n}$$

$$y_1 \qquad \cdots \qquad s_i \qquad y_i \qquad \cdots \qquad y_c$$

(B) $\frac{c(x^n)\log c(x^n)}{n} \longrightarrow \overline{H}$

(B.1)

Pour tout découpage de $x^n = y_1 \cdots y_c$,

$$-\frac{1}{n}\log Q_k(x_1^n|s_1) \geq \frac{1}{n}\sum_{\ell,s} c_{\ell s}\log c_{\ell s}$$

(B.2)

$$\frac{1}{n}\sum_{\ell,s} c_{\ell s}\log c_{\ell s} \geq \frac{c\log c}{n} - \epsilon_k$$

$$\epsilon_k \xrightarrow{n\to\infty} 0$$

(B.1)

$$
\begin{aligned}
-\frac{1}{n}\log Q_k(x_1^n|s_1) &= -\frac{1}{n}\log\prod_{i=1}^{c} P(y_i|s_i) = -\frac{1}{n}\sum_{i=1}^{c}\log P(y_i|s_i) \\
&= -\frac{1}{n}\sum_{\ell s}\sum_{i\,:\,\ell_i=\ell,s_i=s}\log P(y_i|s_i) \\
&= -\frac{1}{n}\sum_{\ell s} c_{\ell s}\sum_{i\,:\,\ell_i=\ell,s_i=s}\frac{1}{c_{\ell s}}\log P(y_i|s_i) \\
\text{(Jensen)} \quad &\geq -\frac{1}{n}\sum_{\ell s} c_{\ell s}\log\sum_{i\,:\,\ell_i=\ell,s_i=s}\frac{1}{c_{\ell s}}P(y_i|s_i) \\
&= -\frac{1}{n}\sum_{\ell s} c_{\ell s}\log\frac{1}{c_{\ell s}}\sum_{i\quad \ell_i=\ell,s_i=s} P(y_i|s) \\
(y_i \neq y_j) \quad &\geq \frac{1}{n}\sum_{\ell s} c_{\ell s}\log c_{\ell s}
\end{aligned}
$$

(B.2)

$$
\begin{aligned}
\frac{1}{n}\sum_{\ell s} c_{\ell s}\log c_{\ell s} &= \frac{c\log c}{n} - \frac{c}{n}H(\pi_{\ell s}) \\
&\geq \frac{c\log c}{n} - \frac{c}{n}\left(H(\pi_{\ell\cdot}) + H(\pi_{\cdot s})\right) \\
&\geq \frac{c\log c}{n} - \frac{c}{n}\left(H(\pi_{\ell\cdot}) + k\right) \\
&\geq \frac{c\log c}{n} - \frac{ck}{n} - \frac{c}{n}\left[(\bar{\ell}+1)\log(\bar{\ell}+1) - \bar{\ell}\log\bar{\ell}\right] \\
&= \frac{c\log c}{n} \quad \underbrace{-\frac{ck}{n}}_{\downarrow\, 0} \quad \underbrace{-\frac{c}{n}\log\frac{c}{n}}_{\downarrow\, 0} \quad \underbrace{-\left(1+\frac{c}{n}\right)\log\left(1+\frac{c}{n}\right)}_{\downarrow\, 0}
\end{aligned}
$$

(B.1)+(B.2)

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

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

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

Optimalité de LZ

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 $x^n \in \{0,1\}^n$,

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n}\ell(x^n) &\leq \lim_{k\to\infty} \limsup_{n\to\infty} \left[-\frac{1}{n} \log \max_{P \in \mathcal{P}_k} Q_k(x_1^n | s_1) \right] \\
&= \lim_{k\to\infty} H(X_j | X_{j-k}^{j-1}) \\
&= \overline{H}(X)
\end{aligned}$$