Théorie de l'information

D. Kohel

11/01/2021

L'entropie

Théorème Soit X un espace de proba fini. Alors $H(X) \leq \log_2 |X|$ avec égalité ssi X est uniforme.

Rappel. On dit que X est uniforme si $p(x) = 1/|X|$ pour tout $x \in X$.

Preuve. Par conséquence de l'inégalité de Gibbs.

Lemme (Gibbs). Soit (X, p) un espace de probabilité discrète, et soit $q : X \longrightarrow [0,1]$ une autre probabilité. Alors

1/ $\sum_{x \in X} p(x) \log(p(x)^{-1}) \leq \sum p(x) \log(q(x)^{-1})$,

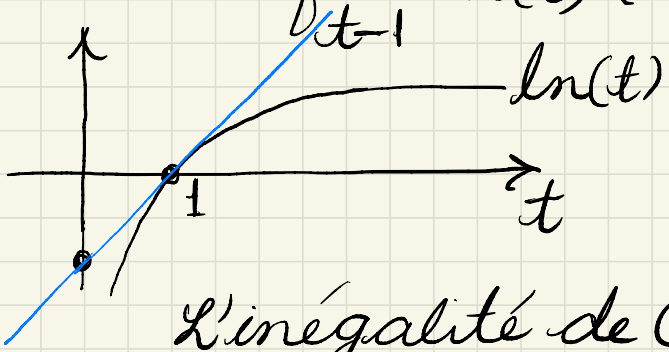
2/ avec égalité ssi $p = q$.

Remarque (Multipliant par $1/\log(2)$.) L'inegalité de Gibbs est

$$H(X) \leq E(\log_2(q(x)^{-1})).$$

Preuve

Il suffit de prendre log = ln, et observer que $\ln(t) \leq t-1$:



L'inégalité de Gibbs est équivalente à

$$\sum_{x \in X} p(x) \ln\left(\frac{q(x)}{p(x)}\right) \leq 0. \quad \circledast$$

Mais on a

$$\sum_{x \in X} p(x) \ln\left(\frac{q(x)}{p(x)}\right) \leq \sum_{x \in X} p(x)\left(\frac{q(x)}{p(x)} - 1\right)$$

$$= \sum_{x \in X} \left(q(x) - p(x)\right) = 1 - 1 = 0.$$

On a démontré $\circledast$, équivalente à l'inégalité de Gibbs.

On observe que $\ln(t) = t-1$,

ssi $t=1$. Par conséquent, l'inégalité $\circledast$ est stricte s'il existe x tel que $p(x) \neq q(x)$. Cela établit la propriété 2/. □

Preuve du théorème. Il suffit de mettre $q(x) = 1/|X|$. Alors

$$H(X) \leqslant E(\log_2 |X|) = \log_2 |X|. \quad \square$$

Théorie de Shannon

Premier théorème de Shannon

Définition. Soit $C : X \longrightarrow A^*$ un codage (où A est un alphabet fini). On dit que i) C est sans pertes si C est injectif, et ii) C est uniquement décodable si l'extension $C^* : X^* \longrightarrow A^*$, définie par $C^*(x_1 \cdots x_n) = C(x_1) \cdots C(x_n)$, est sans pertes.

Exemple. $X=\{A,B,C\}$, $A=\{0,1\}$.
Le codage $C: X \longrightarrow A^*$ donné par: $C(A)=0$, $C(B)=1$, $C(C)=01$.
Alors C est sans pertes, mais pas uniquement décodable, car $C^*(AB)=01=C^*(C)$.
Par contre, $C': X \longrightarrow A^*$ donné par: $C(A)=00$, $C(B)=11$, $C(C)=01$, est uniquement décodable.

Théorème de Shannon (codage source)
Soit $C: X \longrightarrow A^*$ un codage uniquement décodable, et $l_C: X \longrightarrow \mathbb{N}$ la fonction longueur (sur A^*) composée avec C. ($l_C = l \circ C$)
Alors
i) $\frac{H(X)}{\log_2 |A|} \leqslant E(l_C)$, et

ii) il existe un codage C pour lequel $E(\ell_C) \leq \frac{H(X)}{\log_2 |A|} + 1$.

Définition. Un codage est instantané ou préfixe si aucun mot de code est un préfixe d'un autre.

Par exemple: Le code $(= C(X))$

$\{C(A), C(B), C(C)\} = \{0, 1, 01\}$

n'est pas préfixe. Par contre

$\{C(A), C(B), C(C)\} = \{00, 11, 01\}$

est préfixe.

Lemme Un codage préfixe est uniquement décodable.

Preuve Clair (exercice).

Soit $c_1 c_2 \cdots c_n = C(x_1 \cdots x_n)$ —

on peut identifier les mots de code c_1, puis c_2, etc en lisant de gauche à droite. □

Remarque. La définition de codage instantané / préfixe est asymétrique : On peut par symétrie définir un codage suffixe (aucun mot est un suffixe d'un autre). La propriété d'être u.d. reste valable. Mais ...

On peut comparer :

x	$C_1(x)$	$C_2(x)$
A	0	0
B	11	11
C	01	10

Alors C_1 est suffixe et C_2 est préfixe. Les deux sont u.d.

Or si $\overbrace{A\ C}\ \overbrace{B}\ \cdots$

$$C_1^*(x_1 \cdots x_n) = \underbrace{00}\ \underbrace{11}\ \underbrace{\cdots\ 1}$$
$$A\ A\ B\ \ B \cdots$$

il faut lire jusqu'à la fin afin de décider si le début a commencé par $ACB\cdots$ ou par AAB. La difference sera déterminé par la longueur de $C_1^*(x_1 \cdots x_n)$.

Par contre

$$C_2^*(x_1 \cdots x_n) = \overbrace{00}\ \overbrace{11}\ \cdots\ 1,$$
$$A\ A\ B\ B$$

étant préfixe, on peut décoder (de gauche à droite) sans attendre la fin de la transmission du mot de code.

Remarque. Un codage $C: X \to A^n \subseteq A^*$ (de longueur fixée) est préfixe, ssi il est sans perte.

Codage de Huffman

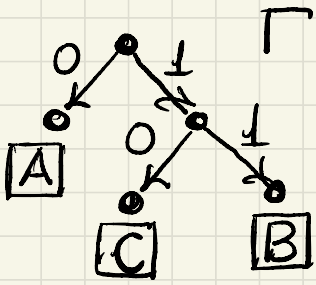
permet de réaliser un codage préfixe qui satisfait la condition ii) du théorème de Shannon. C'est un algorithme de compression. Remarque : la propriété d'être préfixe (et également u.d.) est une propriété du code $C(X)$ (en supposant que C soit sans pertes) et pas de l'espace de probabilité.

Arbre de codage fini

Un arbre de codage est un arbre (graphe sans cycle avec une racine = sommet distingué) dérigé tel que

chaque feuille (sommet terminal) porte une étiquette $x \in X$ et chaque arc d'un sommet porte une étiquette distincte dans A.

Ex. Arbre binaire ($A = \{0, 1\}$)



Si chaque message $x \in X$ apparait une et une seule fois comme étiquette d'une feuille, on dit que Γ est un arbre de codage pour X.

Lemme Il y a un bijection entre les codages préfixes pour X et les arbres de codage pour X.

Construction d'un abre de Huffman

L'algorithme "glouton".

Soit X un ensemble et A un alphabet de cardinal q.

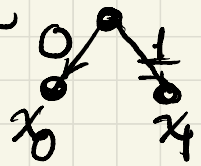
0) Si $|X| \bmod (q-1) \neq 1$, on ajoute des éléments à X avec $p(x)=0$.

i) Mettre $X=\{x_0, x_1, \dots, x_{n-1}\}$ en ordre avec $p(x_0) \leq p(x_1) \leq \dots \leq p(x_{n-1})$.

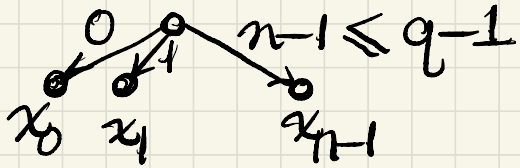
ii) Si $n=|X|=1$, retourne: 

Remarque: On va écrire $A=\{0,1,\dots,q-1\}$ pour les éléments de l'alphabet.

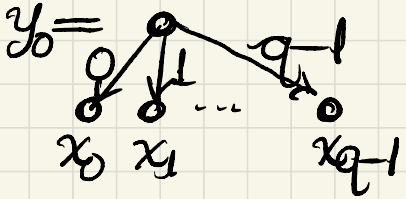
Si $n=2$, retourner



et si ($q>2$ et)
$n \leqslant q$ on met :



Sinon, mettre (construction d'un nouveau espace de proba) :



$y_1 = x_q$, $y_2 = x_{q+1}$, ..., $y_{n-q-1} = x_{n-1}$,
ayant des probabilités :

$$p(y_0) = \sum_{0 \leqslant i < q} p(x_i) \quad \text{et} \quad p(y_j) = p(x_{j+q-1})$$

pour tout $j \geqslant 1$.

On retourne l'arbre de Huffaman pour $\{y_0, \ldots, y_{n-q-1}\}$.

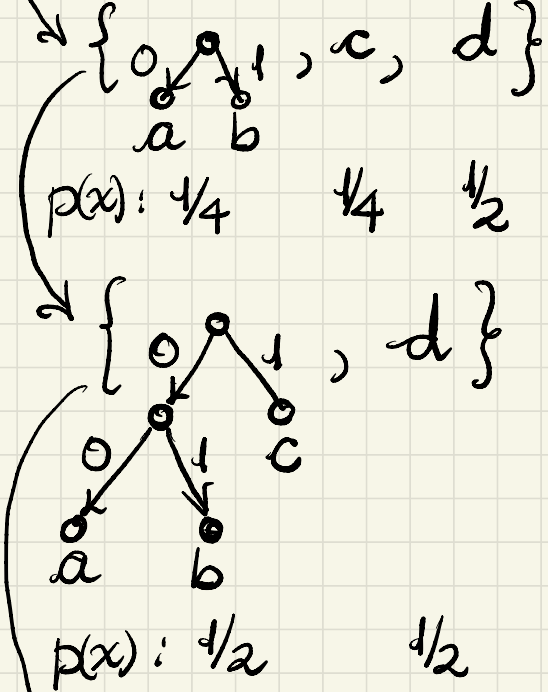
N.B. L'étiquette pour y_0 devient un sous-arbre.

Exemple. $A=\{0,1\}$.

$X=\{a,b,c,d\}$.

$p(x)$: 1/8, 1/8, 1/4, 1/2

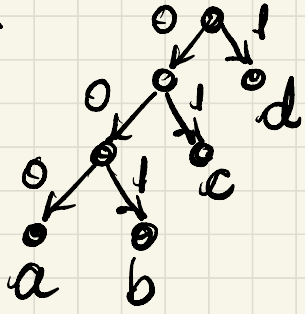
Un arbre de codage est donc :



Remarque

Lors de la construction on peut échanger $\{a,b\}$, ou $\{\{a,b\},c\}$, ou $\{\{\{a,b\}c\},d\}$.

L'arbre de codage de Huffman :



Le codage préfixe de Huffman est donc :

$a \longmapsto 000$ $\quad c \longmapsto 01$

$b \longmapsto 001$ $\quad d \longmapsto 1$

Conclusion. Un codage de Huffman n'est pas unique, mais on peut démontrer que l'espérance mathématique de longueur $E(\ell_C)$ est une invariant de tout codage de Huffman.

Prochain résultat :

Le codage de Huffman satisfait

$$E(\ell_C) \leqslant \frac{H(X)}{\log_2(q)} + 1.$$