Arbre D-aire de hauteur L pondéré

alphabet de codage $\mathcal{D} = \{a_1, a_2, \cdots, a_D\}$

- construction d'un arbre T_L représentant $\mathcal{D}^L$ ($L > 1$)
  - la racine correspond au mot vide
  - les D fils d'un noeud θ sont θa_1, θa_2 , ..., θa_D
  - la construction s'arrête au mots de longueur L
- pondération des noeuds de l'arbre T_L
  la longueur l d'un mot m est égale à la hauteur du noeud le représentant dans l'arbre
  On définit le poids de m par

$$w_T(m) = \frac{1}{D^l}$$

  - le poids de la racine $w(\varepsilon) = 1$
  - le poids d'un noeud interne m est équidistribué entre ses enfants

arbre d'un code "préfixe"

Si L est a longueur maximale du code, les N mots de codes apparaissent dans l'arbre T_L
Construction de l'arbre $T_{\mathcal{C}}$ associé au code préfixé $\mathcal{C}$:

- on marque les mots de code
- pour un code "préfixe" aucun mot marqué n'est l'ancêtre ou le descendant d'un mot marqué
- on élimine de l'arbre tous les noeuds qui ne sont ni des noeuds marqués, ni des ancêtres des mots marqués (les noeuds marqués sont les feuilles de l'arbre)

pondération d'un arbre d'un code "préfixe "

Soit m un mot de l'arbre du code "préfixe" $\mathcal{C}$,

- si m est une feuille de hauteur l alors $w(m) = \frac{1}{D^l}$
- si m est un noeud interne, son poids $w(m)$ est la somme des poids de ses fils
- le poids de l'arbre $w(T_\mathcal{C})$ est défini comme le poids de la racine

Remarque :

- $w(T_\mathcal{C}) = \sum_{i=1}^{N} \frac{1}{D^{l_i}}$
- Si $\mathcal{C} \subset \mathcal{C}'$ alors $w(T_\mathcal{C}) \leq w(T_{\mathcal{C}'})$
- T_L est l'arbre d'un code "préfixe" (D^L états de la sources), $w(T_L) = 1$

Inégalité de Ktraft

Si $\mathcal{D}$ est le cardinal de l'alphabet de codage et $\mathcal{C}$ est un code instantanné composé de mots de longueur $l_1, \ldots, l_N$ alors il satisafait l'inégalité de Kraft :

$$\sum_{i=1}^{i=N} \frac{1}{D^{l_i}} \leq 1$$

Inégalité de Kraft : Réciproquement

Peut-on construire un code "préfixe" des N états de la source par des mots de longueur l_i vérifiant $\sum_{i=1}^{i=N} \frac{1}{D^{l_i}} \leq 1$?
On choisit les mots un par un en s'assurant que le mot construit n'est pas le suffixe d'un mot déjà choisi

- on choisit un mot de longueur l_1 : $\frac{1}{D^{l_1}} < 1$ puisque $\frac{1}{D^{l_1}} \leq \frac{1}{D}$
- par récurrence : soit $l_1 \leq l_2 \leq \cdots \leq l_k$ tel que $\sum_{i=1}^{k} \frac{1}{D^{l_i}} < 1$ on peut choisir $l_{k+1} \geq l_k$ tel que $\sum_{i=1}^{k} \frac{1}{D^{l_i}} \leq 1 - \frac{1}{D^{l_{k+1}}}$, autrement dit tel que $\sum_{i=1}^{k+1} \frac{1}{D^{l_i}} \leq 1$

Montrons alors qu'on peut choisir un mot de code de longueur l_{k+1} qui n'a aucun mot de code préalablement choisi comme préfixe

Inégalité de Kraft : Réciproquement

- il ya $D^{l_{k+1}}$ mots de longueur l_{k+1}
- un mot de code de longueur l_i a $D^{l_{k+1}-l_i}$suffixes de longueur l_{k+1}
- le nombre de mots de longueur l_{k+1} qui ont un mot de code pour préfixe est au plus de

$$\sum_{i=1}^{k} D^{l_{k+1}-l_i} = D^{l_{k+1}} \sum_{i=1}^{k} \frac{1}{D^{l_i}} \leq D^{l_{k+1}} - 1$$

Inégalité de Kraft

Si D est le cardinal de l'alphabet utilisé pour le codage, il existe un code instantanné composé de mots de longueur $l_1, \ldots, l_N$ ssi

$$\sum_{i=1}^{i=N} \frac{1}{D^{l_i}} \leq 1$$

Théorème de Mac Millan

tout code déchiffrable satisfait la condition de Kraft

Théorème de Mac Millan : Démonstration

Si $\mathcal{C}$ est un code déchiffrable de longueur maximale L et de poids $w(\mathcal{C}) = w(t_{\mathcal{C}})$.

$\mathcal{C}.\mathcal{C} = \mathcal{C}^2$ est aussi déchiffrable et $w\left(\mathcal{C}^2\right) = (w(\mathcal{C}))^2$.

Par récurrence pour $r \in \mathbb{N}^*$,
$\mathcal{C}^r$ est aussi déchiffrable et $w(\mathcal{C}^r) = (w(\mathcal{C}))^r$

Théorème de Mac Millan : Démonstration

Supposons $w(\mathcal{C}) > 1$ et prenons r tel que $(w(\mathcal{C}))^r > rL$

- Tous les mots de $\mathcal{C}^r$ sont de longueur au plus rL
- si pour $l = 1, .., rL$ on note $w_l(\mathcal{C}^r)$ la somme des poids des mots de longueur l de $\mathcal{C}^r$ on a

$$w(\mathcal{C}^r) = \sum_{i=1}^{rL} w_l(\mathcal{C}^r)$$

- on a donc $\sum_{i=1}^{rL} w_l(\mathcal{C}^r) > rL$
- il existe donc une longueur $l \leq rL$ tel que $w_l(\mathcal{C}^r) > 1$
- tous les mots de longueur l étant de poids $\frac{1}{D^l}$ cela impose qu'il y ait plus de D^l mots de code de longueur l : **contradiction**

Kraft - Mac Millan : conclusion

S'il existe un code déchiffrable avec des mots de longueur l_i, ce code vérifie la condition de Kraft et donc il existe un code instantané pour le même ensemble de longueur

On peut donc se contenter d'utiliser des codes instantanés sans être pénalisé du point de vue des performances de compression

Efficacité d'un code

notation : entropie de base D de la v.a X :

$$H_D(X) = -\sum_{i=1}^{i=n} p_i log_D(p_i) = \frac{H(X)}{log_2(D)}$$

- l'entropie moyenne par caractère est $\frac{H(X)}{\nu}$
- majorée par $log_2(D)$ entropie de la loi uniforme sur $\mathcal{D}$

Efficacité du code :

$$Eff = \frac{H(X)}{\nu log_2(D)}$$

$$Eff = \frac{H_D(X)}{\nu}$$

Premier Théorème de Shannon

Si X est une source simple sans mémoire , pour tout code déchiffrable d'alphabet de cardinal D et de compacité ν

$$H_D(X) \leq \nu$$

Il existe un code déchiffrable pour lequel

$$H_D(X) \leq \nu \leq H_D(X) + 1$$

La borne de la compacité peut être atteinte asymptotiquement

$H_D(X)$ borne de la compacité

- la borne est atteinte ssi pour tout i on a

$$p_i = \frac{1}{D^{l_i}}$$

- on peut construire un code tel que

$$H_D(X) \leq \nu \leq H_D(X) + 1$$

  - pour tout $x_i \in \mathcal{A}_X$ il existe un unique l_i tel que $D^{-l_i} \leq p(x_i) \leq D^{-l_i+1}$
  - $\sum_1^N D^{-l_i} \leq 1$ donc d'après le théorème de Kraft il existe un code préfixe dont les mots ont pour longueurs $l_1, \cdots, l_N$
  - pour tout $x_i \in \mathcal{A}_X$: $-log_D(p(x_i)) \leq l_i \leq -log_D(p(x_i)) + 1$
  - $H_D(X) \leq \nu \leq H_D(X) + 1$

L'éfficacité du code préfixe produit vérifie

$$1 - \frac{1}{H_D(X) + 1} < E < 1$$

codage par bloc

- on considère la source $Y_i = (X_1, \cdots, X_i)$ discrète et sans mémoire .
- il existe un code déchiffrable tel que $\mathcal{C}_i$ de compacité ν_i tel que
$$H_D(Y_i) \leq \nu_i \leq H_D(Y_i) + 1$$
- X étant une source simple $H_D(Y_i) = iH_D(X)$
- $H_D(X) \leq \frac{\nu_i}{i} \leq H_D(X) + \frac{1}{i}$
- Le codage par bloc permet d'approcher la borne mais entraine un retard et augmente la complexité du codeur

codage par bloc : exemple

- la source X délivre deux symboles A et B avec $p(A) = 0.8$ et $p(B) = 0.2$, alphabet de codage $\mathcal{D} = \{0, 1\}$
  - $H(X) = 0.72 bits$ et $\nu = 1$
  - $0.72 \leq \nu \leq 1.72$
  - un surcoût de 0.28 oit de 39% sur la longueur moyenne des messages codés
- la source $Y = X_1 X_2$ délivre 4 symboles

y	p(y)	$-log_2(p(y))$	l	code
AA	0.64	0.69	1	0
AB	0.16	2.64	2	10
BA	0.16	2.64	3	100
BB	0.04	4.64	3	101

  - Codage des extensions de la source d'ordre 2 : longueur moyenne de 1.56 par paire d'états de la source
  - Longueur moyenne de $\nu = 0.76$ par état de la source $H(X) \leq \nu = .78 \leq H(X) + \frac{1}{2}$ soit un surcout de 0.06, 8% sur la longueur moyenne des mots

Premier théorème de Shannon

Pour toute source discrète sans mémoire, pour tout réel $\varepsilon > 0$ il existe un codage déchiffrable dont l'efficacité est strictement supérieure à $1 - \varepsilon$