Huffman coding

Figure 1: Huffman coding

fiche de cours

Hector Blondel

1 Entropy and other quantities

$H(Y\|X) = \mathbb{E}_{X;Y}(lg(\frac{1}{\mathbb{P}(Y\| X)})) = \sum_i \mathbb{P}(X = x_i)H(Y\| (X = x_i))$ $I(X;Y) = H(X) - H(X\| Y) = H(Y) - H(Y\| X)$

2 Source coding

2.1 Coding methods

Instantaneous codes $\subset$ Decodable codes $\subset$ Regular codes $\subset$ Block codes $\subset$ Codes

- Un code est régulier (non singulier) si deux mots codes correspondant à deux messages distincts sont distincts
- Un code régulier est déchiffrable si à toute **suite** de mots codes ne correspond qu'un seul message de la source
- An instantaneous code (*code instantané*) is a code such that a codeword cannot be the beginng of another codeword.

Kraft inequality : There exists a code $\mathcal{C}$ for a length r codeword alphabet whose codewords are of length $l_1, ..., l_N$ if and only if :

$$\sum_{n=1}^{N} (\frac{1}{r})^{l_i} \leq 1$$

Let S be a source and let $L = \sum_{i=1}^{N} p_i l_i$ be the mean code length. We have :

$$L \geq \frac{H(S)}{\log r} = H_r(S)$$

Efficiency (*rendement*) is $\eta = \frac{H_r(S)}{L}$.

Shannon's source coding theorem : We can build a code such that :

$$\frac{H(S)}{\log r} \leq L \leq \frac{H(S)}{\log r} + 1$$

2.2 Huffman coding :

3 Channel coding

3.1 Channel caracterization

Channel capacity : $C := \max_{P(X)} I(X,Y)$ $C_{awgn} = \log_2(1+snr)$ bits per symbol, or bits/s/Hz $C = W\log_2(1+snr) = W\log_2(1+\frac{P}{N_0W})$ bits/s

- Low SNR : BAndwidth increases a lot the capacity, but power not so much
- High SNR : bandwidth doesn't increase so much the capacity, power increase it a lot Most of the time, we are in Medium to high SNR regime.

The capacity of the continuous WAGN channel ($Y = X+N, N \sim \mathcal{N}(0,\sigma^2)$) is $C = \max_{\mathbb{P}(X)} I(X;Y) = \frac{1}{2}\log(1+\frac{P_X}{P_N})$ It is achieved for a gaussian X following $\mathcal{N}(\mu,\sigma^2)$, $\mathbb{H}(X) = \frac{1}{2}\lg(2\pi\sigma^2)$

Symmetric binary channel : $C = \mathcal{H}_2(\frac{1}{2}) - \mathcal{H}_2(p) (= \mathcal{H}_2(\frac{1}{2}) - \mathcal{H}_2(1-p)) = 1 - \mathcal{H}_2(p)$ where $\mathcal{H}_2(p) = p\lg\frac{1}{p} + (1-p)\lg\frac{1}{1-p}$

3.2 Block error correcting codes

$C : (\mathbb{F}_2)^k \to (\mathbb{F}_2)^n$ code. $C(\mathbb{F}_2^k) \subset \mathbb{F}_2^n$ is the set of codewords The code rate is $\frac{k}{n} \leq 1$ Minimum distance is : $d = \min_{c_1,c_2\in\mathcal{C}} d(c_1,c_2)$, so detection capacity is $d-1$ and correction capacity is $t := \lfloor\frac{d-1}{2}\rfloor$

> /! d is noted D in the course, with d = D-1

3.3 Linear codes

Linear codes are described by equations :

$$\begin{cases} v = sG \\ \Leftrightarrow vH^T = 0 \end{cases}$$

G is the **generator matrix** and H is the **syndrome matrix**.

Systematic code : $G = [I_k\ G'] \Rightarrow H = [-A^T\ I_{n-k}]$

Décodage : La donnée d'une table, dite tableau standard associant à chaque syndrome son unique antécédent par H dans la boule de centre le vecteur nul et de rayon t, permet le décodage. La correction d'erreurs consiste à soustraire l'antécédent e à l'élément de F reçu. Le mot du code recherché est c = x - e. L'implémentation informatique de cette méthode utilise une table de hachage et constitue une méthode rapide de décodage.

Singleton bound :$M \leq q^{N-d+1}$; pour un code linéaire : $d-1 < n-k$. Codes MDS satisfont cette borne Hamming bound : $\sum_{i=0}^{t} C_n^i \leq 2^{n-k}$. Codes parfaits (comme code de Hamming) satisfont cette borne. Codes de Hamming permettent de construire des codes très puissants.

Extension et raccourcissement de codes :

- Si le code $C(n,k)$ est de distance D impaire alors le code étendu $C(n+1,k)$ est de distance $D+1$
- si le code C(n,k) est de distance D alors le code raccourci C(n-s, k-s) est de distance supérieure ou égale à D

3.4 Some code examples

- Hamming codes : systematic, $n = 2^{n-k}-1$. There have **maximum rate and distance 3**. Hamming codes are perfect codes : Un code parfait correcteur de t bits erronés sur n est tel que tout mot de n bits est à une distance inférieure ou égale à t d'un mot du code.

4 Lossy coding

memoryless Gaussian source : $R(D) = \frac{1}{2}\log_2(\frac{\sigma_x^2}{D})$ and $D(R) = 2^{-2R}\sigma_x^2$

For an arbitrary source and under high rate assumption, $D(R) = \varepsilon_X\sigma_X^2 2^{-2R}$

Optimal allocation for N sources after transform :

$$D_{t,k} = \varepsilon_{t,k}\sigma_{t,k}^2 2^{-2R_k} = \left(\prod_{k=1}^{N} \varepsilon_{t,k}\sigma_{t,k}^2\right)^{1/N} 2^{-2R}$$