ACCQ 202 - Information theory

1 Source coding

Def. **Source** d'information : v.a. $X \in \mathcal{X}, |\mathcal{X}| < \infty$ telle que $X \sim P$ avec probabilités $\forall i \in [\![1\,;|\mathcal{X}|]\!], p_i = P(X = i)$.
Def. **Code** pour une source X : $\mathcal{C} \colon \mathcal{X} \to \{0,1\}^*$.
Def. **Longueur moyenne** d'un code $\mathcal{C}$: $\mathcal{L}(\mathcal{C}) = \sum_i p_i l_i$ avec l_i la longueur du i^e^ mot codé.
Def. Un code est **non singulier** si $\forall x_i \neq x_j, \mathcal{C}(x_i) \neq \mathcal{C}(x_j)$.
Def. L'extension d'un code $\mathcal{C}$ est $\forall n, \forall x_1, , \dots, x_n, \mathcal{C}(x_1, \dots, x_n) \triangleq \mathcal{C}(x_1) * \mathcal{C}(x_2) \cdots * \mathcal{C}(x_n)$.
Def. Un code est **à décodage unique** si son extension est non singulière.
Def. Un code est dit **instantané** si aucun mot code n'est le préfixe d'un autre. On dit alors qu'il s'auto-ponctue car on peut décoder en temps réel, symbole par symbole.
Th (Inégalité de Kraft). *Soit $\mathcal{C}$ un code instantané avec longueurs (l_i). Alors $\sum_i l_i \leqslant 1$. Inversement, soit (l_i) une famille de longueurs. Si elle satisfait l'inégalité de Kraft alors il existe un code à décodage unique avec ces longueurs.*
Th (de **McMillan**). *Le théorème précédent reste valable si l'on remplace décodage instantané par décodage unique.*
Cor. $\min_{\mathcal{C} \text{ à décodage unique}} \mathcal{L}(\mathcal{C}) = \min_{\mathcal{C} \text{ à décodage instantané}} \mathcal{L}(\mathcal{C})$.
Th (Borne entropique). *Pour tout $\mathcal{C}$ à décodage unique, $\mathcal{L}(\mathcal{C}) \geqslant H(X)$, où $H(X) = -\sum_i pi \log_2(p_i)$ est l'entropie de la source, avec égalité si et seulement si $\forall i, p_i = 2^{-l_i}$.*
Th (Inégalité de Jensen). *Si f est convexe, alors $\mathbf{E}(f(X)) \geqslant f(\mathbf{E}(X))$. Si la convexité est stricte alors $(\mathbf{E}(f(X)) \geqslant f(\mathbf{E}(X))) \iff (f$ est constante$)$.*
Def. La **divergence de Kullback-Leibler**, ou entropie relative, de deux probabilités P et Q est définie par $D_{KL}(P\|Q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right)$.

C'est une mesure de dissimilarité entre les deux distributions de probabilités.
Cor. *On a $D_{KL}(P\|Q) \geqslant 0$ avec égalité si et seulement si $\forall i, p_i = q_i$.*