

# ACCQ 202 - Information theory

## 1 Source coding

Soit  $X$  une variable aléatoire discrète d'alphabet  $\mathcal{X}$  et de fonction de probabilité  $p$  telle que  $\forall x \in \mathcal{X}, p(x) = \mathbf{P}(X = x)$ . On note  $p(x)$  plutôt que  $p_X(x)$  par commodité, mais par  $p(x)$  et  $p(y)$  on fait référence à deux fonctions de probabilité distinctes.

**Def.**  $X$  est une **source d'information** si  $|\mathcal{X}| < \infty$  et on note  $\forall i \in \llbracket 1; |\mathcal{X}| \rrbracket, p_i = p(x_i) = \mathbf{P}(X = x_i)$ .

**Def.** **Code** pour une source  $X : \mathcal{C} : \mathcal{X} \rightarrow \{0, 1\}^*$ .

**Def.** **Longueur moyenne** d'un code  $\mathcal{C} : \mathcal{L}(\mathcal{C}) = \sum_i p_i l_i$  avec  $l_i$  la longueur du  $i^{\text{e}}$  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-punctue 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  $\sum_i l_i \leq 1$ . 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,  $\mathcal{L}(\mathcal{C}) \geq H(X)$ , où  $H(X) = -\sum_i p_i \log_2(p_i)$  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  $\mathbf{E}(f(X)) \geq f(\mathbf{E}(X))$ . Si la convexité est stricte alors  $(\mathbf{E}(f(X)) \geq f(\mathbf{E}(X))) \iff (f \text{ est constante})$ .

**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  $D_{KL}(P||Q) \geq 0$  avec égalité si et seulement si  $\forall i, p_i = q_i$ .

### Code de Shannon

On construit un code de Shannon en définissant les longueurs selon  $l_i = \left\lceil \log\left(\frac{1}{p_i}\right) \right\rceil$ , qui satisfait l'inégalité de Kraft et peut donc être utilisé pour produire un code instantané.

**Prop.** Soit  $\mathcal{C}$  un code de Shannon pour  $X$ . Alors  $H(X) \leq \mathcal{L}(\mathcal{C}) \leq H(X) + 1$ .

### Codage de Huffman

Pour construire un codage de Huffman on ordonne l'ensemble des  $p_i$ , puis l'on construit itérativement de nouveaux ensembles de probabilités en sommant à chaque étape les deux plus faibles. On repart ensuite à l'inverse : à partir de la dernière probabilité, égale à 1, on va re-diviser les probabilités de sorte à construire un arbre dont les feuilles correspondront aux  $p_i$ . La profondeur de chaque feuille  $i$  s'identifie alors à  $l_i$ .

**Th.** Un code de Huffman minimise  $\mathcal{L}(\mathcal{C})$ .

## 2 Entropie et questionnement

### Entropie et information mutuelle

On remarque que  $\mathcal{L}(\mathcal{C})$  s'identifie au nombre moyen de questions à poser pour identifier une valeur  $X \in \mathcal{X}$ .

**Th.** On a  $0 \leq H(X) \leq \log(|\mathcal{X}|)$ .

**Def.** Soit  $(X, Y) \sim p(x, y)$ . On a  $H(X, Y) = -\sum_{x,y} p(x, y) \log(p(x, y)) = -\mathbf{E}_{p(x,y)}(\log(p(X, Y)))$ . Et pour des v.a.  $X_1, \dots, X_n$  il vient  $H(X_1, \dots, X_n) = -\mathbf{E}_{p(x_1, \dots, x_n)}(\log(p(X_1, \dots, X_n)))$ .

**Def** (Entropie conditionnelle).  $H(Y | X) = \sum_x p(x) H(Y | X = x) = -\sum_{x,y} p(x, y) \log(p(y | x)) = -\mathbf{E}[\log(p(Y | X))]$

**Th** (Chain rule).  $H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i | X^{i-1})$  où  $X^i \triangleq X_1, \dots, X_i$ .

**Prop.** L'information mutuelle  $I(X; Y) = \sum_{x,y} p(x, y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$  vérifie

- $I(X; Y) = H(X) + H(Y) - H(X, Y)$
- $I(X; Y) = H(X) - H(X | Y) = H(Y) - H(Y | X) = I(Y; X)$
- $I(X; X) = H(X)$
- $I(X; Y) = D_{KL}(p_{X,Y} \| p_X \cdot p_Y)$
- $I(X; Y) = 0 \iff X \perp\!\!\!\perp Y$
- $H(Y | X) \leq H(Y)$
- $H(X^n) \leq \sum_{i=1}^n H(X_i)$
- $H(X)$  est concave en  $p_X$
- $H(f(X)) \leq H(x)$  pour toute fonction  $f$  déterministe.

**Def.** On définit  $H(X; Y | Z) = \sum_{x,y,z} p(x, y, z) \log\left(\frac{p(x,y|z)}{p(x|z)p(y|z)}\right) = H(X | Z) - H(X | Y, Z)$ .

**Th** (Chain rule sur  $I$ ).  $I(X_1, \dots, X_n; Y) = \sum_{i=1}^n I(X_i; Y | X^{i-1})$ .

### Typicalité

**Def.** Soit  $X_1, \dots, X_n$  i.i.d.. On appelle  $A_\varepsilon^n = \{x^n \mid 2^{-n(H(x)+\varepsilon)} \leq p(x^n) \leq 2^{-n(H(x)-\varepsilon)}\}$  **ensemble typique**.

**Th.** • Pour  $n$  suffisamment grand,  $\mathbf{P}(A_\varepsilon^n) \geq 1 - \varepsilon$ .  
•  $(1 - \varepsilon)2^{-n(H(x)-\varepsilon)} \leq |A_\varepsilon^n| \leq 2^{-n(H(x)+\varepsilon)}$

**Not.** On note  $a_n \doteq b_n$  si  $\forall \varepsilon > 0, \exists N \in \mathbf{N}, \forall n \geq N, \left|\frac{1}{n} \log\left(\frac{a_n}{b_n}\right)\right| \leq \varepsilon$ .

On a ici  $|A_\varepsilon^n| \doteq 2^{nH(X)}$ .

**Def.** L'ensemble des **séquences conjointement typiques** est

$\tilde{A}_\varepsilon^n = \{(x^n, y^n) \mid x^n \in A_\varepsilon^n, y^n \in A_\varepsilon^n, \left|-\frac{1}{n} \log p(x^n, y^n) - H(X, Y)\right| \leq \varepsilon\}$ , où  $p(x^n, y^n) \triangleq \prod_{i=1}^n p_{X,Y}(x_i, y_i)$ .

**Th.** Soit  $(X^n, Y^n) = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$  i.i.d. selon  $p_{X,Y}(x, y)$ . On a :

- $\forall \varepsilon > 0, \exists N \in \mathbf{N}, \forall n \geq N, \mathbf{P}(\tilde{A}_\varepsilon^n) \geq 1 - \varepsilon$ .
- $|\tilde{A}_\varepsilon^n| \doteq 2^{nH(X,Y)}$ .
- Si  $p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y)$  alors  $\mathbf{P}((X^n, Y^n) \in \tilde{A}_\varepsilon^n) \doteq 2^{-nI(X;Y)}$ .

### Entropie différentielle

**Def.** Soit  $X \sim f_X$ . Son **entropie différentielle** est donnée par  $h(X) = -\int f_X(x) \log(f_X(x)) dx$ .

**Def.** L'entropie différentielle de  $X$  par rapport à  $Y$  est  $h(X | Y) = -\int f_{X,Y}(x, y) \log f_{X|Y}(x | y) dx dy$ .

**Def.** La distance de Kullback-Leibler entre  $X \sim f$  et  $Y \sim g$  est  $D(X \| Y) \triangleq D(f \| g) = \int_{\mathbf{R}} f(x) \log_2\left(\frac{f(x)}{g(x)}\right) dx$ .

**Th.** On a  $D(f \| g) \geq 0$  avec égalité si et seulement si  $f = g$ .

**Def.** Soit  $(X, Y) \sim f_{X,Y}(x, y)$ . Leur **information mutuelle** est  $I(X; Y) = \iint f_{X,Y}(x, y) \log\left(\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\right) dx dy$

**Cor.** On a  $I(X; Y) \geq 0$  avec égalité si et seulement si  $X \perp\!\!\!\perp Y$ .

**Th.** Soit  $X$  une variable aléatoire et  $\bar{X}$  une variable aléatoire vectorielle.

- (i)  $h(X + c) = h(X)$
- (ii)  $h(aX) = h(X) + \log(|a|)$  et  $h(A\bar{X}) = h(\bar{X}) + \log(|\det(A)|)$  avec  $A$  une matrice.
- (iii) Si  $V(X) = \sigma^2$  alors  $h(X) \leq \frac{1}{2} \log(2\pi e \sigma^2)$  (entropie d'une variable gaussienne). Soit  $K = \mathbf{E}(\bar{X}^T \bar{X})$ , alors  $h(\bar{X}) \leq \frac{1}{2} \log((2\pi e)^n |K|)$ .

### 3 Transmission d'information

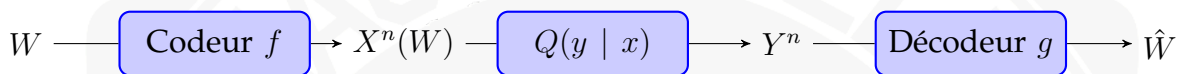
#### Chaîne de transmission

**Def.** Un canal de transmission est la donnée des probabilités  $\{Q(y | x), x \in \mathcal{X}, y \in \mathcal{Y}\}$ .

**Def.** La capacité d'information d'un canal est  $C = \max_{p_X} I(X, Y)$  (maximisation sur les distributions de  $X$ ), où  $(X, Y) \sim p_X \cdot Q_{Y|X}$ .

**Prop.** On a  $0 \leq C \leq \min\{\log(|\mathcal{X}|), \log(|\mathcal{Y}|)\}$ .

On a alors la chaîne de transmission suivante :



où  $W$  et  $\tilde{W}$  sont des fonctions aléatoires à valeurs dans  $\mathcal{M}$ , l'ensemble des mots.  $W$  est supposée suivre une loi uniforme. Pour un canal sans mémoire :  $\mathbf{P}(y^n | x^n(w)) = \prod_{i=1}^n Q(y_i | x_i(w))$ . Une **stratégie de transmission** désigne l'ensemble du codeur et du décodeur.

Nos données sont  $(M, n)$ , avec  $M = |\mathcal{M}|$ .

**Def.** On note  $R = \frac{\log_2 M}{n}$ , en bits d'informations, le **taux de performance**.

On veut maximiser  $R$  tout en minimisant  $P_e = \mathbf{P}(\tilde{W} \neq W) = \frac{1}{M} \sum_{w \in \mathcal{M}} \mathbf{P}(g(Y^n) \neq w | X^n(w))$ .

**Def.** Un taux  $R$  est dit **atteignable** s'il existe une suite de stratégies de codage  $((M = 2^{nR}, n))_{n \geq 1}$  telle que  $P_e^{(n)} \xrightarrow{n \rightarrow \infty} 0$ .

**Th.** Soit un canal  $Q(y | x)$ . On a :

- $\forall R, R < C$ ,  $R$  est atteignable,
- $\forall R, R > C$ ,  $R$  n'est pas atteignable.

**Not.** Si  $X, Y, Z$  forment une chaîne de Markov, on note  $X - Y - Z$ .

**Lem.** Si  $X - Y - Z$ ,  $I(X; Y) \geq I(X; Z)$ .

**Lem (Inégalité de Fano).** Soit une chaîne  $X - Y - \hat{X}$ . On a  $1 + \mathbf{P}(\hat{X} \neq X) \cdot \log(|\mathcal{X}|) \geq H(X | Y)$ .

**Lem.** Soit  $X^n$  et  $Y^n$  l'entrée et la sortie d'un canal donné. Alors  $I(X^n; Y^n) \leq nC$  avec  $C$  la capacité du canal.

#### Canal gaussien

Pour un canal gaussien on a  $Y = X + Z$  avec  $Z \sim \mathcal{N}(0, \sigma^2)$  indépendant de  $X$ .

Contrainte de puissance : le code  $\{x^n(m)\}_{m \in \mathcal{M}}$  doit satisfaire  $\forall m \in \mathcal{M}, \frac{1}{n} \sum_{i=1}^n x_i^2(w) \leq P$ .

**Th.** La capacité du canal gaussien  $(P, \sigma^2)$  est  $C = \max_{X, \mathbf{E}(X^2) \leq P} I(X; Y) = \frac{1}{2} \log \left( 1 + \frac{P}{\sigma^2} \right)$  où  $\frac{P}{\sigma^2}$  est le rapport signal sur bruit (SNR).

#### Codage conjoint source - canal

Soit  $S^n = S_1, S_2, \dots, S_m$  une source i.i.d. d'alphabet  $\mathcal{S}$ . Puisque  $|A_\varepsilon^n| \doteq 2^{mH(S)}$  on a besoin de  $m \cdot H(S)$  bits pour coder les séquences typiques, et l'on peut se restreindre à elles (compression sans erreur). Si l'on veut envoyer ensuite cette information à travers un canal de capacité  $C$  il suffit que la longueur  $n$  des mots code du codage de canal satisfasse  $\frac{\log M}{n} = R \cdot H(S) < C$ , où  $R = \frac{m}{n}$ .

Lorsque l'on veut envoyer  $S^n$  à travers un canal, on peut sans perte d'optimalité décomposer la procédure en 2 parties : compresser  $S^n$  (codage source) pour ensuite la protéger du bruit du canal en lui rajoutant de la redondance (codage de canal).

### Compression avec distorsion

**Def.** Une **fonction de distorsion** est de la forme  $d: \mathcal{X} \times \mathcal{X} \rightarrow \mathbf{R}_+$   
 $(x, \hat{x}) \mapsto d(x, \hat{x})$ .

**Ex.**

- Distorsion de Hamming :  $d(x, \hat{x})$  vaut 0 si  $\hat{x} = x$  et 1 sinon.
- Si  $\mathcal{X} = \mathbf{R}$ ,  $d(x, \hat{x}) = (x - \hat{x})^2$ .

**Def.** La **distorsion** entre  $x^n$  et  $\hat{x}^n$  est donnée par  $d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^n d(x_i, \hat{x}_i)$ .

**Not.** Fonction de codage :  $f_n: \mathcal{X}^n \rightarrow \llbracket 1; 2^{nR} \rrbracket$ . Fonction de décodage :  $g_n: \llbracket 1; 2^{nR} \rrbracket \rightarrow \mathcal{X}^n$ .

La distorsion associée à  $(2^{nR}, n)$  est  $D = \mathbf{E}[d(X^n, g_n \circ f_n(X^n))] = \sum_{x^n} p(x^n) d(x^n, g_n \circ f_n(x^n))$ .

**Def.**  $(R, D)$  est dit atteignable s'il existe  $((2^{nR}, n))_{n \geq 1}$  telle que  $\limsup_{n \rightarrow \infty} \mathbf{E}[d(X^n, g_n \circ f_n(X^n))] \leq D$ . On note  $R(D)$  le **taux atteignable minimal** avec distorsion  $D$ .

**Th.** Pour  $X$  donné,  $R(D) = \min_{p_{\hat{x}|x}} I(X; \hat{X})$ .

**Th.** Pour une source binaire i.i.d. de loi  $\mathcal{B}(p)$  avec  $p \leq \frac{1}{2}$  et une distorsion de Hamming, il vient

$$R(D) = \begin{cases} H_b(p) - H_b(D) & \text{si } 0 \leq D \leq p \\ 0 & \text{sinon} \end{cases}.$$