Introduction à la Théorie de l'Information et au Codage

Marc Lelarge

11 mars 2014

Table des matières

Notations

Pour des variables aléatoires (v.a.) discrètes X et Y à valeurs dans $\mathcal{X}$ et $\mathcal{Y}$ resp. n'ayant pas la même distribution, on utilisera les notations suivantes pour $x \in \mathcal{X}$ et $y \in \mathcal{Y}$:

$$\begin{aligned} p(x) &= P(X = x) \\ p(y) &= P(Y = y) \\ p(x, y) &= P(X = x, Y = y) \\ p(x|y) &= P(X = x|Y = y) = p(x, y)/p(y). \end{aligned}$$

La fonction $p(.)$ est donc définie par son argument : $p(x)$ étant la distribution de la v.a. X et $p(y)$ celle de Y. On a donc par exemple :

$$p(x) = \sum_y p(x, y) \text{ et } p(y) = \sum_x p(x, y).$$

Si deux v.a. X et Y sont indépendantes alors $p(x, y) = p(x)p(y)$. Lorsque ces notations sont ambigues, on pourra écrire $p_X(x), p_Y(y), p_{X,Y}(x, y), p_{X|Y}(x|y)$.

Chapitre 1

Suites typiques et compression de données avec pertes

1.1 Entropie

Une source (discrète) émet une suite de v.a. $\{U_i\}_{i=1}^{\infty}$ à valeurs dans un ensemble fini $\mathcal{U}$ appelé l'alphabet de la source. Si les U_i sont indépendants et identiquement distribués (i.i.d.) de loi P, la source est dite sans mémoire de distribution P. On notera également P la loi d'un vecteur de v.a. i.i.d. $(U_1, U_2, \ldots, U_k) \in \mathcal{U}^k$. Ainsi pour un ensemble A de $\mathcal{U}^k$, on note $P((U_1, U_2, \ldots, U_k) \in A)$ la probabilité que le vecteur appartienne à A.

Définition 1.1.1 *Soit U une variable aléatoire à valeurs dans un ensemble fini $\mathcal{U}$, de distribution de probabilité :*

$$p(u) = P(U = u),\ u \in \mathcal{U}\,.$$

Son entropie est, par définition, la quantité

$$H(U) = -E\left[\log(p(U))\right] = -\sum_{u\in\mathcal{U}} p(u)\log p(u)\,,$$

avec la convention $0\log 0 = 0$.

Le choix de la base du logarithme correspond à un choix d'unité. Sauf mention du contraire, on choisit par défaut la base 2. L'entropie s'exprime alors en bits.

On parle souvent de l'entropie d'une variable aléatoire mais il est clair par la définition que celle-ci ne dépend que de la distribution de cette variable aléatoire. On parlera alors indifféremment de l'entropie d'une distribution de probabilité.

1.2 Ensemble typique

On rappelle la loi faible des grands nombres : étant donné une suite de v.a. $(X, X_i, 1 \leq i)$ i.i.d. de moyenne finie $E[X]$, on a pour tout $\epsilon > 0$ lorsque n tend vers l'infini :

$$P\left(\left|\frac{1}{n}(X_1 + \cdots + X_n) - E[X]\right| \geq \epsilon\right) \to 0$$

On dit que la moyenne empirique $\frac{1}{n}(X_1 + \cdots + X_n)$ converge en probabilité vers sa moyenne $E[X]$.

Dans notre cadre, on a donc :

$$-\frac{1}{n}\log p(U_1, U_2, \ldots, U_n) = -\frac{1}{n}\sum_i p(U_i) \to -E[\log p(U)] = H(U),$$

où la convergence est en probabilité. On notera bien que le terme de gauche est aléatoire alors que celui de droite est déterministe. En particulier, $p(u_1, u_2, \ldots, u_n)$ est la loi du n-uplet $(U_1, U_2, \ldots, U_n)$ qui est donc égale à $\prod_i p(u_i)$. On note donc $p(U_1, U_2, \ldots, U_n)$ la valeur prise par cette fonction au point aléatoire $(U_1, U_2, \ldots, U_n)$ tiré selon la loi $\prod_i p(u_i)$.

Définition 1.2.1 *Pour $n \in \mathbb{N}$ et $\delta > 0$, l'ensemble typique $A_\delta^{(n)}$ par rapport à la distribution $p(u)$ est l'ensemble des suites $(u_1, \ldots, u_n) \in \mathcal{U}^n$ telles que :*

$$2^{-n(H(U)+\delta)} \leq p(u_1, \ldots, u_n) = \prod_{i=1}^{n} p(u_i) \leq 2^{-n(H(U)-\delta)}.$$

Théorème 1.2.1 *Pour tout $n \in \mathbb{N}$, $\delta > 0$, on a :*

1. *Si $(u_1, \ldots, u_n) \in A_\delta^{(n)}$ alors*

$$H(U) - \delta \leq -\frac{1}{n}\log p(u_1, \ldots, u_n) \leq H(U) + \delta.$$

2. *Pour tout $\epsilon > 0$ et pour n suffisamment grand, on a :*

$$P\left(A_\delta^{(n)}\right) = P\left((U_1, \ldots, U_n) \in A_\delta^{(n)}\right) \geq 1 - \epsilon.$$

3. Le cardinal de l'ensemble $A_\delta^{(n)}$ est borné par :

$$\left|A_\delta^{(n)}\right| \leq 2^{n(H(U)+\delta)},$$

et pour tout $\epsilon > 0$, pour n suffisamment grand, ce cardinal est minoré par :

$$\left|A_\delta^{(n)}\right| \geq (1-\epsilon)2^{n(H(U)-\delta)}.$$

Remarque 1.2.1 *Le point 3. du Théorème entraine directement :*

$$H(U) - \delta \leq \liminf_{n\to\infty} \frac{\log\left|A_\delta^{(n)}\right|}{n} \leq \limsup_{n\to\infty} \frac{\log\left|A_\delta^{(n)}\right|}{n} \leq H(U) + \delta.$$

Démonstration. Le point 1 est une application directe de la définition. Le point 2 découle de la loi faible des grands nombres en écrivant :

$$-\frac{1}{n}\log p\left(U_1, U_2, \ldots, U_n\right) = -\frac{1}{n}\sum_{i=1}^{n} \log p(U_i),$$

qui est une somme de v.a. i.i.d. de moyenne $-E\left[\log p(U)\right] = H(U)$.

Pour la première partie du point 3, on écrit :

$$1 \;=\; \sum_{(u_1,\ldots,u_n)\in\mathcal{U}^n} p(u_1,\ldots,u_n) \geq \sum_{(u_1,\ldots,u_n)\in A_\delta^{(n)}} p(u_1,\ldots,u_n) \geq \left|A_\delta^{(n)}\right| 2^{-n(H(U)+\delta)},$$

où la dernière inégalité provient de la définition de l'ensemble typique.

Pour la seconde partie du point 3, on a grâce au point 2 pour n suffisamment grand :

$$1-\epsilon \leq P\left(A_\delta^{(n)}\right) \leq \left|A_\delta^{(n)}\right| 2^{-n(H(U)-\delta)}.$$

□

1.3 Codage de source avec perte

Nous considérons dans ce chapitre une notion très générale de codage que nous préciserons dans le chapitre suivant. Un (k,n)-codage binaire est une paire de fonctions

$$f : \mathcal{U}^k \to \{0,1\}^n, \text{ et } \phi : \{0,1\}^n \to \mathcal{U}^k.$$

Pour une source donnée, la probabilité d'erreur du code (f, ϕ) est

$$e(f, \phi) := P(\phi(f(U^{(k)})) \neq U^{(k)}),$$

avec $U^{(k)} = (U_1, \ldots, U_k) \in \mathcal{U}^k$ les k premiers symboles émis par la source.

Le but de la compression est de trouver des codes avec un ratio n/k petit et une probabilité d'erreur petite. Plus précisément, pour tout k, soit $n(k, \epsilon)$ le plus petit entier n tel qu'il existe un (k, n)-code satisfaisant $e(f, \phi) \leq \epsilon$.

Théorème 1.3.1 *Pour une source discrète sans mémoire de distribution $P(U = u) = p(u)$, on a pour tout $\epsilon \in (0, 1)$:*

$$\lim_{k \to \infty} \frac{n(k, \epsilon)}{k} = H(U) = -\sum_{u \in \mathcal{U}} p(u) \log p(u).$$

Démonstration. L'existence d'un (k, n)-code binaire (f, ϕ) avec $e(f, \phi) \leq \epsilon$ est équivalente à l'existence d'un ensemble $A \subset \mathcal{U}^k$ avec $P(A) \geq 1 - \epsilon$ et $|A| \leq 2^n$. A est alors l'ensemble des suites $u^{(k)} \in \mathcal{U}^k$ reproduites de manière exacte, c-à-d. telles que $\phi(f(u^{(k)})) = u^{(k)}$.

Soit $s(k, \epsilon)$ la taille minimale d'un ensemble $A \subset \mathcal{U}^k$ avec $P(A) \geq 1 - \epsilon$, c.a.d

$$s(k, \epsilon) = \min\{|A|;\ P(A) \geq 1 - \epsilon\}.$$

Pour prouver le théorème, il suffit de montrer que pour $\epsilon \in (0, 1)$,

$$\lim \frac{\log s(k, \epsilon)}{k} = H(U). \tag{1.1}$$

Pour tout $\delta > 0$, en prenant $A = A_\delta^{(k)}$ l'ensemble typique pour la source $p(u)$, on a pour k suffisamment grand $P\left(A_\delta^{(k)}\right) \geq 1 - \epsilon$ et donc :

$$s(k, \epsilon) \leq |A_\delta^{(k)}| \leq 2^{k(H(U)+\delta)},$$

donc

$$\limsup_k \frac{\log s(k, \epsilon)}{k} \leq H(U). \tag{1.2}$$

Inversement, pour tout $A \subset \mathcal{U}^k$ avec $P(A) \geq 1 - \epsilon > 0$, le point 2 du Théorème 1.2.1 implique que pour k suffisamment grand $P\left(A_\delta^{(k)}\right) \geq \frac{1+\epsilon}{2}$ et donc

$$P(A \cap A_\delta^{(k)}) \geq P(A) - (1 - P\left(A_\delta^{(k)}\right)) \geq \frac{1 - \epsilon}{2}.$$

On a donc par définition de $A_\delta^{(k)}$,

$$|A| \geq |A \cap A_\delta^{(k)}| \geq \sum_{u^{(k)} \in A \cap A_\delta^{(k)}} p(u^{(k)}) 2^{k(H(U)-\delta)} \geq \frac{1-\epsilon}{2} 2^{k(H(U)-\delta)},$$

et donc pour tout $\delta > 0$,

$$\liminf_{k\to\infty} \frac{1}{k} \log s(k,\epsilon) \geq H(U) - \delta.$$

Ceci, avec (1.2), implique (1.1). □

Corollaire 1.3.1

$$0 \leq H(U) \leq \log |\mathcal{U}|$$

1.4 Convexité et propriétés de l'entropie

Un vecteur $(p_i : 1 \leq i \leq n)$ est une distribution de probabilité si $p_i \geq 0$ pour tout $1 \leq i \leq n$ et $\sum_{i=1}^n p_i = 1$.

Lemme 1.4.1 *Si $(p_i : 1 \leq i \leq n)$ est une distribution de probabilité alors le minimum de la fonction*

$$G(q_1, \ldots, q_n) = -\sum p_i \log q_i,$$

sur toutes les distributions de probabilité $(q_1, \ldots, q_n)$ est atteint uniquement pour $q_k = p_k$, $1 \leq k \leq n$.

Démonstration. En utilisant l'inégalité de convexité $\log z \leq (z-1)\log e$ qui est une égalité uniquement lorsque $z = 1$, on obtient :

$$\log\left(\frac{q_k}{p_k}\right) \leq \left(\frac{q_k}{p_k} - 1\right) \log e,$$

avec égalité si et seulement si $q_k = p_k$. On a donc

$$G(p_1, \ldots, p_n) - G(q_1, \ldots, q_n) = \sum_k p_k \log\left(\frac{q_k}{p_k}\right) \leq \log e \sum_k (q_k - p_k) = 0.$$

□

On en déduit facilement les théorème suivants :

Théorème 1.4.1 *Pour tout n, $H(p_1, \ldots, p_n) \leq \log n$, avec égalité si et seulement si $p_1 = p_2 = \cdots = p_n = 1/n$.*

Démonstration. D'après le lemme précédent, on a :

$$H(p_1, \ldots, p_n) = G(p_1, \ldots, p_n) \leq G(1/n, \ldots, 1/n) = \log n,$$

avec égalité si et seulement si $p_i = 1/n$ pour tout $1 \leq i \leq n$. □

L'entropie d'une paire (X, Y) ne nécessite pas de nouvelle définition ! On notera $H((X, Y)) = H(X, Y)$ qui est donc l'entropie de la distribution du couple (X, Y) à valeur dans $\mathcal{X} \times \mathcal{Y}$ où $\mathcal{X}$ et $\mathcal{Y}$ sont les alphabets respectifs des v.a. X et Y.

Théorème 1.4.2 *Si X et Y sont des v.a. (discrètes) alors $H(X, Y) \leq H(X) + H(Y)$, avec égalité si et seulement si X et Y sont indépendantes.*

Démonstration. On a

$$\begin{aligned} H(X) + H(Y) &= -\sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{y \in \mathcal{Y}} p(y) \log p(y) \\ &= -\sum_{x,y} p(x, y) \log p(x) - \sum_{x,y} p(x, y) \log p(y) \\ &= -\sum_{x,y} p(x, y) \log p(x)p(y). \end{aligned}$$

Donc par le Lemme 1.4.1, on a

$$H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y) \leq -\sum_{x,y} p(x, y) \log p(x)p(y) = H(X) + H(Y),$$

avec égalité si et seulement si $p(x, y) = p(x)p(y)$, c'est-à-dire si X et Y sont indépendantes. □

1.5 Entropie conditionnelle et information mutuelle

Étant donné une v.a. X sur un espace de probabilité Ω et A un événement dans Ω, on définit l'entropie conditionnelle de X sachant A par

$$H(X|A) = -\sum_{k=1}^{m} P(X = x_k|A) \log P(X = x_k|A).$$

De la même manière si Y est une autre v.a., on définit l'entropie conditionnelle de X sachant Y par

$$\begin{aligned} H(X|Y) &= \sum_{j=1}^{m} H(X|Y=y_j)P(Y=y_j) \\ &= -\sum_{x,y} p(x,y)\log p(x|y). \end{aligned}$$

Il est facile de vérifier les propriétés suivantes :

$$\begin{aligned} H(X|X) &= 0 \\ H(X|Y) &= H(X) \text{ si } X \text{ et } Y \text{ sont indépendantes} \\ H(X|Y) &= 0 \text{ si et seulement si } X = g(Y) \text{ pour une fonction } g. \end{aligned}$$

Pour la dernière propriété, il suffit d'écrire $H(X|Y) = \sum H(X|Y=y_i)P(Y=y_i)$ donc pour que $H(X|Y)=0$, il faut que $H(X|Y=y_i)=0$ pour chaque i, c'est-à-dire qu'il existe x_i tel que $P(X=x_i|Y=y_i)=1$. Donc X est déterminé par Y.

La différence $H(X,Y)-H(X)$ mesure la quantité d'information supplémentaire sur le couple (X,Y) donnée par Y si X est déjà connu. Comme montré dans le théorème suivant, cette différence est l'entropie conditionnelle de Y sachant X.

Théorème 1.5.1 *Pour toute paire de v.a. X, Y, on a $H(X,Y) = H(Y)+H(X|Y)$.*

Démonstration. On écrit

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} p(x,y)\log p(x,y) \\ &= -\sum_{x,y} p(x,y)\log p(y)p(x|y) \\ &= -\sum_{y} p(y)\log p(y) - \sum_{x,y} p(x,y)\log p(x|y), \end{aligned}$$

ce qui est l'égalité souhaitée. □

Corollaire 1.5.1 *Pour toute paire de v.a. X, Y, $H(X|Y) \leq H(X)$ avec égalité si et seulement si X et Y sont indépendantes.*

Démonstration. On a $H(X|Y) = H(X,Y) - H(Y)$ et $H(X,Y) \leq H(X)+H(Y)$ avec égalité si et seulement si X et Y sont indépendantes. Le résultat en découle. □

Définition 1.5.1 *L'information mutuelle entre X et Y est définie par*

$$I(X;Y) = H(Y) - H(Y|X) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y).$$

L'information mutuelle entre X et Y correspond à la diminution d'incertitude sur Y causée par la connaissance de X, c'est-à-dire la quantité d'information sur Y contenue 'dans' X. Elle est symétrique en X et Y.

1.6 Test d'hypothèse

Problème : décider entre deux distributions P et Q à partir d'un échantillon de taille k, i.e. le résultat de k tirages indépendants. Un test est défini par un ensemble $A \subset \mathcal{U}^k$: si l'échantillon $(U_1, \ldots, U_k)$ appartient à A alors le test retourne l'hypothèse P sinon Q.

On considère un scénario où les hypothèses ne sont pas symétriques. On désire une probabilité d'erreur au plus ϵ si P est la vraie distribution : $P(A) \geq 1 - \epsilon$. Le but est alors de minimiser la probabilité d'erreur si l'hypothèse Q est vraie,

$$\beta(k, \epsilon) = \min\{Q(A), \text{ t.q. } A \subset \mathcal{U}^k, \ P(A) \geq 1 - \epsilon\}.$$

Définition 1.6.1 *L'entropie relative ou distance de Kullback-Leibler entre deux distributions p et q est définie par :*

$$D(p\|q) = E_p\left[\log \frac{p(U)}{q(U)}\right] = \sum_{u \in \mathcal{U}} p(u) \log \frac{p(u)}{q(u)},$$

avec les conventions $0 \log \frac{0}{0} = 0$, $0 \log \frac{0}{q} = 0$ *et* $p \log \frac{p}{0} = \infty$.

Le Lemme 1.4.1 implique directement :

Corollaire 1.6.1 *La distance de Kullback-Leibler est non-négative :* $D(p\|q) \geq 0$ *et* $D(p\|q) = 0$ *si et seulement si* $p = q$.

Par contre, la distance de Kullback-Leibler n'est pas une vraie distance car elle n'est pas symétrique.

Théorème 1.6.1 *Pour tout* $\epsilon \in (0, 1)$, *on a*

$$\lim_{k \to \infty} \frac{1}{k} \log \beta(k, \epsilon) = -D(p\|q).$$

Avant de démontrer ce théorème, nous allons généraliser la notion d'ensemble typique. Pour deux distributions p et q avec $q(u) > 0$ pour tout $u \in \mathcal{U}$, on définit l'ensemble $A_\delta^{(n)}(p\|q)$ par l'ensemble des suites $(u_1, \ldots, u_n) \in \mathcal{U}^n$ telles que :

$$2^{-n(D(p\|q)+\delta)} \leq \frac{q(u_1, \ldots, u_n)}{p(u_1, \ldots, u_n)} \leq 2^{-n(D(p\|q)-\delta)}.$$

Noter que s'il existe $u \in \mathcal{U}$ tel que $p(u) = 0$ alors pour $(u_1, \ldots, u_n) \in A_\delta^n(p\|q)$, on a $u_i \neq u$ pour tout i.

Théorème 1.6.2 *Si pour tout $u \in \mathcal{U}$, on a $q(u) > 0$, alors pour tout $\delta > 0$, on a :*

1. *pour tout $n \in \mathbb{N}$, si $(u_1, \ldots, u_n) \in A_\delta^{(n)}(p\|q)$ alors*

$$D(p\|q) - \delta \leq \frac{1}{n} \log \frac{p(u_1, \ldots, u_n)}{q(u_1, \ldots, u_n)} \leq D(p\|q) + \delta.$$

2. *Pour tout $\epsilon > 0$ et pour n suffisamment grand, on a :*

$$P\left(A_\delta^{(n)}(p\|q)\right) \geq 1 - \epsilon.$$

3. *Pour tout $\epsilon > 0$ et pour n suffisamment grand, on a :*

$$(1-\epsilon)2^{-n(D(p\|q)+\delta)} \leq Q\left(A_\delta^{(n)}(p\|q)\right) \leq 2^{-n(D(p\|q)-\delta)}.$$

Démonstration. La démonstration est similaire à celle pour l'ensemble typique. Le premier point découle de la définition et le point 2 de la loi faible des grands nombres. Pour le point 3, on écrit :

$$\begin{aligned} Q\left(A_\delta^{(n)}(p\|q)\right) &= \sum_{(u_1,\ldots,u_n)\in A_\delta^{(n)}(p\|q)} q(u_1, \ldots, u_n) \\ &\leq \sum_{(u_1,\ldots,u_n)\in A_\delta^{(n)}(p\|q)} p(u_1, \ldots, u_n) 2^{-n(D(p\|q)-\delta)} \leq 2^{-n(D(p\|q)-\delta)}, \end{aligned}$$

et pour l'autre borne :

$$Q\left(A_\delta^{(n)}(p\|q)\right) \geq \sum_{(u_1,\ldots,u_n)\in A_\delta^{(n)}(p\|q)} p(u_1, \ldots, u_n) 2^{-n(D(p\|q)+\delta)} \geq (1-\epsilon)2^{-n(D(p\|q)+\delta)},$$

où la dernière inégalité découle du point 2. □

Intuitivement, l'ensemble $A_\delta^{(n)}(p\|q)$ permet de distinguer les distributions p et q : si la suite est tirée selon P, alors elle appartient avec grande probabilité

à $A_\delta^{(n)}(p\|q)$ tandis que si elle est tirée selon Q, la probabilité d'appartenance à $A_\delta^{(n)}(p\|q)$ décroît exponentiellement vite vers zéro.

Démonstration.(du Théorème 1.6.1)
Le cas $q(u) > 0$ pour tout $u \in \mathcal{U}$ découle directement du théorème précédent. En effet, on a directement,

$$\frac{1}{k}\log\beta(k,\epsilon) \leq \frac{1}{k}\log Q\left(A_\delta^{(n)}(p\|q)\right) \leq -D(p\|q) + \delta.$$

Inversement, pour tout $A \subset \mathcal{U}^k$ avec $P(A) \geq 1-\epsilon$, on a $P\left(A \cap A_\delta^{(k)}(p\|q)\right) \geq 1-2\epsilon$ pour k suffisamment grand et donc

$$Q(A) \geq Q\left(A \cap A_\delta^{(k)}(p\|q)\right) \geq (1-2\epsilon)2^{-k(D(p\|q)+\delta)}.$$

On a donc $\frac{1}{k}\log s(k,\epsilon) \geq \frac{1}{k}\log(1-2\epsilon) - D(p\|q) - \delta$.

On considère maintenant le cas où il existe $u \in \mathcal{U}$ tel que $p(u) > 0$ et $q(u) = 0$. L'idée est alors la suivante : si on observe le symbole u alors la vraie distribution ne peut être que P. On définit donc $A = \{u^{(k)} \in \mathcal{U}^k,\ \exists i \leq k,\ u_i = u\}$ de telle sorte que $P(A) = 1-(1-p(u))^k \geq 1-\epsilon$ pour k suffisament grand. On a bien sur $Q(A) = 0$ donc $\beta(k,\epsilon) = 0$ pour k suffisament grand. □

1.7 Exercice : Définition axiomatique de l'entropie

Étant donné une distribution de probabilité $p_1, \ldots, p_n$, on cherche une fonction $H(p_1, \ldots, p_n)$ quantifiant "l'incertitude" associée à cette distribution. On postule les conditions suivantes pour la fonction H :

(A1) $H(p_1, \ldots, p_n)$ est maximum pour $p_1 = p_2 = \cdots = p_n = 1/n$.

(A2) H est une fonction symétrique en ses arguments.

(A3) $H(p_1, \ldots, p_n) \geq 0$ avec égalité quand un des p_i vaut 1.

(A4) $H(p_1, \ldots, p_n, 0) = H(p_1, \ldots, p_n)$.

(A5) $H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) \leq H\left(\frac{1}{n+1}, \ldots, \frac{1}{n+1}\right)$.

(A6) la fonction H est continue.

(A7) pour des entiers m et n,

$$H\left(\frac{1}{mn}, \ldots, \frac{1}{mn}\right) = H\left(\frac{1}{m}, \ldots, \frac{1}{m}\right) + H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right).$$

(A8) pour $p = p_1 + \cdots + p_m$ et $q = q_1 + \cdots + q_n$ où tous les p_i, q_i sont positifs. Si p et q sont strictement positifs tels que $p + q = 1$, on a :

$$H(p_1, \ldots, p_m, q_1, \ldots, q_n) = H(p, q) + pH(p_1/p, \ldots, p_m/p) + qH(q_1/q, \ldots, q_n/q).$$

1. Jusifier les différents axiomes.
2. Montrer que la fonction $g(n) = H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ est de la forme $g(n) = A \ln n$ pour un certain A.
3. En déduire la forme de $H(p, 1-p)$ quand p est rationnel.
4. Conclure.

Chapitre 2

Codage pour des sources discrètes

2.1 Mots code de longueur variable

Définition 2.1.1 *Un code de source C pour une variable aléatoire $U \in \mathcal{U}$ est une fonction de $\mathcal{U}$ vers $\mathcal{D}^*$ (l'ensemble des mots finis sur un alphabet D-aire). $C(u)$ est le mot code correspondant à u, et $l(u)$ sa longueur.*

Définition 2.1.2 *La longueur moyenne $L(C)$ d'un code pour la variable aléatoire U de distribution de probabilité $p(u)$ est :*

$$L(C) = \sum_{u \in \mathcal{U}} p(u) l(u)$$

Remarque 2.1.1 *Dans la suite, on supposera sans perte de généralité : $\mathcal{D} = \{0, 1, \ldots, d-1\}$.*

Exemple 2.1.1: Sur un alphabet binaire, si :

$$\begin{array}{ll} P(U=1) = 1/2 & C(1) = 0 \\ P(U=2) = 1/4 & C(2) = 10 \\ P(U=3) = 1/8 & C(3) = 110 \\ P(U=4) = 1/8 & C(4) = 111 \end{array}$$

on a $H(U) = 1.75$ bits et $L(C) = 1.75$ bits. Par exemple : 0110111100110 se décode en 134213.

Définition 2.1.3 *Un code C est dit* non-ambigu *si :*

$$x \neq y \Rightarrow C(x) \neq C(y).$$

Définition 2.1.4 *L'extension C^* du code C est la fonction des mots finis de $\mathcal{U}$ vers les mots finis de $\mathcal{D}$ définie par*

$$C(u_1 \dots u_n) := C(u_1) \dots C(u_n) \text{ (concaténation)}$$

Définition 2.1.5 *Un code C est dit* uniquement décodable *si son extension C^* est non-ambigüe.*

Définition 2.1.6 *Un code est dit* instantané *si aucun mot code n'est le préfixe d'un autre mot code.*

EXEMPLE 2.1.2: Parmi les codes suivants :

U	code 1	code 2	code 3
1	0	10	0
2	010	00	10
3	01	11	110
4	10	110	111

Le code 3 est instantané.
Le code 2 est uniquement décodable (il suffit de regarder la parité du nombre de 0 après 11).
Le code 1 est non-ambigu, mais non uniquement décodable, par exemple 010 peut se décoder par $2, 14$ ou 31.

Un code instantané est uniquement décodable. De plus un code instantané peut être décodé sans référence aux mots code future puisque la fin d'un mot code est reconnaissable immédiatement.

Théorème 2.1.1 (Inégalité de Kraft) *Pour un code instantané sur un alphabet de taille D, les longueurs des mots code $l_1, \dots, l_m$ doivent vérifier :*

$$\sum_i D^{-l_i} \leq 1$$

Inversement, étant donné une suite de longueurs vérifiant cette inégalité, il existe un code instantané avec des mots code ayant ces longueurs.

Démonstration. Pour prouver le premier point, on peut considérer l'arbre de codage du code C.

Soit l_{max} la longueur du plus long mot code. Un mot code de longueur l_i a $D^{l_{max}-l_i}$ descendants à la profondeur l_{max} qui doivent être disjoints des descendants des autres mots code par la propriété du préfixe. On a donc :

$$\sum_i D^{l_{max}-l_i} \leq D^{l_{max}}.$$

Inversement étant donné des longueurs $\ell_1, \ldots, \ell_m$ satisfaisant l'inégalité de Kraft, on peut toujours construire un arbre de codage comme précédemment : dans l'arbre D-aire, associer au premier (pour l'ordre lexicographique de l'arbre) noeud de profondeur l_1 le mot code 1 et retirer ses descendants. Associer alors au premier noeud restant de profondeur l_2 le mot code 2 et ainsi de suite. □

Théorème 2.1.2 (McMillan) *Les longueurs des mots code d'un code D-aire uniquement décodable doivent satisfaire l'inégalité de Kraft.*

Une conséquence immédiate est que les codes instantanés seront tout aussi performants (pour ce qui concerne leurs longueurs) que les codes uniquement décodables (et pas moins, comme on aurait pu le penser).

Démonstration. Soit k un entier.

On a :

$$\begin{aligned}\left(\sum_{u\in\mathcal{U}} D^{-l(u)}\right)^k &= \sum_{u_1\in\mathcal{U}} \cdots \sum_{u_k\in\mathcal{U}} D^{-l(u_1)-l(u_2)\cdots-l(u_k)} \\ &= \sum_{(u_1,\ldots,u_k)\in\mathcal{U}^k} D^{-l(u_1\ldots u_k)} \\ &= \sum_{m=1}^{kl_{max}} A(m) D^{-m}\end{aligned}$$

où :

$$A(m) = |\{(u_1 \ldots u_k) \in \mathcal{U}^k, l(u_1 \ldots u_k) = m\}|$$

On a $A(m) \leq D^m$ car le code est uniquement décodable, et donc :

$$\sum_{u\in\mathcal{U}} D^{-l(u)} \leq (kl_{max})^{1/k}$$

Or $(kl_{max})^{1/k} \xrightarrow[k\to\infty]{} 1$, ce qui conclut la preuve □

2.2 Un théorème de codage de source

Théorème 2.2.1 *Etant donné une source discrète à valeurs dans $\mathcal{U}$ et d'entropie $H(U)$, et étant donné un alphabet de D symboles pour le code, il est possible de coder chaque lettre de la source de manière instantanée et telle que la longueur moyenne des mots satisfasse : $L(C) < \frac{H(U)}{\log D} + 1$.*

De plus, pour tout code uniquement décodable : $L(C) \geq H(U)/\log D$.

Démonstration. On a :

$$\begin{aligned} H(U) - L(C)\log D &= \sum_u p(u)\log\frac{1}{p(u)} - \sum_u p(u)l(u)\log D \\ &= \sum_u p(u)\log\frac{D^{-l(u)}}{p(u)} \end{aligned}$$

On sait par ailleurs que pour $z > 0$, $\log z \leq (z-1)\log e$, on a donc :

$$\begin{aligned} H(U) - L(C) &\leq (\log e)\big(\sum_u D^{-l(u)} - \underbrace{\sum_u p(u)}_{=1}\big) \\ &\leq 0\text{par (Mcmillan)} \end{aligned}$$

Pour l'autre inégalité, on choisit $l(u)$ tel que $D^{-l(u)} \leq p(u) < D^{-l(u)+1}$.
On a donc :

$$\sum_u D^{-l(u)} \leq 1$$

D'après l'inégalité de Kraft, il existe donc un code instantané avec ces longueurs, de plus

$$\begin{aligned} \log p(u) &< (-l(u)+1)\log D \\ l(u) &< \frac{-\log p(u)}{\log D} + 1 \end{aligned}$$

et

$$L(C) = \sum p(u)l(u) < \frac{H(U}{\log D} + 1$$

□

Théorème 2.2.2 *Pour une source discrète sans mémoire d'entropie $H(U)$ et un alphabet à D symboles, il est possile de coder les suites de k lettres de la source de sorte que :*

1. *La propriété du préfixe soit satisfaite*
2. *La longueur moyenne des mots code par lettre source vérifie :*

$$H(U)/\log D \leq L^k/k < H(U)/\log D + 1/k$$

où $L^k = \sum_{u_1 \dots u_k} l(u_1 \dots u_k) p(u_1 \dots u_k)$

Démonstration. Il suffit de vérifier que $H(U^{(k)}) = kH(U)$, et le résultat découle du théorème précédent.

On a bien $H(U^{(k)}) = \sum_{u_1 \dots u_k} p(u_1 \dots u_k) \log p(u_1 \dots u_k)$ et par la propriété sans mémoire de la source, $p(u_1 \dots u_k) = p(u_1) \dots p(u_k)$, d'où le résultat.

□

2.3 Un codage optimal : le codage de Huffman

On présente ici le codage de Huffman. Il est optimal en ce sens qu'il n'existe pas de code uniquement décodable avec une longueur moyenne inférieure.

Dans toute la suite, on se restreint aux codes instantnés sans perte de généralité par l'inégalité de McMillan.

2.3.1 Cas du code binaire : $D = 2$

On suppose que $\mathcal{U} = \{U_1, \dots, U_k\}$, avec $p(u_1) \geq p(u_2) \geq \dots \geq p(u_k)$.

Lemme 2.3.1 *Pour tout $k \geq 2$, un code binaire optimal existe pour lequel les mots code les moins probables $C(u_k)$ et $C(u_{k-1})$ ont la même longueur et diffèrent par le dernier bit. (disons que $C(u_k)$ finit par 1 et $C(u_{k-1})$ par 0)*

Démonstration. Si $l(u_k) < \max_i l(u_i) := l(u_j)$ alors on obtient un meilleur code en interchangeant les mots codant u_k et u_j. En effet, on change $L(C)$ de :

$$\begin{aligned} \Delta &= p(u_j)l(u_k) + p(u_k)l(u_j) - p(u_k)l(u_k) - p(u_j)l(u_j) \\ &= (p(u_j) - p(u_k))(l(u_k) - l(u_j)) \leq 0. \end{aligned}$$

On peut donc supposer $l(u_k) = \max_i l(u_i)$.

Si maintenant $C(u_k)$ est le seul mot de longueur $l(u_k)$, on obtient un meilleur code instantané en tronquant les derniers bits de $C(u_k)$. Donc il existe u_i tel que $l(u_i) = l(x_k)$ et $C(u_i)$ et $C(u_k)$ diffèrent dans le dernier digit. Donc $l(u_i) \geq l(u_{k-1})$ et par le même argument que précédemment, interchanger $C(u_i)$ et $C(u_{k-1})$ n'augmente pas $L(C)$. □

On a donc réduit le problème de construction d'un code optimal à celui de construire $C(u_1), \ldots, C(u_{k-2})$ et trouver les $l(u_k) - 1$ premiers digits de $C(u_k)$.

On définit maintenant l'ensemble réduit : $\mathcal{U}' = \{u'_1, \ldots, u'_{k-1}\}$ avec la v.a U' associée : $p(u'_j) = p(u_j)$ si $j \leq k-2$ et $p(u'_{k-1}) = p(u_k) + p(u_{k-1})$.

Il y a une bijection entre les codes instantanés pour U' et les codes instantanés pour U pour lesquels $C(u_k)$ et $C(u_{k-1})$ ne diffrent que par le dernier digit, $C(u_k)$ finissant par un 1 et $C(u_{k-1})$ par un 0.

Lemme 2.3.2 *Si un code instantané est optimal pour U', le code instantané correspondant pour U est optimal.*

Démonstration.

$$l(u_j) = \begin{cases} l(u'_j) & \text{si } j \leq k-2 \\ l(u'_{k-1}) + 1 & \text{si } j \geq k-1 \end{cases}$$

Donc,

$$\begin{aligned} L(C) &= \sum p(u_j) l(u_j) \\ &= \sum_{j \leq k-2} p(u'_j) l(u'_j) + (p(u_{k-1}) + p(u_k))(l(u'_{k-1}) + 1) \end{aligned}$$

Or $p(u_{k-1}) + p(u_k) = p(u'_{k-1})$, donc :

$$L(C) = L(C') + p(u'_{k-1}).$$

Comme $p(u'_{k-1})$ ne dépend pas de C', on peut minimiser $L(C)$ sur la classe des codes où $C(u_k)$ et $C(u_{k-1})$ ne diffèrent que sur le dernier digit en minimiasant $L(C')$. Par le lemme 2.3.1, un tel code minimise $L(C)$ sur tous les codes instantanés. □

Application On construit donc l'arbre de codage de Huffman de proche en proche en rassemblant à chaque étape les deux noeuds de plus faible probabilité et en affectant la somme de ces probabilités au noeud père.

Voici un exemple :

mot code	message	$p(u_k)$
00	u_1	0.3
01	u_2	0.25
10	u_3	0.25
110	u_4	0.1
111	u_5	0.1

2.3.2 Extension au cas $D > 2$

On définit un arbre de codage *complet* comme un arbre de codage pour lequel tous les noeuds intermédiaires ont D enfants.

Lemme 2.3.3 *Le nombre de feuilles dans un arbre de codage complet est de la forme $D + m(D-1)$ pour un certain entier m.*

Démonstration. Le plus petit arbre complet a D feuilles, le second plus petit en a D-1+D (on remplace une feuille de l'arbre précédent par un noeud à D enfants), d'où le résultat par récurrence. □

Pour un code instantané, nous complétons son arbre de codage en rajoutant B feuilles (non utilisées par le code).

Pour un code optimal, toutes les feuilles non utilisées doivent être au même niveau que le mot code le plus long, et ne diffèrent que par le dernier digit.

Un code optimal doit donc avoir au plus $D-2$ feuilles inutilisées.

Si K le nombre de mots code et B le nombre de feuilles inutilisées, on doit avoir :

$$B + K \quad = \quad m(D-1) + D \text{ et } B \leq D-2,$$

donc $K - 2 = m(D-1) + (D-2-B)$ et $0 \leq D-2-B \leq D-2$

ainsi, $B = D - 2 - ((K-2) mod (D-1))$.

En suivant le Lemme 2.3.1, un code optimal existe pour lequel les B feuilles inutilisées et les $D-B$ mots code les moins probables diffèrent par le dernier digit. Donc la première étape consiste à grouper les $D-B$ noeuds les moins probables. Ensuite à chaque itération, l'ensemble réduit est de cardinal $D + m(D-1)$ et on regroupe les D noeuds les moins probables.

EXEMPLE 2.3.1: Pour $D = 3$ et $K = 6$, il faut rajouté 1 feuille inutilisée et on obtient dans cet exemple :

mot code	message	$p(u_k)$
0	u_1	0.4
1	u_2	0.3
20	u_3	0.2
21	u_4	0.05
220	u_5	0.03
221	u_6	0.02