Feuille d'exercices 3

Théorie de Shannon

Pour tous les exercices, on note $\mathcal{S}$ désigne un système cryptographique $\mathcal{S} = (\mathcal{M}, \mathcal{C}, \mathcal{K}, e, d)$. On suppose des distributions de probabilité données pour $\mathcal{M}$ et $\mathcal{K}$, et une distribution induite sur $\mathcal{C}$; on note M (pour les messages clairs), C (pour les messages chiffrés), K (pour les clefs) les variables aléatoires correspondantes, de support ($\mathcal{M}$, $\mathcal{C}$, $\mathcal{K}$). On suppose M et K indépendantes, et M, C et K à valeurs strictement positives (sinon il suffit d'éliminer de $\mathcal{M}$, $\mathcal{C}$ et $\mathcal{K}$ les items inutiles). On déduit C de M et K :

$$\begin{aligned} \mathrm{P}(C = c|M = m) &= \sum_{\{k \ / \ e_k(m)=c\}} \mathrm{P}(K = k) \\ \mathrm{P}(C = c) &= \sum_{\{(m,k) \ / \ e_k(m)=c\}} \mathrm{P}(K = k)P(M = m) \end{aligned}$$

On dit que $\mathcal{S}$ est à *confidentialité parfaite* quand M et C sont indépendantes (intuitivement la connaissance de C n'apprend rien sur M) : pour tout $m \in \mathcal{M}$, $P(M = m|C = c) = P(M = m)$.

Exercice 1. Le DES est un chiffrement par bloc de taille 64 bits et qui utilise des clefs de 56 bits. La transmission d'un bloc se fait-elle à confidentialité parfaite ? Plus généralement, montrer que, si $\mathcal{S}$ est à confidentialité parfaite, alors : $|\mathcal{M}| \leq |\mathcal{C}| \leq |\mathcal{K}|$.

Exercice 2. Un système cryptographique $\mathcal{S} = (\mathcal{M}, \mathcal{C}, \mathcal{K}, e, d)$ est à confidentialité par paires si, pour toute paire de messages clairs m et m' et tout cryptogramme c, la probabilité que c soit le chiffré de m égale celle que ce soit le chiffré de m'.

1. Écrire la condition précisément en terme de probabilité sur les clefs, et montrer qu'elle équivaut à :

$$\forall m \in \mathcal{M} \ \forall m' \in \mathcal{M} \ \forall c \in \mathcal{C} \ \mathrm{P}(C = c|M = m) = \mathrm{P}(C = c|M = m') \ .$$

2. Montrer que si $\mathcal{S}$ est à confidentialité parfaite, alors $\mathcal{S}$ est à confidentialité par paire.

3. Réciproquement, montrer que si $\mathcal{S}$ est à confidentialité par paires alors $\mathcal{S}$ à confidentialité parfaite.

Dans la suite on note « log » la fonction logarithme en base 2.

Exercice 3 (Préliminaires : entropie et entropie conditionnelle). Dans tout l'exercice, X et Y sont des variables aléatoires discrètes à support fini $\{x_1, \ldots, x_n\}$ et $\{y_1, \ldots, y_m\}$. On note $\mathrm{P}(X = x_i) = p_i$, $\mathrm{P}(Y = x_i) = q_i$, et on suppose $p_i > 0$ et $q_i > 0$.
On note $\mathrm{H}(X)$ l'entropie de la variable aléatoire discrète X de support fini $\{x_1, \ldots, x_n\}$; en posant $\mathrm{P}(X = x_i) = p_i$:

$$\mathrm{H}(X) = -\sum_{i=1}^{n} p_i \log p_i = \sum_{i=1}^{n} p_i \log \frac{1}{p_i}$$

On note $X|y$ la variable aléatoire de même support que X, qui a pour distribution de probabilités $\mathrm{P}(X = x|Y = y)$. L'entropie conditionnelle de X sachant Y est la moyenne pondérée par les $P(Y = y)$ des $\mathrm{H}(X|y)$:

$$\begin{aligned} \mathrm{H}(X|Y) &= -\sum_y \sum_x \mathrm{P}(Y = y)\mathrm{P}(X = x|Y = y) \log \mathrm{P}(X = x|Y = y) \\ &= -\sum_y \sum_x \mathrm{P}(X = x, Y = y) \log \mathrm{P}(X = x|Y = y) \end{aligned}$$

1. Montrer que $\mathrm{H}(X) \leq \log n$ et qu'il y a égalité si et seulement si X est équidistribuée.

2. Montrer que :

$$\sup(\mathrm{H}(X), \mathrm{H}(Y)) \leq \mathrm{H}(X, Y) \leq \mathrm{H}(X) + \mathrm{H}(Y).$$

et que $\mathrm{H}(X, Y) = \mathrm{H}(X) + \mathrm{H}(Y)$ si et seulement si X et Y sont indépendantes.

3. Montrer que : $H(X, Y) = H(Y) + H(X|Y)$.

4. Montrer que $H(X|Y) \leq H(X)$ et que l'égalité est réalisée si et seulement si X et Y sont indépendantes.

Exercice 4.

1. Calculer l'entropie d'une distribution équiprobable sur les 26 lettres de l'alphabet.

2. Écrire un programme qui calcule les fréquences des 26 lettres de l'alphabet dans un texte et l'entropie de cette distribution (voir la feuille de rappels de C pour la fonction log).

Exercice 5.

1. Montrer que $\mathrm{H}(C|K,M) = \mathrm{H}(M|K,C) = 0$, et que $\mathrm{H}(M,K,C) = \mathrm{H}(M,K) = \mathrm{H}(K,C) = \mathrm{H}(M) + \mathrm{H}(K)$.
2. Montrer que $\mathrm{H}(K|C) = \mathrm{H}(K) + \mathrm{H}(M) - \mathrm{H}(C)$. l'entropie conditionnelle $\mathrm{H}(K|C)$ est une mesure de l'incertitude qui reste sur la clef quand le chiffré est connu.
3. En déduire que $\mathrm{H}(M|C) \leq \mathrm{H}(K)$.
4. Montrer que $\mathcal{S}$ est à confidentialité parfaite si et seulement si $\mathrm{H}(M|C) = \mathrm{H}(M)$.
5. Montrer que si $\mathcal{S}$ est à confidentialité parfaite alors $\mathrm{H}(K) \geq \mathrm{H}(M)$. Interpréter ce résultat en terme d'information.

Distance d'unicité

On s'intéresse aux systèmes $(\mathcal{M}^n, \mathcal{C}^n, \mathcal{K}, e, d)$ (on étend e et d à $\mathcal{M}^n$ et $\mathcal{C}^n$). On note M^n et C^n des variables aléatoires de domaine $\mathcal{M}^n$, en prenant pour M^n la répartition des mots de n lettres dans la langue étudiée.
L'*entropie d'une langue* L (sur l'alphabet $\mathcal{M}$) est par définition :

$$H_L = \lim_{n\to\infty} \frac{\mathrm{H}(M^n)}{n} \ .$$

Concrètement la suite $\mathrm{H}(M^n)$ devient assez rapidement stationnaire, et on calcule ainsi H_L, estimée à au plus 2 pour la langue française (1,25 pour la langue anglaise).

Exercice 6.

1. Calculer l'entropie d'un langage « aléatoire » (toutes les suites de longueur n sur $\mathcal{M}$ équiprobables).
2. Montrer que quelque soit la variable aléatoire X de domaine $\mathcal{X}$, $\mathrm{H}(X^n) \leq n \log |\mathcal{X}|$.
3. Montrer que, si n est suffisament grand : $\mathrm{H}(K|C^n) \geq \mathrm{H}(K) - n(\log|\mathcal{C}| - H_L)$.
4. On fait l'hypothèse supplémentaire que le système de chiffrement est construit de telle façon que les textes chiffrés sont équiprobables (ce qui est souhaitable, mais pas vérifié pour beaucoup de chiffrement « historiques »). Montrer qu'alors l'inégalité précédente est une égalité.

On suppose à partir de maintenant que toutes les clefs sont équiprobables.

Exercice 7 (nombre de clefs possibles). On pose $K_{\overline{c}} = \{k \in \mathcal{K} \ / \ \exists \overline{m} \in \mathcal{M}^n \ P(M^n = \overline{m}) > 0 \text{ et } e_k(\overline{m}) = \overline{c}\}$, l'ensemble des clefs possibles pour le chiffré $\overline{c}$. Le nombre moyen de clefs possibles pour des mots de longueur n est :

$$k_n = \sum_{\overline{c}\in\mathcal{C}^n} P((C^m = \overline{c})|K_{\overline{c}}| \ .$$

1. Montrer que $\mathrm{H}(K|C^n) \leq \log k_n$ (revenir à la définition).
2. En déduire que pour n suffisament grand :

$$\log k_n \geq \log|\mathcal{K}| - n(\log|\mathcal{C}| - H_L) \ .$$

3. Quelles hypothèses faire pour que l'égalité soit réalisée ?

La *distance d'unicité* d'un système de chiffrement pour une langue naturelle donnée est intuitivement le nombre de lettres d'un texte chiffré nécessaire pour que la clef soit déterminée de façon unique, sachant que le clair est dans la langue donnée. On la définit comme le plus petit entier d tel que $k_d = 1$ (une seule clef possible en moyenne pour des mots de longueur d).

Exercice 8.

1. En reprenant les exercices précédant, donner des hypothèses sur le chiffrement suffisantes pour que l'on ait :

$$d = \frac{\log|\mathcal{K}|}{\log|\mathcal{C}| - H_L}.$$

2. Appliquer cette formule, pour $H_L = 2$, $H_L = 1,5$ au chiffrement par substitution (les hypothèses sont elles satisfaites ?), au chiffrement de César (idem, et le résultat semble-t-il pertinent ?)
3. On considère qu'un caractère est codé sur un octet. LE DES utilise des blocs de 64 bits et des clefs de 56 bits. L'AES utilise des blocs de 128 bits et des clefs de 128 bits ou 256 bits. Calculer la distance d'unicité (en nombre de blocs) pour les trois chiffrements, pour $H_L = 2$ et $H_L = 1,5$.