La théorie de l'information

Comment mesurer la sécurité apportée par un système de cryptographie? C'est le mathématicien Claude Shannon qui en 1947 a répondu à cette question en développant la théorie de l'information.

Entropie

On considère une épreuve aléatoire, et une variable aléatoire X associée à cette épreuve. Comment mesurer **l'information moyenne** apportée par la connaissance de X sur l'épreuve aléatoire?

Prenons l'exemple suivant : on lance un dé non pipé et on considère les 3 variables aléatoires suivantes :

X_1 qui vaut 0 si le nombre tiré est pair, 1 s'il est impair.

X_2 qui vaut 0 si le nombre tiré est 1 ou 2, 1 si le nombre tiré est 3 ou 4, 2 si le nombre tiré est 5 ou 6.

X_3 qui vaut le nombre tiré.

Il est intuitivement clair que la connaissance de X_3 renseigne plus sur le déroulement de l'épreuve aléatoire que la connaissance de X_2, qui elle-même renseigne plus que celle de X_1.

La **notion d'entropie** permet de mathématiser cette heuristique :

Définition 1 *Si X est une variable aléatoire discrète, l'entropie de X est définie par :*

$$H(X) = -\sum_x P(X = x) \log(P(X = x))$$

Dans cette définition, la base du logarithme est souvent choisie égale à 2. L'entropie se mesure alors en shannons, ou en bits.

Il nous faut encore mesurer l'incertitude ou le désordre, liée à l'expérience aléatoire. Si $\{a_1, \ldots, a_N\}$ est l'ensemble des issues possibles de l'expérience, l'entropie vaut :

$$H(E) = -\sum_i P(\{a_i\}) \log(P(\{a_i\}))$$

On retrouve ici la définition physique de l'entropie : mesure du désordre d'un système. La variable aléatoire X renseigne totalement sur le déroulement de l'expérience E si $H(X) = H(E)$.

Exemples :

- Dans l'exemple du dé, on vérifie que :

$$H(X_1) = \log 2 < H(X_2) = \log 3 < H(X_3) = \log 6$$

Dans cet exemple, l'incertitude totale liée à l'expérience est $\log 6$.

- Si X est une variable aléatoire équidistribuée qui peut prendre n valeurs, l'entropie de X est :

$$H(X) = -\sum_{k=1}^{n} \frac{1}{n} \log\left(\frac{1}{n}\right) = \log n$$

Il est facile de voir que parmi les variables à n valeurs, l'entropie $H(X)$ est maximale lorsque X est équirépartie : une variable aléatoire apporte en moyenne un maximum d'informations lorsqu'elle peut prendre chaque valeur avec une égale probabilité.

En effet

A cause de la concavité de la fonction $t \mapsto t \log t$ sur l'intervalle [0, 1], l'entropie $H(X)$ est maximale si et seulement si X est équidistribuée.

En effet, si X n'est pas équidistribuée, il existe deux valeurs de x telles que

$$t_1 = P(X = x_1) \neq P(X = x_2) = t_2.$$

Notons $f(t) = t \log t$,

$$f(t_1) + f(t_2) > 2f(\frac{t_1 + t_2}{2}).$$

Si on remplace X par la variable X' avec $P(X = x_1) = P(X = x_2) = \frac{t_1+t_2}{2}$, on obtient une entropie plus grande: $H(X') > H(X)$.

Pour la cryptographie, la quantité importante va être **l'entropie conditionnelle** de X sachant Y. Elle est définie par :

$$H(X|Y) = -\sum_{x,y} P(X = x, Y = y) \log(P(X = x|Y = y))$$

où la somme est étendue à toutes les valeurs x et y prises par X et Y.

$H(X|Y)$ représente **l'incertitude qu'il reste sur X lorsque l'on connait Y**.

Elle verifie la propriété:

$$H(X, Y) = H(Y) + H(X|Y)$$

qui signifie que l'information apportée par les 2 variables X et Y vaut

l'information apportée par Y seule
plus
l'information apportée par X connaissant déjà la valeur de Y.

Proposition 1 *On a:*

$$H(X,Y) = H(Y) + H(X|Y)$$

où $H(X,Y)$ *désigne l'entropie de la variable aléatoire* (X,Y)*.*

Démonstration: Il suffit de se rappeler que

$$P(X=x, Y=y) = P(X=x|Y=y)P(Y=y).$$

On a alors:

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

La quantité H a un certain nombre de **propriétés intéressantes** qui justifient plus loin qu'elle soit prise comme **mesure raisonnable de choix ou d'information.**

1. $H = 0$ si et seulement si tous les p_i sauf un sont nuls, celui-ci ayant la valeur 1. Ainsi seulement quand nous sommes certains des résultats, $H = 0$.

Autrement H est positif.

2. Pour n donné, H est un maximum et une égale à $\log n$ quand tous les p_i sont égaux (c.-à-d., $1/n$). C'est intuitivement la situation la plus incertaine.

3. Supposons qu'il y ait deux événements, x et y en question avec m possibilités pour la première et le n pour la seconde. Soit $p(i,j)$ la probabilité de l'occurrence commune de i pour le premier et de j pour la seconde. L'entropie de l'événement commun est

$$H(x,y) = -\sum p(i,j)\log p(i,j)$$

tandis que

$$H(x) = -\sum_i (\sum_j p(i,j))\log \sum_j p(i,j)$$

$$H(y) = -\sum_j (\sum_i p(i,j))\log \sum_i p(i,j)$$

il est facilement montré que

$$H(x,y) \leq H(x) + H(y)$$

avec **l'égalité seulement si les événements sont indépendants** (c.-à-d., $p(i,j) = p(i)p(j)$). L'incertitude d'un événement commun est inférieur ou égal à la somme des différentes incertitudes.

4. N'importe quel changement vers l'égalisation des probabilités $p_1, p_2, \ldots, p_n$ augmente H. Ainsi si $p_1 < p_2$ et si nous augmentons p_1, décroissions p_2 du même montant de telle sorte que p_1 et p_2 soient plus voisins, alors H augmente.

5. L'incertitude (ou l'entropie) de l'événement commun x, y est l'incertitude de x plus l'incertitude de y quand x est connu.

$$H(x, y) = H(x) + H(y|x)$$

6. D'après 3 et 5 nous avons

$$H(x) + H(y) \geq H(x, y) = H(x) + H(y|x)$$

Donc

$$H(y) \geq H(y|x)$$

L'incertitude de y n'est jamais augmentée par la connaissance de x. Elle sera diminuée à moins que x et y soient des événements indépendants, dans ce cas elle n'est pas changée.

Application à la cryptographie

Shannon a modélisé les systèmes cryptographiques de la façon suivante :

– le message clair m est défini comme une variable aléatoire à valeurs dans l'ensemble des messages M possibles;

– le message chiffré c est défini comme une variable aléatoire à valeurs dans l'ensemble C des messages chiffrés.

– m, k et c sont reliés par la relation

$$c = e_k(m),$$

où e est la fonction de chiffrement.

L'incertitude liée au système est l'entropie $H(m)$.

Un cryptanalyste doit donc obtenir de l'ordre de $H(m)$ shannons d'information pour retrouver m. A priori, il ne peut connaitre que le système cryptographique utilisé, et c.

Proposition 2 *Pour qu'un système cryptographique soit parfaitement sûr il faut et il suffit que*

$$H(m|c) = H(m)$$

où m *et* c *sont des variables aléatoires.*

Démonstration

$$H(m|c) = H(m)$$
$$\Updownarrow$$
$$H(m|c) + H(c) = H(c) + H(m)$$
$$\Updownarrow$$
$$H(m,c) = H(c) + H(m)$$
$$\Updownarrow$$
m et c sont indépendantes
$$\Updownarrow$$
$$P(m|c) = P(m)$$

Proposition 3 *L'équivocation du message et de la clé sont liés par*

$$H(k|c) = H(m|c) + H(k|m,c)$$

Démonstration

Par la propriété 5 de l'entropie, on a

$$\begin{aligned} H(m|c) &= H(m,c) - H(c) \\ &= H(m,k,c) - H(k|m,c) - H(c) \end{aligned}$$

$$\begin{aligned} H(k|c) &= H(k,c) - H(c) \\ &= H(m,k,c) - H(m|k,c) - H(c) \end{aligned}$$

Maintenant, puisqu'un système de cryptographie est réversible, la connaissance de c et k détermine complètement m. Donc: $H(m|(c,k)) = 0$.

Par conséquent

$$\begin{aligned} H(k|c) &= H(m,k,c) - H(c) \\ &= H(m|c) + H(k|m,c) \end{aligned}$$

Corollaire 1 *On a* $H(k|c) \geq H(m|c)$

Théorème 1 *Si un système cryptographique est parfait, alors :*

$$H(k) \geq H(m)$$

Démonstration

$$H(k) \geq H(k|c) \geq H(m|c) = H(m)$$

En d'autres termes, l'information contenue dans la clé est au moins aussi grande que l'information contenue dans le message en clair. Si t est la taille de la clé, on a $H(K) \leq t$.

Il ne sert à rien de vouloir inventer un système de cryptographie à clé réduite aussi sûr que le chiffre de Vernam. La théorie de Shannon prouve que cela est impossible.

Distance d'unicité

Shannon a aussi considéré le cas des **systèmes imparfaits,** à clé courte, en cherchant a déterminer la **quantité d'information nécessaire** au cryptanalyste pour **retrouvrer la clé à partir du texte chiffré.**

Prenons le cas, par exemple, d'un système de chiffrement par substitution sur l'alphabet latin. On a

$$M = C = A = \{A, B, \ldots, Z\}$$

et l'ensemble K des clés est l'ensemble des permutations de 26 éléments. Il est clair que si le cryptanalyste intercepte un texte chiffré trop court, par exemple d'une seule lettre, il ne saura en déduire la clé. Combien de lettres lui faut-il intercepter en moyenne pour qu'il puisse en déduire la clé ?

Plus généralement, supposons qu'une même clé K, de longueur fixe, soit utilisée pour chiffrer un texte constitué de n messages $M_1, \ldots, M_n$, auxquels sont associés les cryptogrammes $C_1, \ldots, C_n$. On appelle **distance d'unicité d** le plus petit entier n tel que

$$H(K|(C_1, \ldots, C_n)) = 0.$$

Il s'agit du plus petit nombre moyen de cryptogrammes $C_1, \ldots, C_n$ tel que, connaissant $C_1, \ldots, C_n$, il n'y ait plus aucune incertitude résiduelle sur la clé. On a:

$$\begin{aligned} H(K|(C_1, \ldots, C_n)) &= H(K, C_1, \ldots, C_n) - H(C_1, \ldots, C_n) \\ &= H(M_i, \ldots, M_n, K, C_1, \ldots, C_n) - H(C_1, \ldots, C_n) \\ &= H(M_i \ldots, M, K) - H(C_1, \ldots, C_n) \\ &= H(M_i, \ldots, M_n) + H(K) - H(C_1, \ldots, C_n) \end{aligned}$$

la dernière égalité provenant de ce que les M_i sont indépendants de K.

On vient donc de montrer que :

$$H(M_i, \ldots, M_d) + H(K) - H(C_1, \ldots, C_d) = 0.$$

Comment utiliser cette égalité pour évaluer d?

On pourra faire l'hypothèse que $H(C_1, \ldots, C_d) = d \log(\#C)$.

Cela signifie que tous les cryptogrammes possibles sont équiprobables, ce qui est clairement une caractéristique souhaitable d'un système cryptographique.

Posons

$$H = \frac{1}{d} H(M_1, \ldots, M_d).$$

Si l'on peut considérer les messages M_i comme indépendants, alors on a $H = H(M_i)$.

Si les messages M_i sont les lettres d'un texte écrit dans un langue naturelle, l'hypothèse d'indépendance est inexacte. Mais si une lettre dépend fortement des lettres voisines elle dépend en général très peu des lettres plus éloignées, autrement dit la suite

$$H(M_1),\ H(M_1, M2)/2,\ H(M_1, M_2, M_3)/3 \ldots$$

est **rapidement stationnaire**. On pourra donc, par exemple, prendre un échantillon de texte, calculer la fréquence p_L d'apparition de chaque suite L de i lettres et poser

$$H = -\frac{1}{i} \sum_{L \in A^i} p_L \log p_L.$$

Dans ces conditions, l'égalité

$$H(M_i, \ldots, M_d) + H(K) - H(C_1, \ldots, C_d) = 0.$$

se réécrit

$$d\,H + H(K) - d\log(\#C) = 0$$

c'est-à-dire

$$d = \frac{H(K)}{\log(\#C) - H}$$

Reprenons notre exemple de chiffrement par substitution sur l'alphabet latin. Dans ce cas on a $H(K) = \log(26!)$, et $\log(\#C) = \log(\#A) = \log 26$. Pour des textes en clair écrits en anglais ou en français une analyse fine des fréquences en prenant i de l'ordre de 8 permet d'estimer l'entropie par lettre à $H \simeq 2$ bits. On en déduit :

$$d \simeq 30.$$

Cela veut dire que l'on doit savoir retrouver la clé dès que le texte chiffré dépasse une trentaine de lettres. l'expérience confirme ce calcul assez précisément.

On peut retenir de cette analyse que, si l'on chiffre des messages qui contiennent une certaine redondance, comme celle des langues naturelles, et que l'on utilise des clés de longueur fixée, alors il est inévitable, quel que soit le système de chiffrement utilisé, qu'avec suffisamment de cryptogrammes interceptés, le cryptanalyste aura à sa disposition assez d'information pour retrouver la clé. Cela ne préjuge cependant pas de l'effort de calcul qui lui sera nécessaire.

Du point de vue des utilisateurs d'un système de chiffrement, on peut également retenir qu'il est souhaitable, pour augmenter la distance d'unicité, de réduire la redondance des messages en clair avant de les chiffrer, par exemple par un algorithme de compression.