Codage de l ’information (1)

- Codage de l ’information ≡ représentation codée de l ’information

- **Rôles multiples du codage**
  - préparation de la transformation message => signal transmis
  - adaptation débit source - capacité du canal (compression)
  - codage de protection contre les erreurs de transmission (détection / correction d’erreurs)
  - chiffrage (secret des communications)
  - tatouage (marquage du titre de propriété)
  - transcodage (changement d ’alphabet: contraintes de transmission)

Un système de communication a pour but d’acheminer des messages d’un expéditeur (source d’information) vers un destinataire (utilisateur de l’information). Le signal supportant l’information à transmettre doit être compatible avec les caractéristiques du canal de transmission. Le **codage de l’information** doit établir une correspondance injective entre le message transmis par la source et la suite des symboles d’informations $\{b_n\}$ transmis à l’émetteur.

Ce codage a des rôles multiples. Il prépare la transformation du message en un signal (effectué dans la partie de l’émetteur : codage information à signal). Il adapte la source d’information à la capacité de transmission du canal. Il permet de protéger l’information contre les erreurs de transmission (dans une certaine limite) en vue de leur détection ou/et de leur correction (dues aux perturbations dans le canal). Il permet également dans certains cas d’assurer le secret des communications (le chiffrage), et de tatouer pour marquer une propriété.

Un canal de transmission donné doit pouvoir transporter des messages de nature variée d’où la nécessité d’un transcodage permettant de transformer la représentation des messages à partir d’un dictionnaire M_k utilisant l’alphabet A_k en une représentation des mêmes messages à partir d’un dictionnaire M_0 utilisant l’alphabet A_0. Cet alphabet particulier est souvent l’ensemble binaire {0, 1}, mais ce n’est pas le seul.

Codage de l ’information (2)

- Définitions
  - Sources S de messages: production d’une suite de messages, chacun d’eux étant sélectionné dans un ensemble M des messages

    (M : dictionnaire de messages possibles M = { m_1 , m_2 , ….}, les m_i sont aussi appelés « mots »)
  - Message : suite finie de symboles

    (caractères pris dans A : alphabet)
  - Alphabet : ensemble fini de symboles A = { a_1, a_2, ……, a_k }

♦ *Définitions* :

On appelle **message** toute suite finie de caractères pris dans un **alphabet** A : ensemble fini de **symboles** (par exemple : lettres, chiffres, …).

On appelle **source de messages** S, le système qui produit une suite temporelle de messages m_i, chacun d’eux étant pris dans l’ensemble M des messages possibles. M est appelé **dictionnaire de messages** (ou mots). Le message transmis est en fait un texte formé par des règles syntaxiques de messages élémentaires appelés mots : M = {m_1, m_2, … }, chaque mot s’écrit par une suite finie fixée de symboles pris dans un alphabet A.

En fonction des applications envisagées, les sources de message S peuvent utiliser des dictionnaires de nature très différentes : depuis les messages écrits utilisant les caractères de l’alphabet, les chiffres, les caractères de ponctuation et de tabulation jusqu’aux messages visuels où les messages sont des images numérisées où, par exemple, chaque mot est un pixel s’écrivant comme la suite de 8 symboles (binaires) pris sur l’alphabet {0, 1} (octet).

Entropie d 'une source (*SHANNON 1948*)

- Définition de l '***incertitude*** et de l '***entropie***
  - **Incertitude** I d 'un événement E :

    $I(E) = -\log_2 Pr\{E\}$ *Unités*: bit (BInary uniT si $\log_2$)

    nat (NAtural uniT si Log_e): 1 nat=1,443 bit

    si source simple s_n => $I(s_n) = \sum_{i=1;n} I(m_{\alpha i})$
  - **Entropie** H d 'une variable aléatoire discrète X :

    $H(X) = E_X [I(X)] = \sum_{i=1;n} p_i I(X_i) = - \sum_{i=1;n} p_i \log_2(p_i)$
  - Propriétés de l 'entropie
    - $H \geq 0$; H est continue , symétrique ; $H(p_1, \ldots, p_N) \leq \log_2 n$
    - si $(p_1, \ldots, p_n)$ et $(q_1, \ldots q_n)$ sont 2 répartitions de probabilités

      $\Longrightarrow \sum_{i=1;n} p_i \log_2(q_i / p_i) \leq 0$ car $\log x < x - 1$

L'***incertitude*** I d'un événement E de probabilité Pr(E) est définie par :

$$I(E) = \log_2 \frac{1}{Pr(E)} = - \log_2 Pr(E)$$

- Remarques :

- si Pr(E) = 1/2 alors I (E) = 1 (incertitude unité)
- si Pr(E) = 1 alors I (E) = 0 : l'incertitude est nulle pour un événement certain.
- L'unité est le ***bit*** (*Binary Unit*) à ne pas confondre avec le bit : *Binary digit*.
- On peut utiliser la fonction logarithme népérien à la place du logarithme en base 2, dans ce cas l'unité est le ***nat*** (Natural Unit = 1,443 bit).

On considère maintenant que les événements E sont en fait des réalisations d'une variable aléatoire discrète X. On définit alors l'***entropie*** H comme étant l'incertitude moyenne de la variable aléatoire X. Si on considère en fait chaque événement x_i, $i \in [1, n]$, comme une réalisation d'une variable aléatoire X (i.e. X est une variable aléatoire à valeurs dans $\{ x_1, x_2, \ldots, x_n \}$) :

$$H(X) = E_X \{ I(X) \} = \sum_{i=1..n} \Pr\{X = x_i\} \, . \, I(x_i) = \sum_{i=1..n} p_i . I(x_i), \text{ avec } p_i = \Pr\{X = x_i\}$$

L'entropie dépend de la loi de probabilité de X mais n'est pas fonction des valeurs prises par X. Elle s'exprime en bits (ou nats) et représente le nombre moyen de bits nécessaires à coder en binaire les différentes réalisations de X.

On considère maintenant une source d'informations S définie par un ensemble de message m_i possibles (dictionnaire) : $S\{m_1, m_2, \ldots, m_N\}$, et par un mécanisme d'émission de messages tel que :

$s_n = \{m_{\alpha 1}, m_{\alpha 2}, \ldots, m_{\alpha n}\}$ avec $m_{\alpha 1}$: 1^{er} message émis, …, $m_{\alpha n}$: $n^{ième}$ message émis.

Attention : l'indice « i » dans αi définit l'indice temporel dans la suite de messages émis par la source. αi définit l'indice du $i^{ème}$ message émis dans le dictionnaire M des messages possibles, en général : $N \neq n$.

Le choix des $m_{\alpha i}$ s'effectue selon une loi de probabilité donnée. L'émission d'une source discrète d'informations correspond donc à une suite de variables aléatoires X_i, $i \in [1, n]$: une variable aléatoire associée à la réalisation de chaque message émis $m_{\alpha i}$ de la source.

La probabilité de s_n s'exprime alors comme un produit de probabilités conditionnelles :

$$\Pr(s_n) = \Pr\{X_1 = m_{\alpha 1}\} \Pr\{X_2 = m_{\alpha 2} / X_1 = m_{\alpha 1}\} \ldots \Pr\{ X_n = m_{\alpha n} / X_1 = m_{\alpha 1}, \ldots, X_{n-1} = m_{\alpha n-1}\}$$

Dans le cas de sources simples, les n variables aléatoires X_i sont indépendantes et de même loi donc :

$$\forall (i, j) \in [1, n] \times [1, N], \Pr\{ X_i = m_j \} = p_j, \text{ et } \Pr\{s_n\} = p_{\alpha 1} . p_{\alpha 2} \ldots p_{\alpha n}$$

$$\Rightarrow I (s_n) = - \log_2 \Pr\{s_n\} = - \log_2 (p_{\alpha 1} . p_{\alpha 2} \ldots p_{\alpha n}) = \sum_{i=1..n} - \log_2 p_{\alpha i} = \sum_{i=1..n} I (m_{\alpha 1})$$

$$\boxed{I (s_n) = \sum_{i=1..n} I (m_{\alpha 1})}$$

Dans le cas d'une source discrète de « n » messages m_i, ou chaque message m_i est associé à la probabilité p_i, l'entropie H de la source S est donnée par :

$$\boxed{H(S) = - \sum_{i=1}^{n} p_i \, . \log_2 p_i}$$

♦ *Propriétés de l'entropie :*

- Comme $0 \leq p_i \leq 1$ et que $\sum_{i=1}^{n} p_i = 1$, alors $H(X) > 0$: l'entropie est positive.
- Soient $(p_1, p_2, \ldots, p_n)$ et $(q_1, q_2, \ldots, q_n)$ deux lois de probabilité, alors $\sum_{i=1}^{n} p_i \log_2 \frac{q_i}{p_i} \leq 0$.

  En effet, $\forall x > 0$, on a $\ln x \leq x - 1$. D'où $\ln \frac{q_i}{p_i} \leq \frac{q_i}{p_i} - 1$, soit $\log_2 \frac{q_i}{p_i} \leq \frac{1}{\ln 2}(\frac{q_i}{p_i} - 1)$

  donc $\sum_{i=1}^{n} p_i \log_2 \frac{q_i}{p_i} \leq \frac{1}{\ln 2} \sum_{i=1}^{n} p_i \left(\frac{q_i}{p_i} - 1 \right) = \frac{1}{\ln 2} \left(\sum_{i=1}^{n} q_i - \sum_{i=1}^{n} p_i \right) = \frac{1}{\ln 2} (1 - 1) = 0$.
- L'entropie d'une variable aléatoire X à n valeurs possibles est maximale et vaut $\log_2 n$ lorsque la loi de X est uniforme. En prenant $q_1 = q_2 = \ldots = q_n = \frac{1}{n}$ (loi uniforme), dans la propriété précédente :

$$\sum_{i=1}^{n} p_i \log_2 \frac{q_i}{p_i} \leq 0 \quad \Leftrightarrow -\sum_{i=1}^{n} p_i \log_2 p_i \leq -\sum_{i=1}^{n} p_i \log_2 q_i$$

$$\Leftrightarrow H(X) \leq -\sum_{i=1}^{n} p_i \log_2 \frac{1}{n}$$

$$\Leftrightarrow H(X) \leq -\log_2 \frac{1}{n} \sum_{i=1}^{n} p_i = \log_2 n$$

- L'entropie est continue et symétrique.

Dans la suite de ce cours, nous utiliserons systématiquement la base 2 du logarithme.

Exemple simple :

On considère une source S, à loi uniforme, qui envoie des messages à partir de l'alphabet français à 26 caractères (a,b,c, …, z). On ajoute à cet alphabet le caractère « espace » comme séparateur des mots.

L'alphabet est donc constitué de 27 caractères : $H(S) = -\sum_{i=1}^{27} \frac{1}{27} \log_2 \frac{1}{27} = \log_2(27) = 4{,}75$ bits d'information par caractère. En réalité, l'entropie est voisine de 4 bits d'information par caractère sur un très grand nombre de textes français.

Codage de l 'information (3)

- **Objectifs**
  - transcription des informations pour faciliter le codage
    code $\Rightarrow$ signal *(Transcodage)*
  - Compression d 'information
    réduction de la quantité d 'information
  - Protection contre les erreurs de transmission
    contre les dégradations et les erreurs de décision
  - Assurer le secret des informations transmises
    Chiffrement
- **Définition d 'un code**
  application de S dans $\mathcal{A} = \{ a_1, a_2, \ldots\ldots, a_q \}$
  message $m_i \in S \quad \Rightarrow \quad$ mot-code $M_i \in \mathcal{M}$ suites finies de $\mathcal{A}$

Le codage des informations consiste à transcrire les messages d'une source d'information sous la forme d'une suite de caractères pris dans un alphabet prédéfini. Les objectifs du codage sont principalement de quatre ordres :

- de ***transcrire*** les informations sous une forme permettant d'élaborer assez facilement le signal qui supportera les informations, ou de manipuler aisément ces informations automatiquement. Pour cela, différents codes de représentation de l'information sont utilisés en fonction des différentes applications envisagées et on utilise assez souvent des opérations de transcodage ;

- de ***réduire*** la quantité des symboles d'information (en terme de nombre total de symboles utilisés) nécessaires à la représentation des informations : c'est un rôle économique ;

- de ***lutter*** contre les dégradations (distorsions, bruit) amenées par le canal de transmission et conduisant à des erreurs de reconstruction de l'information en sortie du canal de transmission (en réception) ;

- d'assurer un secret dans la transmission des informations de façon à rendre l'information inintelligible sauf au destinataire de celle-ci.

- *Définition du code :*

Soit un ensemble, nommé alphabet A formé de q caractères a_i : $\mathcal{A}$ = { a_1, a_2, …, a_q }et $\mathcal{M}$ l'ensemble fini des suites finies M_i de caractères (par exemple : $M_i = a_{10}\ a_4\ a_7$).
Soit un ensemble fini de messages provenant d'une source S de messages : S={ m_1, …, m_N }.

On appelle ***code*** toute application de S dans $\mathcal{A}$: codage de S par l'alphabet A.

L'élément M_i de $\mathcal{M}$ qui correspond au message m_i de S est appelé le ***mot-code*** de m_i. Sa ***longueur***, notée n_i, est le nombre de caractères appartenant à $\mathcal{A}$ qui le forment.

Le ***décodage*** d'une suite de messages m_i envoyés implique qu'on soit en mesure de séparer les mots-codes dans la suite reçue de mots-codes M_i d'où, parfois, l'utilisation dans l'alphabet d'un caractère particulier d'espacement.

Codage de l 'information (4)

- Alphabet A = { a_1, a_2, ……, a_q }
- Ensemble fini de messages S = { m_1 , m_2 , …., m_i ,…… , m_N }

Codage ↓

C = { M_1 , M_2 , …., M_i ,…….. , M_N }

- Longueur des mots-codes : $n_i = n(M_i)$
- Longueur moyenne des mots-codes : $E(n) = \sum_{i=1;N} p_i n_i$
- Entropie de la source H : $H(p_1, \dots, p_N) \leq \log_2 N$
- Quantité moyenne d 'information par caractère = H / E(n)

  or $H / E(n) \leq \log_2 q \Rightarrow E(n) \geq H / \log_2 q$
- Débit d 'une source d'informations codée en moyenne avec D caractères par seconde : R = D H/E(n)

  => **$R \leq D \log_2 q$** *R en bit/seconde*

On appelle dorénavant **m_i** les **messages** produits par la source d'information et **M_i** les **mots-codes** associés à ceux-ci.
On appelle **n_i** = $n(M_i)$ le nombre de caractères appartenant à **l'alphabet** $\mathcal{A}$ (Card($\mathcal{A}$) = q) nécessaires au codage de m_i, n_i est donc la **longueur** du mot-code M_i. Si la source utilise N messages différents possibles, la longueur moyenne des mots-codes est donnée par :

$$E(n) = \sum_{i=1}^{N} p_i n_i \text{ , où } p_i = Pr\{ m_i \}.$$

H est l'incertitude moyenne (i.e. l'entropie) de la source S **par message** envoyé, donc l'incertitude moyenne (i.e. l'entropie) **par caractère** (de l'alphabet $\mathcal{A}$) est égale à $\frac{H}{E(n)}$ et telle que : $\frac{H}{E(n)} \leq \log_2 q$ (car on a q caractères dans l'alphabet $\mathcal{A}$), donc : $E(n) \geq \frac{H}{\log_2 q}$.
Enfin, si la source codée d'information produit D caractères par seconde pris dans l'alphabet $\mathcal{A}$, $\frac{H}{E(n)}$ étant l'information moyenne transportée par caractère en bit/caractère, le débit R d'information est : $R = D.\frac{H}{E(n)}$. Ce débit est donc limité par : $R \leq D.\log_2 q$.

Codage et décodage de l 'information (5)

- **Efficacité** η d 'un code : $\eta = n_{min} / E(n)$ => $\eta = H / (E(n) \log_2 q)$
- **Redondance** ρ d 'un code : $\rho = 1 - \eta$
- *Exemples simples: codes C_1 et C_2*
- **Contraintes** : séparation des mots-codes & lecture non-ambiguë des mots-codes => *codes réguliers et déchiffrables*
- **Code régulier** : si $m_i \neq m_j \Longrightarrow M_i \neq M_j$ *(application injective)*
- **Code déchiffrable** : 2 suites de messages distincts
  ==> 2 suites de codes distincts
  si $(m_{\alpha 1},\ldots, m_{\alpha i}) \neq (m_{\beta 1},\ldots, m_{\beta j})$ => $(M_{\alpha 1},\ldots, M_{\alpha i}) \neq (M_{\beta 1},\ldots, M_{\beta j})$
  exemples: codes à longueur fixe; codes avec séparateur
- **Code irréductible**: code déchiffrable pouvant être décodé sans artifice M_i n 'est pas préfixe de M_j $\forall$ i , j

Quelques définitions et propriétés liées au codage et au décodage de l'information :

Efficacité :

Pour un alphabet A donné, on appelle efficacité d'un code le rapport η donné par :

$$\eta = \frac{n_{min}}{E(n)} = \frac{\min E(n)}{E(n)} = \frac{\frac{H}{\log_2 q}}{E(n)} = \frac{H}{E(n)\log_2 q}, \eta \in [0, 1]$$

Redondance :

On définit la redondance mathématique par le facteur $\rho = 1 - \eta$. La redondance peut être utilisée pour accroître la robustesse du codage face aux erreurs de transmission de l'information codée (détection et correction des erreurs).

Voici un exemple simple : on considère une source à 4 messages possibles $\{m_1, m_2, m_3, m_4\}$ de probabilités respectives : $p_1 = 0.5$; $p_2 = 0.25$; $p_3 = p_4 = 0.125$.
Soient les deux codes C_1 (codage binaire simple) et C_2 suivants (code à longueur variable) :

Messages / Codes	m_1	m_2	m_3	m_4
C_1	0 0	0 1	1 0	1 1
C_2	0	1 0	1 1 0	1 1 1

Pour C_1 : $\eta = \frac{1.75}{2} = 0.875$ et $\rho = 1 - \eta = 0.125$.

Pour C_2 : $\eta = \frac{1.75}{1.75} = 1$ et $\rho = 1 - \eta = 0$.

Le code C_2 est d'efficacité maximale (unité) alors que le code C_1 ne l'est pas.

Code régulier :

Tout mot-code donné n'est associé qu'à un seul message possible (l'application $S \rightarrow A$ est bijective) : si $m_i \neq m_j$ alors $M_i \neq M_j$.

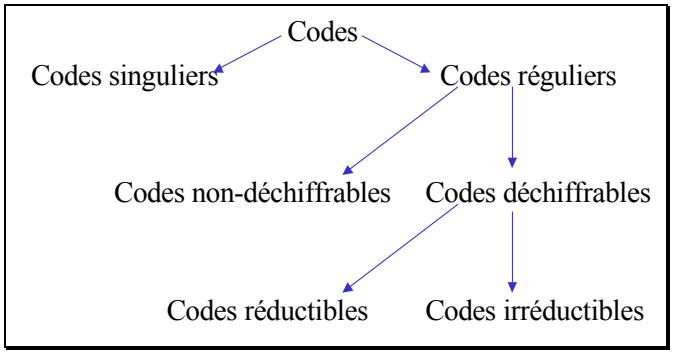
Code déchiffrable :

Un code est déchiffrable si deux textes distincts $(m_{\alpha 1}, \ldots, m_{\alpha i})$ et $(m_{\beta 1}, \ldots, m_{\beta j})$ conduisent nécessairement à des codages distincts (par exemple les codes à longueur fixe tels que C_1 et les codes avec séparateur). Un code déchiffrable est donc un cas particulier de code régulier.

Code irréductible :

C'est un code déchiffrable pouvant être lu directement sans artifices particuliers (code de longueur fixe, séparateur). Pour cela, tout mot-code M_i d'un message m_i ne doit avoir aucun préfixe qui soit un autre mot-code M_j.

On peut ainsi créer une classification hiérarchique pour caractériser le type d'un code :



Exemples de codes

Codes réguliers / Codes déchiffrables / Codes irréductibles

Messages Proba.	m_1 0.5	m_2 0.25	m_3 0.125	m_4 0.125
C_1	1	1	0	00
C_2	0	1	11	01
C_3	1	01	001	000
C_4	1	10	100	1000

- C_1 est non régulier
- C_2 est non-déchiffrable
- C_3 est déchiffrable et irréductible
- C_4 est seulement déchiffrable

Voici quatre codes C_1, C_2, C_3 et C_4 donnés en exemples des définitions et des propriétés précédentes. On suppose que les quatre messages m_1, m_2, m_3, et m_4 sont distincts.

Le code C_1 est non régulier : $m_1 \neq m_2$ mais $C_1(m_1) = C_1(m_2)$, et de même $C_1(m_3) = C_1(m_4)$.

Le code C_2 est non-déchiffrable : les deux textes $\{m_1, m_2\}$ et $\{m_4\}$ sont différents, mais ils conduisent au même code « 01 ».

Le code C_3 est déchiffrable et irréductible : deux textes distincts composés de suites de messages, par exemple $\{m_1, m_3, m_4\}$ et $\{m_1, m_2\}$, conduisent toujours à des codes différents, et aucun mot-code $M_i = C_3(m_i)$ n'est préfixe d'un autre mot-code $M_j = C_3(m_j)$

Le code C_4 est déchiffrable mais non-irréductible : deux textes distincts conduisent toujours à des codes différents, mais les mots-codes $M_i = C_4(m_i)$ sont les préfixes de tous les mots-codes $M_j = C_4(m_j)$ dès que $i < j$.