

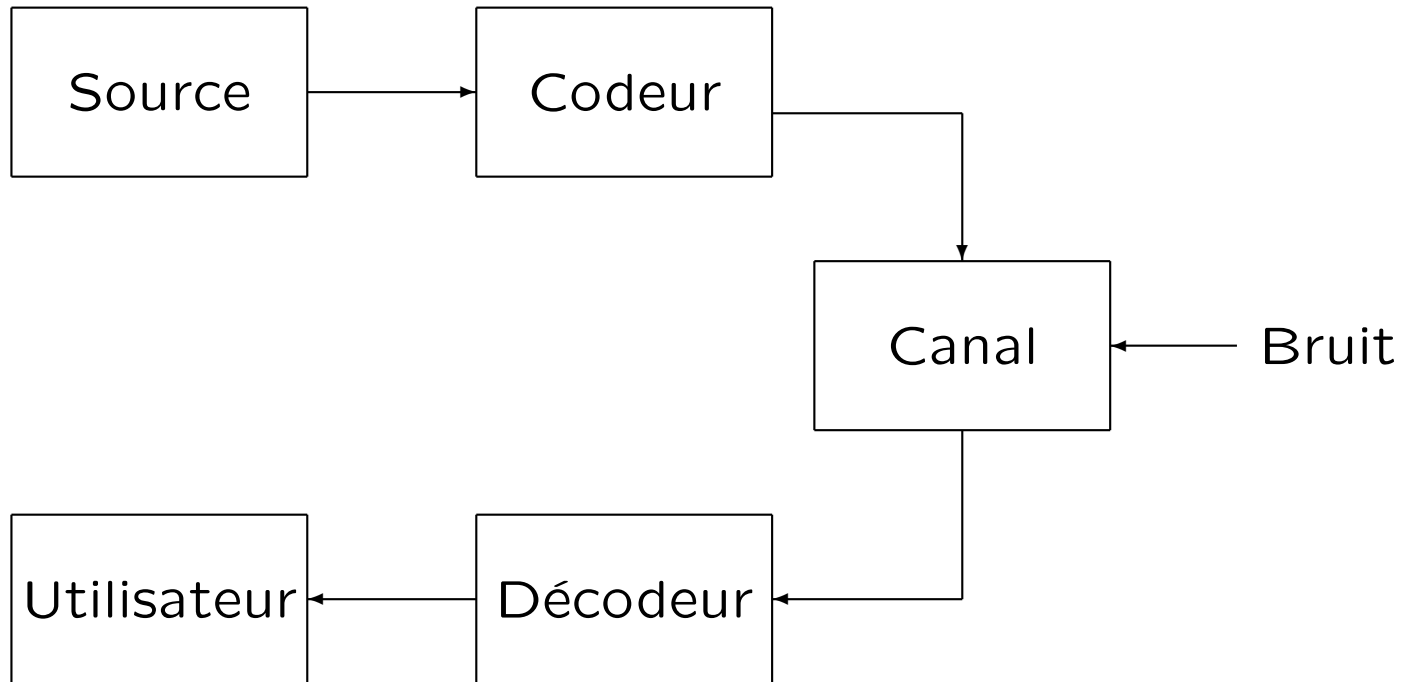
# Introduction à la théorie de l'information

Nicolas Sendrier

Journée TIPE 2009  
ENSTA, 4 novembre 2008

# I. Introduction

## Systeme de communication

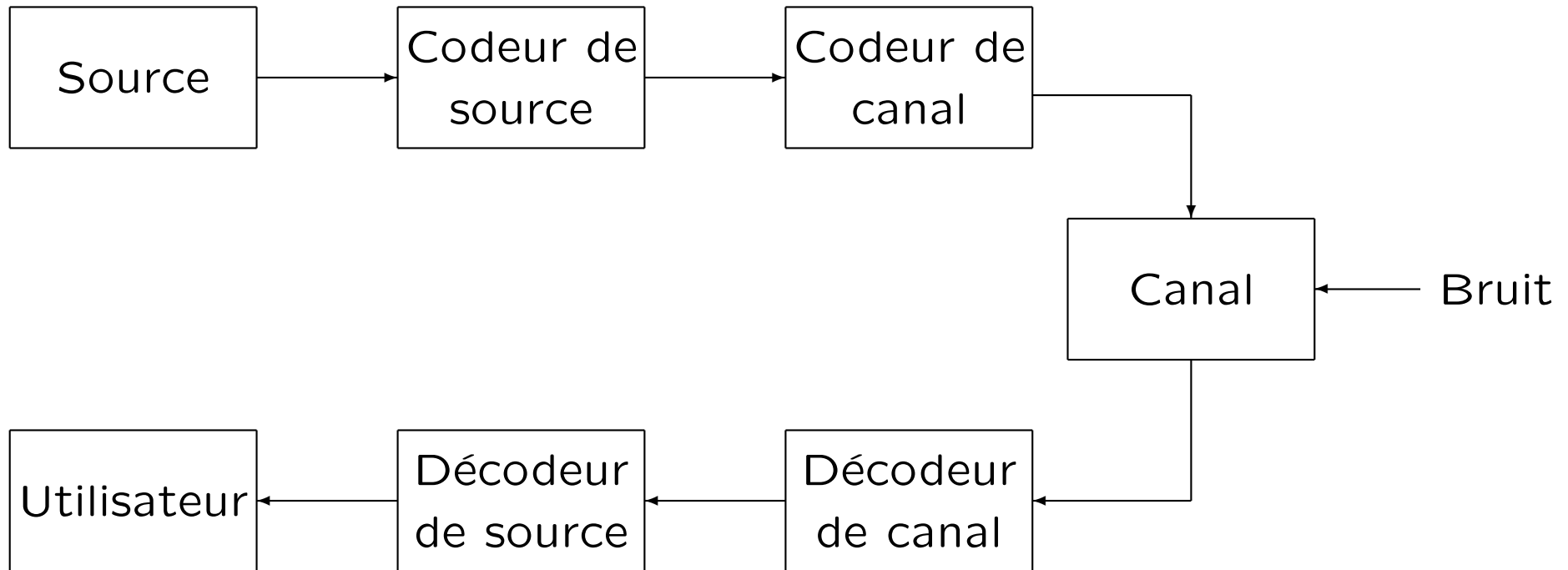


**Source** : voix, musique, image (fixe ou animée), texte, ...

**Canal** : radio, fil, fibre optique, support magnétique ou optique, ...

**Bruit** : perturbations electromagnétiques, rayures, ...

## Codage de source et de canal



**Efficacité** : Pour faire parvenir une quantité donnée d'information à l'utilisateur, utiliser le minimum de ressources.

**Fiabilité** : Restituer à l'utilisateur une information *suffisamment* fidèle à celle produite par la source.

## Exemple de codage de source

$$\mathcal{X} = \{a_1, a_2, a_3, a_4\}, \text{ Loi 1 : } \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right), \text{ Loi 2 : } \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right)$$

$$\text{Code A : } \begin{cases} a_1 \rightarrow 00 \\ a_2 \rightarrow 01 \\ a_3 \rightarrow 10 \\ a_4 \rightarrow 11 \end{cases} \quad \text{Code B : } \begin{cases} a_1 \rightarrow 0 \\ a_2 \rightarrow 10 \\ a_3 \rightarrow 110 \\ a_4 \rightarrow 111 \end{cases}$$

### Longueur moyenne :

- pour le Code A, on trouve 2 dans les deux cas
- pour le Code B
  - avec la Loi 1 :  $(1 + 2 + 3 + 3) \times \frac{1}{4} = \frac{9}{4} = 2.25$
  - avec la Loi 2 :  $1 \times \frac{1}{2} + 2 \times \frac{1}{4} + (3 + 3) \times \frac{1}{8} = \frac{7}{4} = 1.75$

*Le meilleur code dépend de la loi d'émission de la source*

## Entropie d'une source discrète

Source discrète sans mémoire  $X = (\mathcal{X}, p_X)$  :

- Alphabet fini  $\mathcal{X} = \{a_1, \dots, a_K\}$
- Loi de probabilité  $p_X(a_1), \dots, p_X(a_K)$

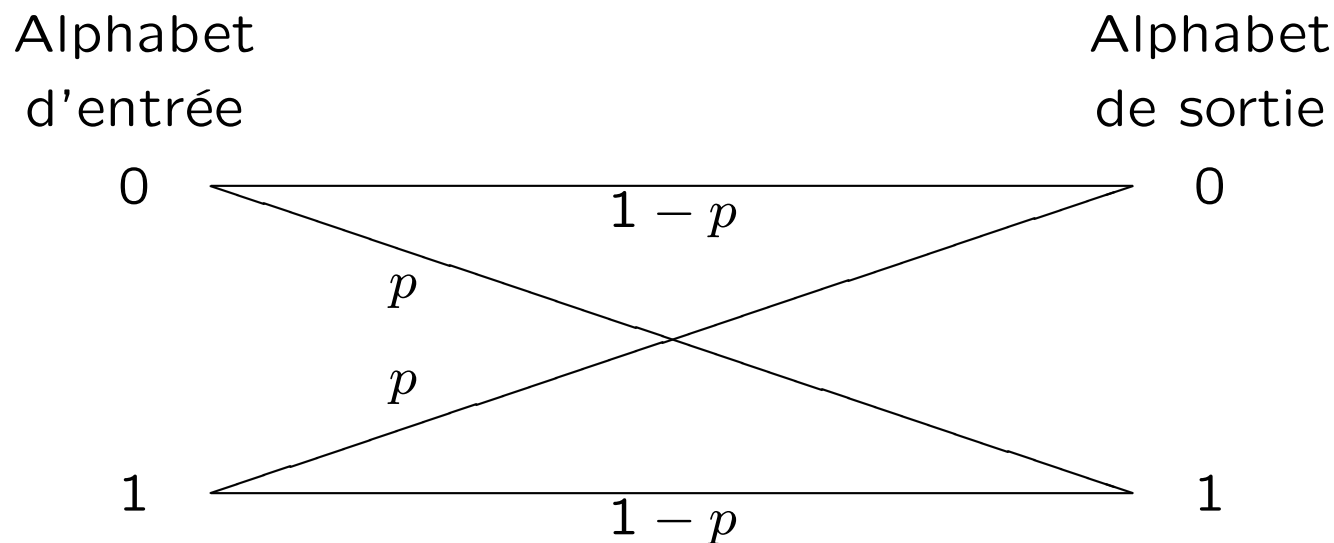
La **longueur moyenne** de tout code est au moins égale à l'**entropie** :

$$H(X) = \sum_{k=1}^K p_X(a_k) \log_2 \frac{1}{p_X(a_k)}$$

Dans l'exemple précédent

- Loi 1 :  $H(X) = 4 \times \left(\frac{1}{4} \times 2\right) = 2$
- Loi 2 :  $H(X) = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = \frac{7}{4} = 1.75$

## Canal binaire symétrique



$p$  est la probabilité d'erreur du canal.

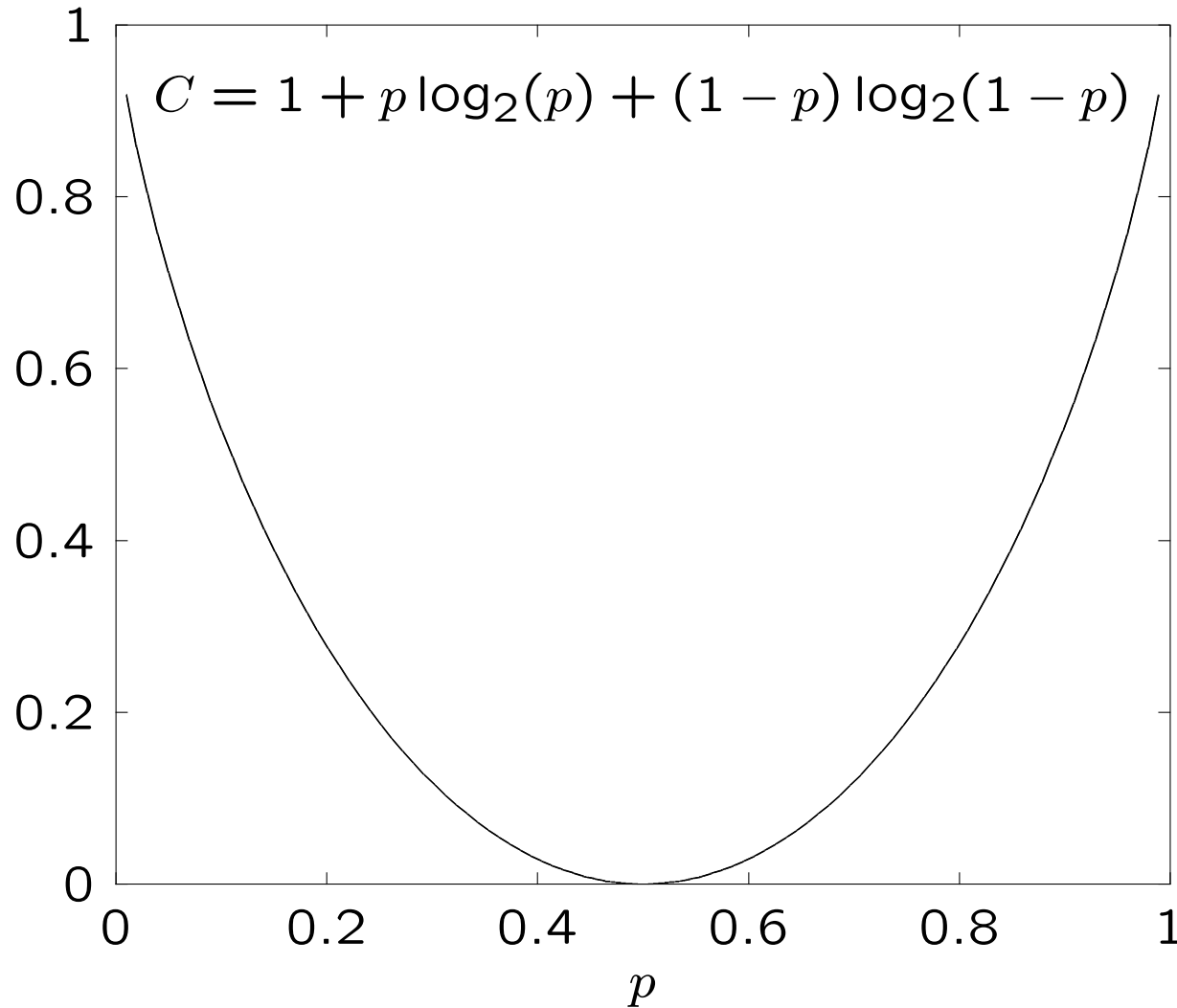
Pour combattre les effets du bruit on ajoutera de la redondance. Par exemple, le code à répétition de longueur 3 :

$$0 \mapsto 000$$

$$1 \mapsto 111$$

Ce code à un taux de transmission 0.33 et corrige une erreur.

## Capacité du canal binaire symétrique



La capacité est le **taux de transmission maximal** du code à utiliser pour transmettre de l'information « dans de bonnes conditions ».

Par ex.  $C(0.01) = 0.919$ . Il y a moyen de faire (beaucoup) mieux que le code à répétition.

## Deux résultats importants

[Shannon, 1948]

### **Premier théorème de Shannon** (Codage de source)

1. On peut coder toute source en utilisant un nombre de bits par lettre aussi proche que l'on veut de son entropie.
2. On ne peut pas faire mieux.

### **Deuxième théorème de Shannon** (Codage de canal)

1. On peut transmettre de l'information de façon fiable en utilisant un code correcteur d'erreur de taux de transmission inférieur à la capacité du canal utilisé.
2. On ne peut pas faire mieux.

*Ces résultats ne sont pas constructifs*

## Modèles discrets – Le monde numérique

Nous n'examinerons ici que des modèles discrets et même finis pour les sources et les canaux

Le monde réel est analogique, les modèles que nous allons étudier ici appartiennent au monde numérique.

Le passage du continu au discret relève d'une autre problématique que nous n'aborderons pas ici.

## II. Une mesure de l'information

## Espace probabilisé discret

L'alphabet est  $\mathcal{X}$  (fini en pratique)

Variable aléatoire  $X$  à valeurs dans  $\mathcal{X}$

Loi de probabilité  $p_X(x), x \in \mathcal{X}$

Moyenne d'une variable aléatoire réelle

$$\bar{V} = \sum_{x \in \mathcal{X}} p_X(x) V(x) = \sum_x p(x) V(x)$$

## Espace probabilisé joint

Alphabet  $\mathcal{X} \times \mathcal{Y}$  muni de la loi produit  $p_{XY}(x, y)$ ,  $(x, y) \in \mathcal{X} \times \mathcal{Y}$

Variations aléatoires  $X$  et  $Y$  à valeurs dans  $\mathcal{X}$  et  $\mathcal{Y}$  respectivement.

Lois marginales

$$p_X(x) = \sum_y p_{XY}(x, y)$$

$$p_Y(y) = \sum_x p_{XY}(x, y)$$

Probabilité conditionnelle

$$\Pr_{XY}[X = x \mid Y = y] = \frac{p_{XY}(x, y)}{p_Y(y)}$$

$$\Pr_{XY}[Y = y \mid X = x] = \frac{p_{XY}(x, y)}{p_X(x)}$$

En l'absence d'ambiguïté, nous noterons  $p(x), p(y), p(x \mid y), p(y \mid x)$ , ou parfois  $p(x \mid y) = p_{X|Y}(x \mid y)$  et  $p(y \mid x) = p_{Y|X}(y \mid x)$ .

$X$  et  $Y$  sont indépendantes si

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, p_{XY}(x, y) = p_X(x)p_Y(y)$$

## Incertitude et information

La quantité d'information obtenue lorsque l'évènement  $X = x$  se réalise est liée à l'incertitude sur cet évènement. Nous cherchons

- fonction positive et décroissante de la probabilité :  $I(x) = f(p(x))$
- l'évènement certain ne produit aucune information :  $f(1) = 0$
- un évènement impossible fournit une quantité infinie d'information :  
 $f(0) = \infty$
- fonction additive : l'information de deux évènements indépendants s'additionne :  $f(p(x)p(y)) = f(p(x)) + f(p(y))$

## Information propre – Entropie

Nous utiliserons comme mesure de l'incertitude la quantité suivante :

$$I(x) = \log_b \frac{1}{p(x)}$$

qui sera appelée *information propre* de  $x$ . La base du logarithme est arbitraire. L'unité d'information que nous utiliserons est le *bit*, défini par

*Un bit est égal à la quantité d'information fournie par le choix d'une alternative parmi deux équiprobables.*

Autrement dit nous utiliserons le logarithme en base 2.

**Définition** (Entropie) - Moyenne de l'information propre

$$H(X) = - \sum_x p(x) \log_2 p(x)$$

## Information mutuelle

Si nous voulons quantifier la corrélation entre deux évènements, il faut se demander comment la réalisation de l'un d'entre eux «  $Y = y$  » va modifier l'incertitude sur l'autre «  $X = x$  ».

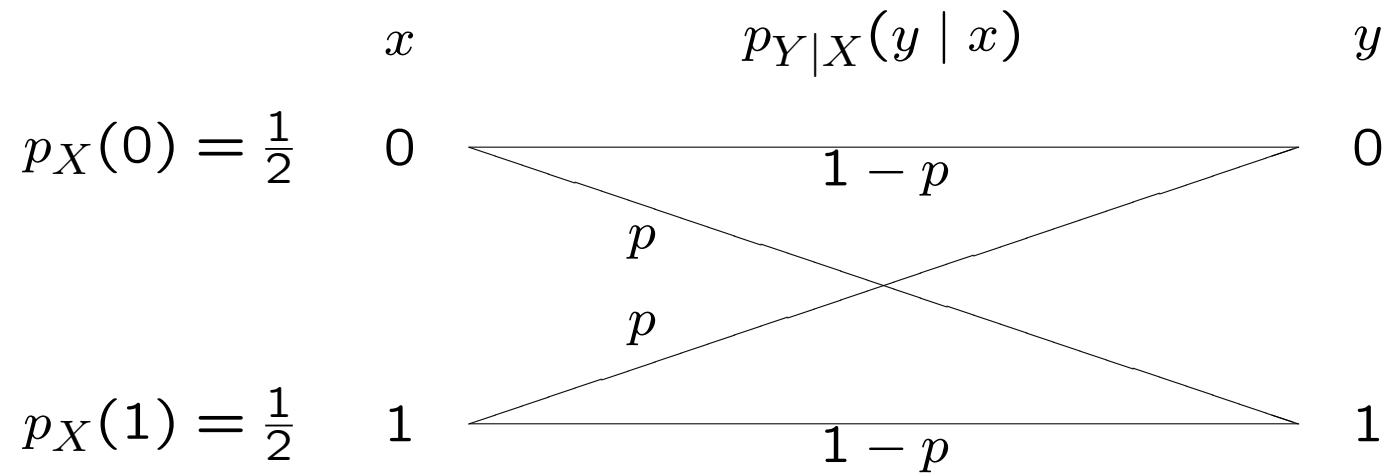
La probabilité *a priori*  $p(x) = \Pr[X = x]$  va devenir la probabilité *a posteriori*  $p(x | y) = \Pr[X = x | Y = y]$ .

La différence entre les deux « quantités d'incertitude » correspondantes sera l'*information mutuelle* entre  $x$  et  $y$  :

$$I(x; y) = I(x) - I(x | y) = \log_2 \frac{\Pr[X = x | Y = y]}{\Pr[X = x]} = \log_2 \frac{p(x | y)}{p(x)}$$

## Exemple

Soit le canal binaire symétrique de probabilité de transition  $p \leq \frac{1}{2}$ . Les symboles 0 et 1 sont émis selon une loi uniforme.



$$I(0; 0) = I(1; 1) = \log_2(2 - 2p) \geq 0$$

$$I(0; 1) = I(1; 0) = \log_2(2p) \leq 0.$$

## Information mutuelle moyenne

L'information mutuelle est symétrique

$$I(x; y) = I(y; x) = \log_2 \frac{p(x, y)}{p(x)p(y)} = \log_2 \frac{p(x | y)}{p(x)}$$

- $I(x; y) \geq 0$  ssi  $p(x | y) \geq p(x)$ ,
- $I(x; y) \leq 0$  ssi  $p(x | y) \leq p(x)$ ,
- $I(x; y) = 0$  ssi  $p(x | y) = p(x)$ .

**Définition** (information mutuelle moyenne)

$$I(X; Y) = \sum_{x,y} p(x, y) I(x; y) = \sum_{x,y} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}$$

**Théorème**

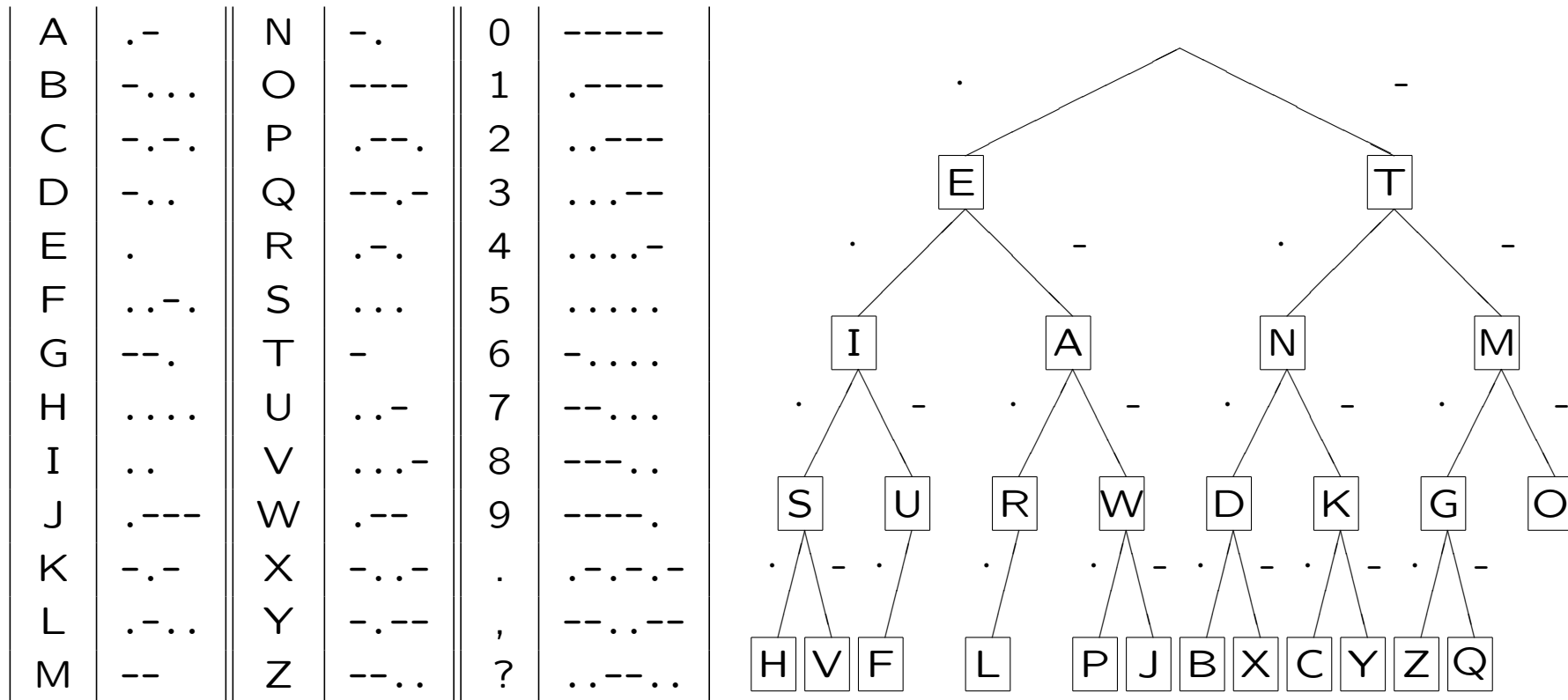
$$I(X; Y) \geq 0$$

avec égalité si et seulement si  $X$  et  $Y$  sont indépendantes

### III. Codage des sources discrètes

## Exemple du code Morse

L'idée générale : lettres fréquentes → mots de code courts

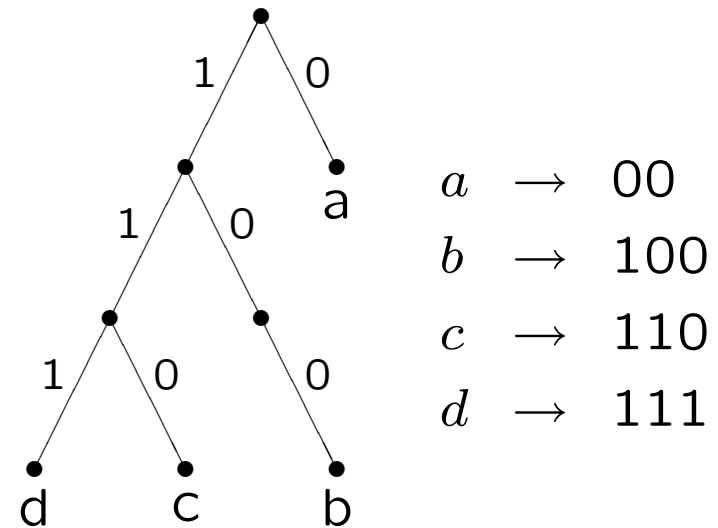
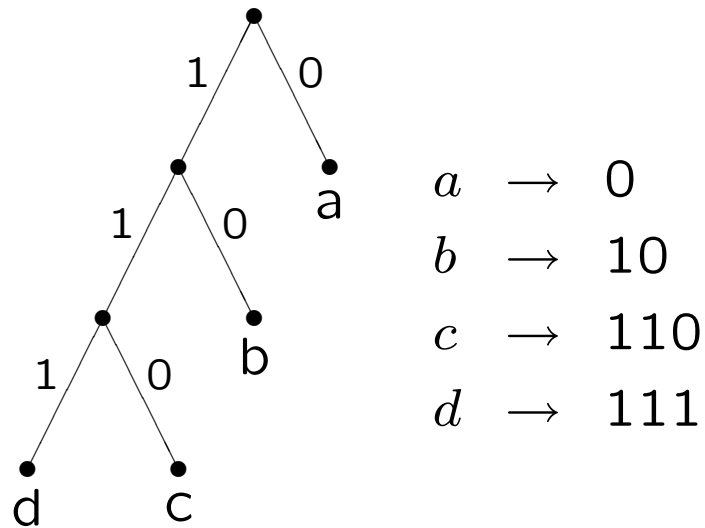


Problème : codage ambigu ("BAM" et "NIJ" → "-.....-")

Il s'agit en fait d'un code ternaire

## Codes préfixes

À tout code binaire (ensemble de mots binaires) on associe le plus petit arbre binaire contenant tout les mots de code



Un code est préfixe si aucun de ses mots n'est le début d'un autre.

Un code est préfixe ssi ses mots sont les feuilles de son arbre associé

## Code et codage

$(\mathcal{A}^* = \bigcup_i \mathcal{A}^i$  les mots de  $\mathcal{A}$ ,  $\parallel$  la concaténation)

Soit un alphabet (fini)  $\mathcal{X}$ .

### Définition

- *code* :  $\varphi : \mathcal{X} \rightarrow \{0, 1\}^*$
- *codage* :  $\psi : \mathcal{X}^* \rightarrow \{0, 1\}^*$
- *codage associé* à un code :  $(x_1, \dots, x_L) \rightarrow (\varphi(x_1) \parallel \dots \parallel \varphi(x_L))$   
(tout codage n'est pas associé à un code)
- Un code est à *décodage unique* si son codage associé est injectif

**Proposition** Tout code préfixe est à décodage unique.

## Inégalité de Kraft – Théorème de Mac Millan

**Théorème** (Kraft) Il existe un code préfixe dont les  $K$  mots ont pour longueur  $n_1, n_2, \dots, n_K$  si et seulement si

$$\sum_{k=1}^K \frac{1}{2^{n_k}} \leq 1.$$

**Théorème** (Mac Millan) Il existe un code à décodage unique dont les  $K$  mots ont pour longueur  $n_1, n_2, \dots, n_K$  si et seulement si

$$\sum_{k=1}^K \frac{1}{2^{n_k}} \leq 1.$$

**Corollaire** On peut se contenter des codes préfixes

## Sources discrètes

D'une façon générale, une source discrète se décrit à l'aide d'un processus stochastique, c'est-à-dire une suite  $(X_i)_{i>0}$  de variables aléatoires, à valeur dans un alphabet fini  $\mathcal{X}$ .

- sans mémoire : les v.a. sont indépendantes
- stationnaire : comportement invariant par décalage dans le temps
- markovien :  $P(X_i | X_1, \dots, X_{i-1}) = P(X_i | X_{i-1})$  (ordre 1)

**Définition** (Entropie d'une source discrète sans mémoire)

$$H(X) = \sum_{x \in \mathcal{X}} -p(x) \log_2 p(x).$$

**Définition** (Entropie par lettre d'une source discrète)

$$H(\mathcal{X}) = \lim_{L \rightarrow \infty} \frac{1}{L} H(X_1, \dots, X_L) = \lim_{L \rightarrow \infty} H(X_L | X_{L-1}, \dots, X_1)$$

Si ces limites existent.

## Longueur moyenne par lettre

La *longueur moyenne par lettre* du codage  $\psi$  sera définie par la limite suivante (si elle existe)

$$\mathcal{L}(\psi) = \lim_{L \rightarrow \infty} \frac{1}{L} \sum_{x_1, \dots, x_L} p(x_1, \dots, x_L) |\psi(x_1, \dots, x_L)|$$

Si  $X$  est une source sans mémoire d'alphabet,  $\varphi$  un code de  $\mathcal{X}$  et  $\psi$  le codage associé à  $\psi$

$$|\varphi| = \sum_x p(x) |\varphi(x)| = \mathcal{L}(\psi)$$

L'*efficacité* d'un codage  $\psi$  pour la source  $X$  est égale à

$$E(\psi) = \frac{H(\mathcal{X})}{\mathcal{L}(\psi)} \left( = \frac{H(X)}{|\varphi|} \right)$$

## Premier théorème de Shannon

**Proposition** Soit une source  $X = (\mathcal{X}, p)$  sans mémoire d'entropie  $H$

1. Il existe un code préfixe  $\varphi$  de  $\mathcal{X}$  tel que  $H \leq |\varphi| < H + 1$ .
2. Tout code  $\varphi$  de  $\mathcal{X}$  à décodage unique vérifie  $|\varphi| \geq H$ .

**Théorème** (de codage de source de Shannon) Pour toute source discrète dont l'entropie par lettre est définie

1. il existe un codage injectif dont l'efficacité est arbitrairement proche de 1.
2. il n'existe pas de codage injectif dont l'efficacité soit  $> 1$ .

## Preuves

- Source sans mémoire  $X = (\mathcal{X}, p)$  d'entropie  $H$   
À l'aide du théorème de Mac Millan en considérant des blocs de lettres de  $\mathcal{X}$  de longueur  $L \rightarrow \infty$
- Cas général  
Séquences typiques et l'AEP (*Asymptotic Equipartition Property*)

## Les séquences typiques

Soit une source (un processus stochastique) constituée de la suite de variables aléatoires  $X_1, X_2, \dots, X_n, \dots$  à valeur dans un alphabet  $\mathcal{X}$ . On suppose que l'entropie par lettre de cette source est définie

$$\mathcal{H} = H(\mathcal{X}) = \lim_{n \rightarrow \infty} \frac{1}{n} \log_2 H(X_1, \dots, X_n)$$

**Définition** Ensemble de *séquences typiques* (de longueur  $n$ )

$$A_\varepsilon^{(n)} = \left\{ (x_1, \dots, x_n) \in \mathcal{X}^n, \left| \frac{1}{n} \log_2 \frac{1}{p(x_1, \dots, x_n)} - \mathcal{H} \right| \leq \varepsilon \right\}$$

**Définition** (*Asymptotic Equipartition Property*)

Un processus stochastique (une source) vérifie l'AEP si :

$$\forall \varepsilon > 0, \lim_{n \rightarrow \infty} \Pr \left[ A_\varepsilon^{(n)} \right] = 1.$$

## Propriétés des séquences typiques

**Proposition** Pour toute source vérifiant l'AEP

- (i)  $\frac{1}{n} \log_2 \frac{1}{p(x_1, \dots, x_n)} \xrightarrow{n \rightarrow \infty} \mathcal{H}$  presque sûrement
- (ii)  $\Pr \left[ A_\varepsilon^{(n)} \right] 2^{n(\mathcal{H}-\varepsilon)} \leq \left| A_\varepsilon^{(n)} \right| \leq 2^{n(\mathcal{H}+\varepsilon)}$

Autrement dit :

Il y a environ  $2^{n\mathcal{H}}$  séquences typiques, ces séquences sont équiprobables de probabilité  $\approx 2^{-n\mathcal{H}}$ .

## Preuve du théorème de Shannon

Grandes lignes :

1. Pour tout  $\varepsilon > 0$ , il existe un codage des séquences typiques dont la longueur moyenne par lettre n'excède pas  $\mathcal{H} + \varepsilon$
2. Pour un codage donné seule la contribution des séquences typiques à la longueur moyenne par lettre est significative
3. Pour tout  $\varepsilon > 0$ , il existe un codage dont la longueur moyenne par lettre n'excède pas  $\mathcal{H} + \varepsilon$
4. Contraposée : un codage de longueur moyenne par lettre  $< \mathcal{H}$  ne peut pas être injectif

## AEP en pratique

**Proposition** Une source sans mémoire vérifie l'AEP.

**Proposition** Une source markovienne irréductible vérifie l'AEP.

**Proposition** Une source stationnaire ergodique vérifie l'AEP.

# Algorithmes de codage de source

Les statistiques de la source sont connues

- Code de Huffman
  - + Optimal
  - + Simple
  - Code précalculé
  - Source sans mémoire uniquement
- Codage arithmétique
  - + Presque optimal
  - + Pas de précalcul
  - + Source markovienne
  - Implémentation difficile

# Algorithmes de compression universelle

Les statistiques de la source n'ont pas besoin d'être connues

- Huffman adaptatif (Source sans mémoire)
- Codage arithmétique adaptatif
- Lempel-Ziv
- Burrows-Wheeler

## IV. Capacité et codage de canal

## Canal discret sans mémoire

**Définition** Un canal discret est défini par la donnée de

- un alphabet d'entrée  $\mathcal{X} = \{a_1, \dots, a_K\}$
- un alphabet de sortie  $\mathcal{Y} = \{b_1, \dots, b_J\}$
- une loi de transition  $P_{Y|X}$ , i.e. une matrice stochastique

$$\Pi = \begin{pmatrix} P(b_1 | a_1) & \dots & P(b_J | a_1) \\ \vdots & \ddots & \vdots \\ P(b_1 | a_K) & \dots & P(b_J | a_K) \end{pmatrix}$$

Nous parlerons du canal  $\mathcal{T} = (\mathcal{X}, \mathcal{Y}, \Pi)$ .

Le canal est *sans mémoire* si la loi de transition est constante

## Capacité

La capacité d'un canal est la quantité maximale d'information pouvant transiter à travers le canal par unité de temps. Autrement dit :

*Quelle quantité d'information puis-je obtenir au maximum sur  $X$  en observant  $Y$  ?*

Cette quantité est l'information mutuelle moyenne de  $X$  et  $Y$ , et le maximum est pris par rapport à la seule chose susceptible de changer : la loi d'émission.

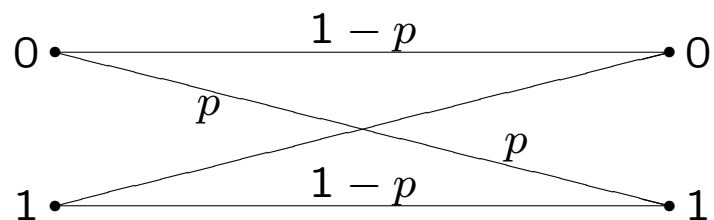
$$C = \max_{x \mapsto P(x)} I(X; Y)$$

On remarquera que  $I(X; Y)$  peut s'écrire en fonction des seules lois de transition et d'émission :

$$I(X; Y) = \sum_{x,y} P(y | x) P(x) \log_2 \frac{P(y | x)}{P(y)} \quad \text{et} \quad P(y) = \sum_x P(y | x) P(x).$$

## Exemples

Canal binaire symétrique

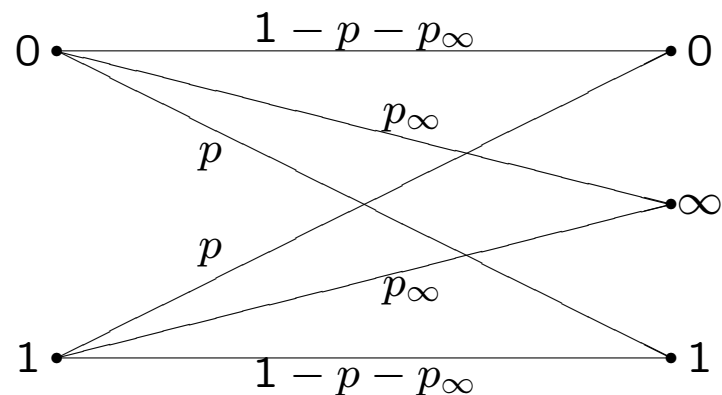


$$\Pi = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$

$$C = 1 + H_2(p)$$

où  $H_2(x) = -x \log_2(x) - (1-x) \log_2(1-x)$

Canal binaire symétrique à effacement



$$\Pi = \begin{pmatrix} 1-p-p_\infty & p_\infty & p \\ p & p_\infty & 1-p-p_\infty \end{pmatrix}$$

$$C = (1-p_\infty) \left( 1 + H_2 \left( \frac{p}{1-p_\infty} \right) \right)$$

## Codage de canal

Nous considérons un canal discret  $\mathcal{T} = (\mathcal{X}, \mathcal{Y}, \Pi)$

**Définition** Un *code en bloc* de longueur  $n$  et de cardinal  $M$  est  $M$  séquences de  $n$  lettres de  $\mathcal{X}$ . Nous parlerons de code  $(M, n)$ . Le *taux de transmission* d'un code est égal à

$$R = \frac{\log_{|\mathcal{X}|} M}{n} \leq 1$$

Un *codeur* est une procédure qui associe à toute séquence binaire finie une séquence finie de lettres de  $\mathcal{X}$ .

## Exemples

**Code à répétition** de longueur 3

$$C = \{000, 111\}$$

**Code de parité** de longueur 4

$$C = \{0000, 0011, 0101, 0110, 1001, 1010, 1100, 1111\}$$

**Code de Hamming** de longueur 7

$$C = \{0000000, 1101000, 0110100, 0011010, \\ 0001101, 1000110, 0100011, 1010001, \\ 1111111, 0010111, 1001011, 1100101, \\ 1110010, 0111001, 1011100, 0101110\}$$

## Performance d'un code – Décodage

Soit  $\mathcal{C}$  un code en bloc  $(M, n)$  utilisé dans un canal discret  $(\mathcal{X}, \mathcal{Y}, \Pi)$

**Définition** Un *algorithme de décodage* de  $\mathcal{C}$  est une procédure qui a tout bloc de  $n$  lettres de  $\mathcal{Y}$  associe un mot de code de  $\mathcal{C}$ .

L'événement « mauvais décodage » pour un algorithme de décodage et un canal donné est défini par :

*Un mot de code  $\mathbf{x} \in \mathcal{C} \subset \mathcal{X}^n$  est transmis à travers le canal, le mot  $\mathbf{y} \in \mathcal{Y}^n$  est reçu et est décodé en  $\tilde{\mathbf{x}} \neq \mathbf{x}$ .*

**Définition** Le *taux d'erreur* de  $\mathcal{C}$  (dans le canal considéré) noté  $P_e(\mathcal{C})$  est le minimum de la probabilité de mauvais décodage pour tous les algorithmes de décodage.

## Second théorème de Shannon

**Théorème** Soit un canal discret sans mémoire de capacité  $C$ . Pour tout  $R < C$ , il existe une suite de codes en bloc  $(\mathcal{C}_n(M, n))_{n>0}$  de taux de transmission  $R_n$  telle que

$$\lim_{n \rightarrow \infty} R_n = R \quad \text{et} \quad \lim_{n \rightarrow \infty} P_e(\mathcal{C}_n) = 0$$

**Théorème (réciproque)** Soit un canal discret sans mémoire de capacité  $C$ . Tout code  $\mathcal{C}$  de taux de transmission  $R > C$  vérifie  $P_e(\mathcal{C}) > K(C, R)$ , où  $K(C, R) > 0$  dépend du canal et du taux de transmission mais est indépendant de la longueur du code.

## AEP conjointe

**Définition** Ensemble des séquences typiques conjointes

$$A_\varepsilon^{(n)} = \left\{ (x, y) \in \mathcal{X}^n \times \mathcal{Y}^n, \left| \frac{1}{n} \log_2 \frac{1}{P(x)} - H(X) \right| \leq \varepsilon, \right. \\ \left. \left| \frac{1}{n} \log_2 \frac{1}{P(y)} - H(Y) \right| \leq \varepsilon, \left| \frac{1}{n} \log_2 \frac{1}{P(x, y)} - H(X, Y) \right| \leq \varepsilon \right\}$$

Le processus  $X \times Y$  vérifie l'AEP conjointe si

$$\forall \varepsilon > 0, \lim_{n \rightarrow \infty} Pr(A_\varepsilon^{(n)}) = 1.$$

## Codage de canal, distance Hamming et théorie des codes

Canal symétrique,  $\mathcal{X} = \mathcal{Y} = A$ , lignes de  $\Pi$  égales à permutation près

Distance de Hamming :  $x = (x_1, \dots, x_n)$  et  $y = (y_1, \dots, y_n)$  dans  $A^n$

$$d_H(x, y) = |\{x_i \neq y_i \mid i = 1, \dots, n\}|$$

**Proposition** Dans un canal symétrique sans mémoire, si la loi d'émission des mots de code est uniforme, le mot  $x \in \mathcal{C}$  le plus probablement émis connaissant le mot reçu  $y \in A^n$  est un mot réalisant le minimum de  $d_H(x, y)$ .

*Preuve :*  $P(x \mid y)$  fonction décroissante de  $d_H(x, y)$

→ théorie algébrique des codes

- trouver des codes ayant une bonne distance minimale
- facile à coder (linéaires)
- facile à décoder (structure algébrique)

## Bibliographie

- [CT91]** T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.
- [Gal68]** R. G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, 1968.
- [Sen07]** N. Sendrier. *Introduction à la théorie de l'information*. Notes de cours, 2007.  
[http ://www-rocq.inria.fr/secret/Nicolas.Sendrier/thinfo.pdf](http://www-rocq.inria.fr/secret/Nicolas.Sendrier/thinfo.pdf)