Cours 3. Théorèmes fondamentaux de la théorie de l'information

A. Désilles

3 avril 2009

Résumé

Résumé

Résumé

Résumé

Résumé

Modèle de communication

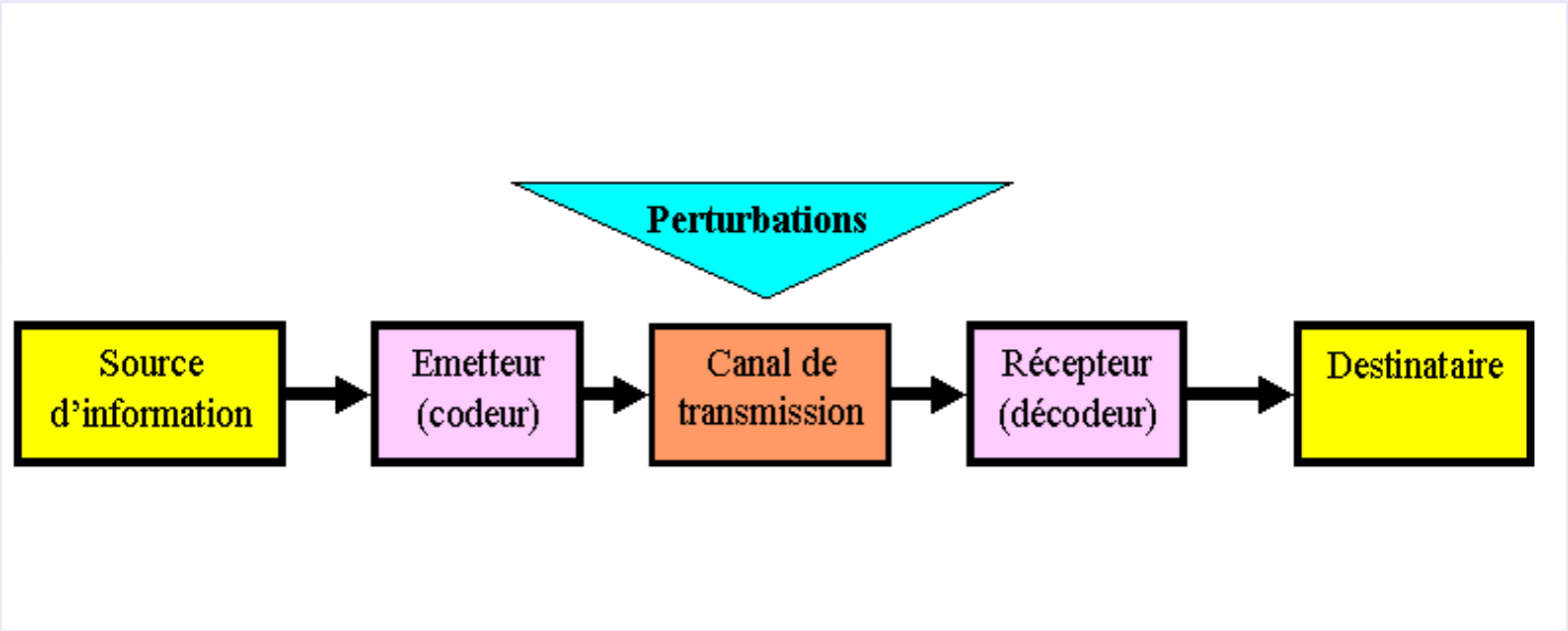

Figure: Paradigme de Shannon

Modèle d'une source d'information

Une source d'information X est décrite par un couple (Ω_X, P_X) où Ω_X est un alphabet fini et P_X est une distribution de probabilités sur Ω_X.

Exemple important : une source binaire

- $\Omega_X = \{0, 1\}$
- $P_X = \{p, 1-p\}$ avec $p = P_X(0)$

Modèle d'une source d'information

Une source d'information X est décrite par un couple (Ω_X, P_X) où Ω_X est un alphabet fini et P_X est une distribution de probabilités sur Ω_X.

Exemple important : une source binaire

- $\Omega_X = \{0, 1\}$
- $P_X = \{p, 1-p\}$ avec $p = P_X(0)$

Entropie d'une source. Définition.

Entropie d'une source

Soient $\Omega_X = \{x_1, \ldots, x_m\}$ l'alphabet fini d'une source et X la variable aléatoire associée t.q. $P[\omega_i] = p_i, \quad i = 1, \ldots, m$. On appelle **entropie** ou encore **quantité moyenne d'information** de la source la quantité

$$H(X) = H(p_1, p_2, \ldots, p_n) = E[h(x)] = -\sum_{i=1}^{m} p_i \log(p_i)$$

L'unité de mesure de cette quantité est le "bit par symbole".

D'une manière générale le codage peut être vu comme une transformation de symboles d'un alphabet donné $\Omega_1 = \{s_1, \ldots, s_n\}$ en suites de symboles d'un autre alphabet $\Omega_2 = \{c_1, \ldots, c_d\}$.

Deux problèmes de codage

1. **Codage de source ou encore codage sans bruit.** Sous cette hypothèse le meilleur code sera celui qui permettra la transmission la plus rapide possible. **Le premier théorème de Shannon** donne la solution à ce problème.
2. **Codage de canal ou encore codage en présence de bruit.** On cherchera une méthode de codage permettant une transmission aussi rapide que possible tout en minimisant la probabilité des erreurs. **Le second théorème de Shannon** donne la solution à ce problème.

Deux problèmes de codage

1. **Codage de source ou encore codage sans bruit.** Sous cette hypothèse le meilleur code sera celui qui permettra la transmission la plus rapide possible. **Le premier théorème de Shannon** donne la solution à ce problème.
2. **Codage de canal ou encore codage en présence de bruit.** On cherchera une méthode de codage permettant une transmission aussi rapide que possible tout en minimisant la probabilité des erreurs. **Le second théorème de Shannon** donne la solution à ce problème.

Vocabulaire

- **Lettre, symbole ou caractère** Tout élément d'un alphabet donné ;
- **Message ou mot** Une séquence fini m de caractères d'un alphabet donné ;
- **Longueur de mot** Le nombre $l(m)$ de caractères d'un mot m ;

Vocabulaire

- **Lettre, symbole ou caractère** Tout élément d'un alphabet donné ;
- **Message ou mot** Une séquence fini m de caractères d'un alphabet donné ;
- **Longueur de mot** Le nombre $l(m)$ de caractères d'un mot m ;

Vocabulaire

- **Lettre, symbole ou caractère** Tout élément d'un alphabet donné ;
- **Message ou mot** Une séquence fini m de caractères d'un alphabet donné ;
- **Longueur de mot** Le nombre $l(m)$ de caractères d'un mot m ;

Un code

- Soit une source S d'alphabet $\Omega_S = \{s_1, \ldots, s_n\}$ et de distribution de probabilité $P_S = \{p_1, \ldots, p_n\}$
- Un code est un ensemble $\{m_1, m_2, \ldots, m_n\}$ de n mots codes correspondants chacun à un symbole de l'alphabet de la source : $\forall\ i = 1, \ldots, n,\ \ m_i = m(s_i)$.
- Soit $l_i = l(m_i)$ les longueurs des mots m_i du code. On définit alors lalongueur moyenne du code par

$$\bar{L} = E[L] = \sum_{i=1}^{n} p_i l_i$$

Un code

- Soit une source S d'alphabet $\Omega_S = \{s_1, \ldots, s_n\}$ et de distribution de probabilité $P_S = \{p_1, \ldots, p_n\}$
- Un code est un ensemble $\{m_1, m_2, \ldots, m_n\}$ de n mots codes correspondants chacun à un symbole de l'alphabet de la source : $\forall\ i = 1, \ldots, n,\ \ m_i = m(s_i)$.
- Soit $l_i = l(m_i)$ les longueurs des mots m_i du code. On définit alors lalongueur moyenne du code par

$$\bar{L} = E[L] = \sum_{i=1}^{n} p_i l_i$$

Un code

- Soit une source S d'alphabet $\Omega_S = \{s_1, \ldots, s_n\}$ et de distribution de probabilité $P_S = \{p_1, \ldots, p_n\}$
- Un code est un ensemble $\{m_1, m_2, \ldots, m_n\}$ de n mots codes correspondants chacun à un symbole de l'alphabet de la source : $\forall\ i = 1, \ldots, n,\ \ m_i = m(s_i)$.
- Soit $l_i = l(m_i)$ les longueurs des mots m_i du code. On définit alors lalongueur moyenne du code par

$$\bar{L} = E[L] = \sum_{i=1}^{n} p_i l_i$$

Propriétés d'un code

- **Régularité** Un code $\{m_1, m_2, \ldots, m_n\}$ est dit régulier si tous les mots qui le composent sont distincts : $m_i \neq m_k, \quad \forall i \neq k$. Un code qui n'est pas régulier est dit **singulier ou irréversible**.
- **Déchiffrabilité** Un code régulier est dit déchiffrable (ou encore à décodage unique) si pour toute suite de mots de code $m^1 m^2 \ldots m^k$ il est possible de distinguer sans ambiguïté tous les mots et donc identifier les symboles $s^j, \ j = 1, \ldots, k$ composant le message.

Propriétés d'un code

- **Régularité** Un code $\{m_1, m_2, \ldots, m_n\}$ est dit régulier si tous les mots qui le composent sont distincts : $m_i \neq m_k, \quad \forall i \neq k$. Un code qui n'est pas régulier est dit **singulier ou irréversible**.
- **Déchiffrabilité** Un code régulier est dit déchiffrable (ou encore à décodage unique) si pour toute suite de mots de code $m^1 m^2 \ldots m^k$ il est possible de distinguer sans ambiguïté tous les mots et donc identifier les symboles $s^j, \ j = 1, \ldots, k$ composant le message.

Propriétés d'un code. Exemples

Soit $\Omega_S = \{a, b, c, d\}$ de distribution de probabilité
$P_S = \{0.4, 0.3, 0.2, 0.1\}$. L'entropie de cette source est $H(S) \simeq 1.85$.

S	Proba	Code 1	Code 2	Code 3	Code 4	Code 5	Code 6
a	0.4	1	00	0	0	0	0
b	0.3	0	01	10	01	10	11
c	0.2	1	10	01	011	110	100
d	0.1	0	11	010	0111	1110	101
	Long. Moy.	1	2	1.7	2	2	1.9

Le code 1 n'est pas régulier. Le code 2 est un code de longueur fixe .

Propriétés d'un code. Exemples

Soit $\Omega_S = \{a, b, c, d\}$ de distribution de probabilité
$P_S = \{0.4, 0.3, 0.2, 0.1\}$. L'entropie de cette source est $H(S) \simeq 1.85$.

S	Proba	Code 1	Code 2	Code 3	Code 4	Code 5	Code 6
a	0.4	1	00	0	0	0	0
b	0.3	0	01	10	01	10	11
c	0.2	1	10	01	011	110	100
d	0.1	0	11	010	0111	1110	101
	Long. Moy.	1	2	1.7	2	2	1.9

Le code 1 n'est pas régulier. Le code 2 est un code de longueur fixe .

Propriétés d'un code. Exemples

Soit $\Omega_S = \{a, b, c, d\}$ de distribution de probabilité $P_S = \{0.4, 0.3, 0.2, 0.1\}$. L'entropie de cette source est $H(S) \simeq 1.85$.

S	Proba	Code 1	Code 2	Code 3	Code 4	Code 5	Code 6
a	0.4	1	00	0	0	0	0
b	0.3	0	01	10	01	10	11
c	0.2	1	10	01	011	110	100
d	0.1	0	11	010	0111	1110	101
	Long. Moy.	1	2	1.7	2	2	1.9

Le code 1 n'est pas régulier. Le code 2 est un code de longueur fixe .

Propriétés d'un code. Exemples

Soit $\Omega_S = \{a, b, c, d\}$ de distribution de probabilité
$P_S = \{0.4, 0.3, 0.2, 0.1\}$. L'entropie de cette source est $H(S) \simeq 1.85$.

S	Proba	Code 1	Code 2	Code 3	Code 4	Code 5	Code 6
a	0.4	1	00	0	0	0	0
b	0.3	0	01	10	01	10	11
c	0.2	1	10	01	011	110	100
d	0.1	0	11	010	0111	1110	101
	Long. Moy.	1	2	1.7	2	2	1.9

Le code 1 n'est pas régulier. Le code 2 est un code **de longueur fixe** .

Un code indéchiffrable

S	Proba	Code 1	Code 2	Code 3	Code 4	Code 5	Code 6
a	0.4	1	00	0	0	0	0
b	0.3	0	01	10	01	10	11
c	0.2	1	10	01	011	110	100
d	0.1	0	11	010	0111	1110	101
	Long. Moy.	1	2	1.7	2	2	1.9

Le code 3 défini par $\{0, 10, 01, 010\}$ est régulier mais pas déchiffrable.
La séquence 010 correspond à la fois à trois messages différents : "d", "ca" et "ab".

Un code indéchiffrable

S	Proba	Code 1	Code 2	Code 3	Code 4	Code 5	Code 6
a	0.4	1	00	0	0	0	0
b	0.3	0	01	10	01	10	11
c	0.2	1	10	01	011	110	100
d	0.1	0	11	010	0111	1110	101
	Long. Moy.	1	2	1.7	2	2	1.9

Le code 3 défini par $\{0, 10, 01, 010\}$ est régulier mais pas déchiffrable.

La séquence 010 correspond à la fois à trois messages différents : "d", "ca" et "ab".

Un code indéchiffrable

S	Proba	Code 1	Code 2	Code 3	Code 4	Code 5	Code 6
a	0.4	1	00	0	0	0	0
b	0.3	0	01	10	01	10	11
c	0.2	1	10	01	011	110	100
d	0.1	0	11	010	0111	1110	101
	Long. Moy.	1	2	1.7	2	2	1.9

Le code 3 défini par $\{0, 10, 01, 010\}$ est régulier mais pas déchiffrable.
La séquence 010 correspond à la fois à trois messages différents : "d", "ca" et "ab".

Décodage unique : Solution 1.

Codes de longueur fixe

Un code régulier de longueur fixe peut toujours être décodé sans ambiguïté.

Désavantage : la longueur moyenne n'est pas optimale.

Décodage unique : Solution 1.

Codes de longueur fixe

Un code régulier de longueur fixe peut toujours être décodé sans ambiguïté.

Désavantage : la longueur moyenne n'est pas optimale.

Décodage unique : Solution 1.

Codes de longueur fixe

Un code régulier de longueur fixe peut toujours être décodé sans ambiguïté.

Désavantage : la longueur moyenne n'est pas optimale.

Codes de longueur fixe. Exemple. La norme ISO-8859-1

Les formats courants d'encodage de caractères d'un texte sont les codes à longueur fixe]

- Le code ASCII (American Standard Code for Information Interchange) utilise 8 bits (1 octet) dont un bit de parité pour chaque caractère. Il est possible de représenter 128 (2^7) caractères.
- La norme ISO 8859-1 représente 191 caractères, chacun sur 1 octet.
- Les standards du consortium UNICODE, utilisent une représentation des caractères sur 2 octets (16 bits). Le format UTF-16 (Universal Transformation Format, 16 bits) utilise 2 octets pour chaque caractère. UTF-8 utilise un nombre variable d'octets (de 2 à 4) en fonction du numéro de caractère.

Codes de longueur fixe. Exemple. La norme ISO-8859-1

Les formats courants d'encodage de caractères d'un texte sont les codes à longueur fixe]

- Le code ASCII (American Standard Code for Information Interchange) utilise 8 bits (1 octet) dont un bit de parité pour chaque caractère. Il est possible de représenter 128 (2^7) caractères.
- La norme ISO 8859-1 représente 191 caractères, chacun sur 1 octet.
- Les standards du consortium UNICODE, utilisent une représentation des caractères sur 2 octets (16 bits). Le format UTF-16 (Universal Transformation Format, 16 bits) utilise 2 octets pour chaque caractère. UTF-8 utilise un nombre variable d'octets (de 2 à 4) en fonction du numéro de caractère.

Codes de longueur fixe. Exemple. La norme ISO-8859-1

Les formats courants d'encodage de caractères d'un texte sont les codes à longueur fixe]

- Le code ASCII (American Standard Code for Information Interchange) utilise 8 bits (1 octet) dont un bit de parité pour chaque caractère. Il est possible de représenter 128 (2^7) caractères.
- La norme ISO 8859-1 représente 191 caractères, chacun sur 1 octet.
- Les standards du consortium UNICODE, utilisent une représentation des caractères sur 2 octets (16 bits). Le format UTF-16 (Universal Transformation Format, 16 bits) utilise 2 octets pour chaque caractère. UTF-8 utilise un nombre variable d'octets (de 2 à 4) en fonction du numéro de caractère.

Codes de longueur fixe. Exemple. La norme ISO-8859-1

ISO/CEI 8859-1															
	x0	**x1**	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**	**x8**	**x9**	**xA**	**xB**	**xC**	**xD**	**xE**	**xF**
0x	*caractères de contrôle et divers non imprimables*															
1x																
2x		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	
8x	*caractères de contrôle et divers non imprimables*															
9x																
Ax		¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬		®	¯
Bx	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
Cx	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
Dx	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
Ex	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
Fx	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

Décodage unique : Solution 2.

Utilisation d'un séparateur

Pour un canal binaire, on peut coder le i-ème symbole de la source s_i à l'aide de i caractères "1" et utiliser "0" comme séparateur.

S	Proba	Code 1
a	0.4	10
b	0.3	110
c	0.2	1110
d	0.1	11110
	Long. Moy.	3

et la séquence "abc" donnerait "101101110"

On constate que la longueur moyenne qui tient compte du séparateur est plus élevée que tous les autres codes.

Décodage unique : Solution 2.

Utilisation d'un séparateur

Pour un canal binaire, on peut coder le i-ème symbole de la source s_i à l'aide de i caractères "1" et utiliser "0" comme séparateur.

S	Proba	Code 1
a	0.4	10
b	0.3	110
c	0.2	1110
d	0.1	11110
	Long. Moy.	3

et la séquence "abc" donnerait "101101110"

On constate que la longueur moyenne qui tient compte du séparateur est plus élevée que tous les autres codes.

Décodage unique : Solution 2.

Utilisation d'un séparateur

Pour un canal binaire, on peut coder le i-ème symbole de la source s_i à l'aide de i caractères "1" et utiliser "0" comme séparateur.

S	Proba	Code 1
a	0.4	10
b	0.3	110
c	0.2	1110
d	0.1	11110
	Long. Moy.	3

et la séquence "abc" donnerait "101101110"

On constate que la longueur moyenne qui tient compte du séparateur est plus élevée que tous les autres codes.

Décodage unique : Solution 2.

Utilisation d'un séparateur

Pour un canal binaire, on peut coder le i-ème symbole de la source s_i à l'aide de i caractères "1" et utiliser "0" comme séparateur.

S	Proba	Code 1
a	0.4	10
b	0.3	110
c	0.2	1110
d	0.1	11110
	Long. Moy.	3

et la séquence "abc" donnerait "101101110"

On constate que la longueur moyenne qui tient compte du séparateur est plus élevée que tous les autres codes.

Décodage unique : Solution 3

Préfixe

On dit qu'un mot W est **un préfixe** d'un autre mot V s'il existe un mot U tel que $V = WU$. Autrement dit, le mot V commence par le mot W.

Un code instantané ou sans préfixe

On dit qu'un code donné est **sans préfixe** ou **instantané** si aucun mot du code n'est un préfixe d'un autre.

Le code 6 défini par $\{0, 11, 100, 101\}$ est sans préfixe.

Décodage unique : Solution 3

Préfixe

On dit qu'un mot W est **un préfixe** d'un autre mot V s'il existe un mot U tel que $V = WU$. Autrement dit, le mot V commence par le mot W.

Un code instantané ou sans préfixe

On dit qu'un code donné est **sans préfixe** ou **instantané** si aucun mot du code n'est un préfixe d'un autre.

Le code 6 défini par $\{0, 11, 100, 101\}$ est sans préfixe.

Décodage unique : Solution 3

Préfixe

On dit qu'un mot W est **un préfixe** d'un autre mot V s'il existe un mot U tel que $V = WU$. Autrement dit, le mot V commence par le mot W.

Un code instantané ou sans préfixe

On dit qu'un code donné est **sans préfixe** ou **instantané** si aucun mot du code n'est un préfixe d'un autre.

Le code 6 défini par $\{0, 11, 100, 101\}$ est sans préfixe.

Décodage d'un code sans préfixe

Prenons une séquence $W = 011010011101$ du code 6 donné par $\{0, 11, 100, 101\}$.

1. **Pas 1** Le premier mot du code formé en lisant de gauche à droite est $m^1 = "0"$. Donc le premier symbole est $s^1 = a$. On sépare le mot m^1 de la séquence. On obtient la nouvelle séquence $W_1 = 1101011101$.
2. Pas 2 $m^2 = "11" \Rightarrow s^2 = b$ et $W_2 = 01011101$.
3. Pas 3 $m^3 = "0" \Rightarrow s^3 = a$ et $W_3 = 1011101$.
4. Pas 4 $m^4 = "100" \Rightarrow s^4 = c$ et $W_4 = 1101$.
5. Pas 5 $m^5 = "11" \Rightarrow s^5 = b$ et $W_5 = 101$.
6. Pas 6 $m^6 = "101" \Rightarrow s^6 = d$ et $W_6 = \emptyset$.

On obtient en symboles de l'alphabet de la source : *abacbd*.

Décodage d'un code sans préfixe

Prenons une séquence $W = 011010011101$ du code 6 donné par $\{0, 11, 100, 101\}$.

1. **Pas 1** Le premier mot du code formé en lisant de gauche à droite est $m^1 = "0"$. Donc le premier symbole est $s^1 = a$. On sépare le mot m^1 de la séquence. On obtient la nouvelle séquence $W_1 = 11010011101$.
2. Pas 2 $m^2 = "11" \Rightarrow s^2 = b$ et $W_2 = 010011101$.
3. Pas 3 $m^3 = "0" \Rightarrow s^3 = a$ et $W_3 = 10011101$.
4. Pas 4 $m^4 = "100" \Rightarrow s^4 = c$ et $W_2 = 11101$.
5. Pas 5 $m^5 = "11" \Rightarrow s^5 = b$ et $W_2 = 101$.
6. Pas 6 $m^6 = "101" \Rightarrow s^6 = d$ et $W_6 = \emptyset$.

On obtient en symboles de l'alphabet de la source : $abacbd$.

Décodage d'un code sans préfixe

Prenons une séquence $W = 011010011101$ du code 6 donné par $\{0, 11, 100, 101\}$.

1. **Pas 1** Le premier mot du code formé en lisant de gauche à droite est $m^1 = "0"$. Donc le premier symbole est $s^1 = a$. On sépare le mot m^1 de la séquence. On obtient la nouvelle séquence $W_1 = 11010011101$.
2. **Pas 2** $m^2 = "11" \Rightarrow s^2 = b$ et $W_2 = 010011101$.
3. Pas 3 $m^3 = "0" \Rightarrow s^3 = a$ et $W_3 = 10011101$.
4. Pas 4 $m^4 = "100" \Rightarrow s^4 = c$ et $W_2 = 11101$.
5. Pas 5 $m^5 = "11" \Rightarrow s^5 = b$ et $W_2 = 101$.
6. Pas 6 $m^6 = "101" \Rightarrow s^6 = d$ et $W_6 = \emptyset$.

On obtient en symboles de l'alphabet de la source : $abacbd$.

Décodage d'un code sans préfixe

Prenons une séquence $W = 011010011101$ du code 6 donné par $\{0, 11, 100, 101\}$.

1. **Pas 1** Le premier mot du code formé en lisant de gauche à droite est $m^1 = "0"$. Donc le premier symbole est $s^1 = a$. On sépare le mot m^1 de la séquence. On obtient la nouvelle séquence $W_1 = 11010011101$.
2. **Pas 2** $m^2 = "11" \Rightarrow s^2 = b$ et $W_2 = 010011101$.
3. **Pas 3** $m^3 = "0" \Rightarrow s^3 = a$ et $W_3 = 10011101$.
4. Pas 4 $m^4 = "100" \Rightarrow s^4 = c$ et $W_2 = 11101$.
5. Pas 5 $m^5 = "11" \Rightarrow s^5 = b$ et $W_2 = 101$.
6. Pas 6 $m^6 = "101" \Rightarrow s^6 = d$ et $W_6 = \emptyset$.

On obtient en symboles de l'alphabet de la source : $abacbd$.

Décodage d'un code sans préfixe

Prenons une séquence $W = 01101001111101$ du code 6 donné par $\{0, 11, 100, 101\}$.

1. **Pas 1** Le premier mot du code formé en lisant de gauche à droite est $m^1 = "0"$. Donc le premier symbole est $s^1 = a$. On sépare le mot m^1 de la séquence. On obtient la nouvelle séquence $W_1 = 11010011101$.
2. **Pas 2** $m^2 = "11" \Rightarrow s^2 = b$ et $W_2 = 010011101$.
3. **Pas 3** $m^3 = "0" \Rightarrow s^3 = a$ et $W_3 = 10011101$.
4. **Pas 4** $m^4 = "100" \Rightarrow s^4 = c$ et $W_2 = 11101$.
5. **Pas 5** $m^5 = "11" \Rightarrow s^5 = b$ et $W_2 = 101$.
6. **Pas 6** $m^6 = "101" \Rightarrow s^6 = d$ et $W_6 = \emptyset$.

On obtient en symboles de l'alphabet de la source : $abacbd$.

Décodage d'un code sans préfixe

Prenons une séquence $W = 01101001101101$ du code 6 donné par $\{0, 11, 100, 101\}$.

1. **Pas 1** Le premier mot du code formé en lisant de gauche à droite est $m^1 = "0"$. Donc le premier symbole est $s^1 = a$. On sépare le mot m^1 de la séquence. On obtient la nouvelle séquence $W_1 = 11010011101$.
2. **Pas 2** $m^2 = "11" \Rightarrow s^2 = b$ et $W_2 = 010011101$.
3. **Pas 3** $m^3 = "0" \Rightarrow s^3 = a$ et $W_3 = 10011101$.
4. **Pas 4** $m^4 = "100" \Rightarrow s^4 = c$ et $W_2 = 11101$.
5. **Pas 5** $m^5 = "11" \Rightarrow s^5 = b$ et $W_2 = 101$.
6. **Pas 6** $m^6 = "101" \Rightarrow s^6 = d$ et $W_6 = \emptyset$.

On obtient en symboles de l'alphabet de la source : $abacbd$.

Décodage d'un code sans préfixe

Prenons une séquence $W = 011010011101$ du code 6 donné par $\{0, 11, 100, 101\}$.

1. **Pas 1** Le premier mot du code formé en lisant de gauche à droite est $m^1 = "0"$. Donc le premier symbole est $s^1 = a$. On sépare le mot m^1 de la séquence. On obtient la nouvelle séquence $W_1 = 11010011101$.
2. **Pas 2** $m^2 = "11" \Rightarrow s^2 = b$ et $W_2 = 010011101$.
3. **Pas 3** $m^3 = "0" \Rightarrow s^3 = a$ et $W_3 = 10011101$.
4. **Pas 4** $m^4 = "100" \Rightarrow s^4 = c$ et $W_2 = 11101$.
5. **Pas 5** $m^5 = "11" \Rightarrow s^5 = b$ et $W_2 = 101$.
6. **Pas 6** $m^6 = "101" \Rightarrow s^6 = d$ et $W_6 = \emptyset$.

On obtient en symboles de l'alphabet de la source : *abacbd*.

Décodage d'un code sans préfixe

Prenons une séquence $W = 011010011101$ du code 6 donné par $\{0, 11, 100, 101\}$.

1. **Pas 1** Le premier mot du code formé en lisant de gauche à droite est $m^1 = "0"$. Donc le premier symbole est $s^1 = a$. On sépare le mot m^1 de la séquence. On obtient la nouvelle séquence $W_1 = 11010011101$.
2. **Pas 2** $m^2 = "11" \Rightarrow s^2 = b$ et $W_2 = 010011101$.
3. **Pas 3** $m^3 = "0" \Rightarrow s^3 = a$ et $W_3 = 10011101$.
4. **Pas 4** $m^4 = "100" \Rightarrow s^4 = c$ et $W_2 = 11101$.
5. **Pas 5** $m^5 = "11" \Rightarrow s^5 = b$ et $W_2 = 101$.
6. **Pas 6** $m^6 = "101" \Rightarrow s^6 = d$ et $W_6 = \emptyset$.

On obtient en symboles de l'alphabet de la source : *abacbd*.

Décodage d'un code sans préfixe

Prenons une séquence $W = 011010011101$ du code 6 donné par $\{0, 11, 100, 101\}$.

1. **Pas 1** Le premier mot du code formé en lisant de gauche à droite est $m^1 = "0"$. Donc le premier symbole est $s^1 = a$. On sépare le mot m^1 de la séquence. On obtient la nouvelle séquence $W_1 = 11010011101$.
2. **Pas 2** $m^2 = "11" \Rightarrow s^2 = b$ et $W_2 = 010011101$.
3. **Pas 3** $m^3 = "0" \Rightarrow s^3 = a$ et $W_3 = 10011101$.
4. **Pas 4** $m^4 = "100" \Rightarrow s^4 = c$ et $W_2 = 11101$.
5. **Pas 5** $m^5 = "11" \Rightarrow s^5 = b$ et $W_2 = 101$.
6. **Pas 6** $m^6 = "101" \Rightarrow s^6 = d$ et $W_6 = \emptyset$.

On obtient en symboles de l'alphabet de la source : *abacbd*.

Codes instantanés : existence

Problème

Soient l'alphabet de la source $\Omega_S = \{s_1, \ldots, s_n\}$ de taille n et et l'alphabet du canal $\Omega_C = \{c_1, \ldots, c_d\}$ de taille d. Étant donnés n nombres entiers positifs $(l_1, l_2, \ldots, l_n) \in \mathbb{Z}_+^*$ existe-t-il un code régulier instantané de n mots $\{m_1, \ldots, m_n\}$ tel que chaque nombre l_i soit la longueur du mot de code m_i ?

Inégalité de Kraft

Théorème

Un code instantané de longueurs de mots données $l_1, \ldots, l_n$ existe si et seulement si

$$\sum_{i=1}^{n} d^{-l_i} \leq 1$$

où d est la taille de l'alphabet du canal.

Problème de codage sans bruit

Problème

Soit une source S d'alphabet $\Omega_S = \{s_1, \ldots, s_n\}$ de taille n et de distribution de probabilités $P_S = \{p_1, \ldots, p_n\}$. Soit un canal d'alphabet $\Omega_C = \{c_1, \ldots, c_d\}$ de taille d, sans bruit, stationnaire et sans mémoire. Existe-t-il un code qui minimise la longueur moyenne de mots $\bar{L}$?

Borne inférieure

Théorème

Soit une source S d'alphabet $\Omega_S = \{s_1, \ldots, s_n\}$ de taille n et de distribution de probabilités $P_S = \{p_1, \ldots, p_n\}$. Soit un canal d'alphabet $\Omega_C = \{c_1, \ldots, c_d\}$ de taille d, sans bruit, stationnaire et sans mémoire. Soit un code déchiffrable $\{m_1, \ldots, m_n\}$ de longueurs de mots $\{l_1, \ldots, l_n\}$. Alors la longueur moyenne de mots de code vérifie :

$$\bar{L} = \sum_{i=1}^{n} p(s_i) l_i \geq \frac{H(S)}{\log(d)}$$

L'égalité n'est possible que si $\forall i = 1, \ldots, n, \quad p_i = d^{-l_i}$.

Code absolument optimal

Un code dont la longueur moyenne de mots atteint la borne inférieure s'appelle **absolument optimal**.

Un code dont la longueur moyenne de mots atteint la borne inférieure s'appelle **absolument optimal**. Un exemple de code absolument optimal est donné par la tableau suivant

S	Proba	Code
a	0.5	0
b	0.25	10
c	0.125	110
d	0.125	111

et on a $H(S) = \bar{L} = \frac{7}{4}$.

Code absolument optimal

Un code dont la longueur moyenne de mots atteint la borne inférieure s'appelle **absolument optimal**.
Un code dont la longueur moyenne de mots atteint la borne inférieure s'appelle **absolument optimal**. Un exemple de code absolument optimal est donné par la tableau suivant

S	Proba	Code
a	0.5	0
b	0.25	10
c	0.125	110
d	0.125	111

et on a $H(S) = \bar{L} = \frac{7}{4}$.

Borne supérieure

Un code absolument optimal n'est pas toujours réalisable.

Les longueurs de mots doivent vérifier $p_i = d^{-l_i}$ et donc $l_i = \dfrac{\log p_i}{\log d}$.

Théorème

Soit une source S d'alphabet $\Omega_S = \{s_1, \ldots, s_n\}$ de taille n et de distribution de probabilités $P_S = \{p_1, \ldots, p_n\}$. Soit un canal d'alphabet $\Omega_C = \{c_1, \ldots, c_d\}$ de taille d, sans bruit, stationnaire et sans mémoire. Alors il existe un code déchiffrable dont la longueur moyenne de mots de code vérifie :

$$\frac{H(S)}{\log(d)} \leq \bar{L} = \sum_{i=1}^{n} p(s_i) l_i < \frac{H(S)}{\log(d)} + 1$$

Borne supérieure

Un code absolument optimal n'est pas toujours réalisable.

Les longueurs de mots doivent vérifier $p_i = d^{-l_i}$ et donc $l_i = \dfrac{\log p_i}{\log d}$.

Théorème

Soit une source S d'alphabet $\Omega_S = \{s_1, \ldots, s_n\}$ de taille n et de distribution de probabilités $P_S = \{p_1, \ldots, p_n\}$. Soit un canal d'alphabet $\Omega_C = \{c_1, \ldots, c_d\}$ de taille d, sans bruit, stationnaire et sans mémoire. Alors il existe un code déchiffrable dont la longueur moyenne de mots de code vérifie :

$$\frac{H(S)}{\log(d)} \leq \bar{L} = \sum_{i=1}^{n} p(s_i) l_i < \frac{H(S)}{\log(d)} + 1$$

Borne supérieure

Un code absolument optimal n'est pas toujours réalisable.

Les longueurs de mots doivent vérifier $p_i = d^{-l_i}$ et donc $l_i = \frac{\log p_i}{\log d}$.

Théorème

Soit une source S d'alphabet $\Omega_S = \{s_1, \ldots, s_n\}$ de taille n et de distribution de probabilités $P_S = \{p_1, \ldots, p_n\}$. Soit un canal d'alphabet $\Omega_C = \{c_1, \ldots, c_d\}$ de taille d, sans bruit, stationnaire et sans mémoire. Alors il existe un code déchiffrable dont la longueur moyenne de mots de code vérifie :

$$\frac{H(S)}{\log(d)} \leq \bar{L} = \sum_{i=1}^{n} p(s_i) l_i < \frac{H(S)}{\log(d)} + 1$$

Exemple

- Soit une source X d'alphabet $\omega_X = \{a, b\}$ et de distribution de probabilité $P_X = \{3/4, 1/4\}$.
- L'entropie de cette source est

$$H(X) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = \simeq 0.811$$

- Un code simple : $m(a) = 0,\ m(b) = 1$
- Longueur moyenne de mots $\bar{L}_1 = 1$ bit par caractère.
- Un message T de longueur initiale $l(T)$ sera alors codé par une suite binaire de $l(T)$ bits.

Exemple

- Soit une source X d'alphabet $\omega_X = \{a, b\}$ et de distribution de probabilité $P_X = \{3/4, 1/4\}$.
- L'entropie de cette source est

$$H(X) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = \simeq 0.811$$

- Un code simple : $m(a) = 0,\ m(b) = 1$
- Longueur moyenne de mots $\bar{L}_1 = 1$ bit par caractère.
- Un message T de longueur initiale $l(T)$ sera alors codé par une suite binaire de $l(T)$ bits.

Exemple

- Soit une source X d'alphabet $\omega_X = \{a, b\}$ et de distribution de probabilité $P_X = \{3/4, 1/4\}$.
- L'entropie de cette source est

$$H(X) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = \simeq 0.811$$

- Un code simple : $m(a) = 0,\ m(b) = 1$
- Longueur moyenne de mots $\bar{L}_1 = 1$ bit par caractère.
- Un message T de longueur initiale $l(T)$ sera alors codé par une suite binaire de $l(T)$ bits.

Exemple

- Soit une source X d'alphabet $\omega_X = \{a, b\}$ et de distribution de probabilité $P_X = \{3/4, 1/4\}$.
- L'entropie de cette source est

$$H(X) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = \simeq 0.811$$

- Un code simple : $m(a) = 0,\ m(b) = 1$
- Longueur moyenne de mots $\bar{L}_1 = 1$ bit par caractère.
- Un message T de longueur initiale $l(T)$ sera alors codé par une suite binaire de $l(T)$ bits.

Exemple

- Soit une source X d'alphabet $\omega_X = \{a, b\}$ et de distribution de probabilité $P_X = \{3/4, 1/4\}$.
- L'entropie de cette source est

$$H(X) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = \simeq 0.811$$

- Un code simple : $m(a) = 0,\ m(b) = 1$
- Longueur moyenne de mots $\bar{L}_1 = 1$ bit par caractère.
- Un message T de longueur initiale $l(T)$ sera alors codé par une suite binaire de $l(T)$ bits.

Coder les couples de symboles

- Avec un alphabet de taille 2 il est possible de former $2^2 = 4$ couples de caractères différents.
- On peut considérer cela comme un nouvel alphabet, d'une nouvelle source, $Y = (X_1, X_2)$
- Compte tenu de l'indépendance de X_1 et X_2 on peut construire la distribution de probabilité de Y :

Y	aa	ab	ba	bb
P_Y	$p(a)^2 = 9/16$	$p(a)p(b) = 3/16$	$p(a)p(b) = 3/16$	$p(b)^2 = 1/16$

Coder les couples de symboles

- Avec un alphabet de taille 2 il est possible de former $2^2 = 4$ couples de caractères différents.
- On peut considérer cela comme un nouvel alphabet, d'une nouvelle source, $Y = (X_1, X_2)$
- Compte tenu de l'indépendance de X_1 et X_2 on peut construire la distribution de probabilité de Y :

Y	aa	ab	ba	bb
P_Y	$p(a)^2 = 9/16$	$p(a)p(b) = 3/16$	$p(a)p(b) = 3/16$	$p(b)^2 = 1/16$

Coder les couples de symboles

- Avec un alphabet de taille 2 il est possible de former $2^2 = 4$ couples de caractères différents.
- On peut considérer cela comme un nouvel alphabet, d'une nouvelle source, $Y = (X_1, X_2)$
- Compte tenu de l'indépendance de X_1 et X_2 on peut construire la distribution de probabilité de Y :

Y	aa	ab	ba	bb
P_Y	$p(a)^2 = 9/16$	$p(a)p(b) = 3/16$	$p(a)p(b) = 3/16$	$p(b)^2 = 1/16$

Coder les couples de symboles

- Avec un alphabet de taille 2 il est possible de former $2^2 = 4$ couples de caractères différents.
- On peut considérer cela comme un nouvel alphabet, d'une nouvelle source, $Y = (X_1, X_2)$
- Compte tenu de l'indépendance de X_1 et X_2 on peut construire la distribution de probabilité de Y :

Y	aa	ab	ba	bb
P_Y	$p(a)^2 = 9/16$	$p(a)p(b) = 3/16$	$p(a)p(b) = 3/16$	$p(b)^2 = 1/16$

Coder les couples de symboles

- Avec un alphabet de taille 2 il est possible de former $2^2 = 4$ couples de caractères différents.
- On peut considérer cela comme un nouvel alphabet, d'une nouvelle source, $Y = (X_1, X_2)$
- Compte tenu de l'indépendance de X_1 et X_2 on peut construire la distribution de probabilité de Y :

Y	aa	ab	ba	bb
P_Y	$p(a)^2 = 9/16$	$p(a)p(b) = 3/16$	$p(a)p(b) = 3/16$	$p(b)^2 = 1/16$

Coder les couples de symboles

- L'entropie de Y peut aussi être déduite de l'indépendance de X_1 et X_2 :

$$H(Y) = H(X_1, X_2) = H(X_1) + H(X_2) = 2H(X)$$

- Il est possible de trouver un code déchiffrable dont la longueur moyenne de mots de code vérifie

$$H(Y) \leq \bar{L}_2 =< H(Y) + 1 \quad \Leftrightarrow \quad 2H(X) \leq \bar{L}_2 =< 2H(X) + 1$$

- On a alors la relation :

$$H(X) \leq \frac{\bar{L}_2}{2} =< H(X) + \frac{1}{2}$$

Coder les couples de symboles

- L'entropie de Y peut aussi être déduite de l'indépendance de X_1 et X_2 :

$$H(Y) = H(X_1, X_2) = H(X_1) + H(X_2) = 2H(X)$$

- Il est possible de trouver un code déchiffrable dont la longueur moyenne de mots de code vérifie

$$H(Y) \leq \bar{L}_2 =< H(Y) + 1 \quad \Leftrightarrow \quad 2H(X) \leq \bar{L}_2 =< 2H(X) + 1$$

- On a alors la relation :

$$H(X) \leq \frac{\bar{L}_2}{2} =< H(X) + \frac{1}{2}$$

Coder les couples de symboles

- L'entropie de Y peut aussi être déduite de l'indépendance de X_1 et X_2 :

$$H(Y) = H(X_1, X_2) = H(X_1) + H(X_2) = 2H(X)$$

- Il est possible de trouver un code déchiffrable dont la longueur moyenne de mots de code vérifie

$$H(Y) \leq \bar{L}_2 =< H(Y) + 1 \quad \Leftrightarrow \quad 2H(X) \leq \bar{L}_2 =< 2H(X) + 1$$

- On a alors la relation :

$$H(X) \leq \frac{\bar{L}_2}{2} =< H(X) + \frac{1}{2}$$

Avec ce code on a $\frac{27}{17}$ bits par symbole de Y en moyenne.

En unités de mesure de la source X on a $\bar{L}_2 = \frac{27}{2 \cdot 17} \simeq 0.8$ bits par symbole de X.

Un message T de longueur $l(T)$ sera codé avec le code 2, en moyenne par $0.8l(T)$ bits binaires au lieu de $l(T)$ bits avec le premier code.

Avec ce code on a $\frac{27}{17}$ bits par symbole de Y en moyenne.

En unités de mesure de la source X on a $\bar{L}_2 = \frac{27}{2 \cdot 17} \simeq 0.8$ bits par symbole de X.

Un message T de longueur $l(T)$ sera codé avec le code 2, en moyenne par $0.8l(T)$ bits binaires au lieu de $l(T)$ bits avec le premier code.

Avec ce code on a $\frac{27}{17}$ bits par symbole de Y en moyenne.

En unités de mesure de la source X on a $\bar{L}_2 = \frac{27}{2 \cdot 17} \simeq 0.8$ bits par symbole de X.

Un message T de longueur $l(T)$ sera codé avec le code 2, en moyenne par $0.8l(T)$ bits binaires au lieu de $l(T)$ bits avec le premier code.

Extension de source

Définition

Soit une source X d'alphabet $\Omega_X = \{x_1, \ldots, n\}$. On appelle extension d'ordre s de la source X la source $Y = (X_1, \ldots, X_s)$ où $X_i,\ i = 1, \ldots, s$ sont les variables aléatoires indépendantes et identiquement distribuées selon la distribution de X.

Premier théorème de Shannon

Théorème

Soit une source X d'alphabet $\Omega_X = \{x_1, \ldots, x_n\}$ de taille n et de distribution de probabilités $P_X = \{p_1, \ldots, p_n\}$. Soit un canal d'alphabet $\Omega_C = \{c_1, \ldots, c_d\}$ de taille d, sans bruit, stationnaire et sans mémoire. Alors il existe un procédé de codage déchiffrable dont la longueur moyenne de mots de code est aussi voisine que l'on souhaite de la borne inférieure $\frac{H(S)}{\log(d)}$.