Codes correcteurs

Contents

Part I
Généralités

1 Transmission de messages

Entre téléphones, ordinateurs, clés USB, CD, sondes spatiales...

Exemple de message

That's one small step for a man, a giant leap for mankind

Un premier exercice

Message reçu :
" les plofs, ils ne fcnt riin, ils sont touuuurs en vacancts ".
Trouver l'original.

Un deuxième

0660124983

Différence entre les deux ?

Que faire pour corriger les numéros de téléphone erronés ?

Le schéma

$$m \underset{codage}{\longrightarrow} c \underset{transmission}{\approx\approx} c' \underset{correction}{\longrightarrow} c \underset{décodage}{\longrightarrow} m$$

L'erreur

On note

$$e = c' - c$$

C'est l'erreur causée par la transmission.

2 Le NIR

Le numéro d'inscription au répertoire (numéro de sécurité sociale) :
1 chiffre pour le sexe, 2 pour l'année de naissance... .
Au total, un nombre n à 13 chiffres, et une clé q à 2 chiffres.

La clé

C'est l'opposé de n modulo 97 :

$$n + q \equiv 0\,[97]$$

La clé peut servir à déceler une erreur, pas à corriger.

3 Le code de répétition

Codage

Chaque bit est répété trois fois ; pourquoi pas deux fois ?
$m = 1 \to c = 111$
$m = 0 \to c = 000$

Correction

$c' = 111$, ou 110, ou 101, ou $011 \to c = 111$

Le CBS

(canal binaire symétrique) : on fait l'hypothèse suivante :
Chaque bit a une probabilité $p \ll 1$ d'être modifié par la transmission.

En utilisant le code de répétition

Pour chaque bit, probabilité de ne pas être corrigé correctement :

$$p^3 + \binom{3}{2} p^2(1-p) = 3p^2 - 2p^3$$

Par exemple, si 1 sur 1000 est incorrect $\left(p = 10^{-3}\right)$, seulement 3 sur un million avec codage et correction.

4 La notion de codage

On suppose $K = \mathbb{F}_2 = \mathbb{Z}/2\mathbb{Z}$, $1 \leq k \leq n$; on note $E = K^n$.
Un codage est une application f de K^k dans E *injective*.

$$R = \frac{k}{n}$$

est appelé taux de transmission ou rendement (rate).

Codage systématique

On dit que le codage est systématique si m est toujours un préfixe de

$$c = f(m)$$

Le codage est donc obtenu en ajoutant $n-k$ bits dits de redondance au message m.

Les codes linéaires

Dans la suite, on supposera f linéaire.

On note C l'image de f ; c'est un SEV de E de dimension $k \geq 1$: l'ensemble des messages codés ; k est la longueur du message m, n est la longueur de $c = f(m)$.

5 La matrice génératrice

Notons $(e_1, ..., e_k)$ la base canonique de K^k.

On peut exprimer le codage à l'aide d'une matrice G, dite matrice génératrice. Les lignes de G sont les

$$L_i = f(e_i)$$

Elles constituent une base de C ; le codage f :

$$m \to m.G = c$$

Pour le code de répétition : $n = 3$, $k = 1$. $G = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$.

Le code de répétition sur 2 bits

Chaque mot est répété trois fois ; $n = 6$, $k = 2$; codage :

$m = 10 \to c = 101010$

$m = 00 \to c = 000000$

$m = 01 \to c = 010101$

La matrice génératrice :

$$G = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

Le codage systématique

Dans le cas précédent, la matrice G est de la forme

$$[I_k, Q]$$

On peut toujours écrire

$$G = [A, B]$$

avec A carrée de taille k ; si A est inversible, on peut se ramener à un codage systématique, sans changer le code C.

Dans le cas général, ce n'est possible qu'après permutation des colonnes, donc des variables.

6 La distance de Hamming

6.1 Le poids

Soit $x \in E$; on appelle poids de x, noté $w(x)$, le nombre de bits non nuls dans x ; on vérifie aisément que

$$w(x+y) \leq w(x) + w(y)$$

La distance

On pose

$$d(x, y) = w(x - y)$$

On définit ainsi une distance sur E.

Pour un code C, on appelle distance minimum du code, notée $d(C)$, ou d, la plus petite distance entre deux éléments distincts de C.

Remarque

C étant un SEV, $d(C)$ est aussi le minimum des poids des éléments non nuls de C.

6.2 Loi suivie par $w(e)$

Le nombre de bits transmis incorrectement suit une loi, laquelle ?

Réponse

$$w(e) \sim B(n, p)$$

6.3 Propriété

Les boules fermées de centre dans C de rayon t sont disjointes si et seulement si $2t + 1 \leq d$; dans ce cas, C peut corriger t erreurs.

Plus précisément, cela signifie que si l'on est sûr qu'il y au plus t erreurs au cours de la transmission, alors on peut retrouver c à partir de c'.

Indication

Supposons $d \leq 2t$.

Supposons par exemple l'existence de $c \in C \setminus \{0\}$ tel que $c[2t:]$ soit nul. On construit alors x en remplaçant $c[:t]$ par des zéros.

On constate que x est dans l'intersection des boules de rayon t, de centres 0 et c.

Exemple

Pour les codes de répétition, $d = 3$; ce code permet de corriger une erreur.

Le cardinal d'une boule

Dans $E = K^n$, le cardinal d'une boule de rayon t est

$$\sum_{k=0}^{t} \binom{n}{k}$$

Plus généralement, si le corps est de cardinal q :

$$\sum_{k=0}^{t} \binom{n}{k} (q-1)^k$$

7 La matrice de contrôle

Rappel

Les lignes d'une matrice donnent un système d'équations du noyau.

7.1 La matrice de contrôle

C étant de dimension k dans $E = K^n$, on peut aussi le décrire par $n - k$ équations indépendantes :

$$C = \ker H$$

où $H \in M_{n-k,n}(K)$, de rang $n - k$.

On dit que H est une matrice de contrôle ; on notera $H_1, ..., H_n$ ses colonnes. En résumé

$$c \in C \iff H.c^T = 0$$

Pour le code de répétition sur 2 bits

$x_1 = x_3$, $x_2 = x_4$, $x_1 = x_5$, $x_2 = x_6$; ce qui équivaut à $H.X = 0$ avec

$$H = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Remarque

Si $G = \begin{bmatrix} I_k & Q \end{bmatrix}$, une possibilité pour H est $\begin{bmatrix} -Q^T & I_{n-k} \end{bmatrix}$.

7.2 La distance du code

Sur l'exemple, $c = (1, 0, 1, 0, 1, 0) \in C = \ker H$, ce qui se traduit pour H par

$$H_1 + H_3 + H_5 = 0$$

Plus généralement, d est le plus petit nombre de colonnes de H formant une famille liée.

8 Les codes parfaits

On dit qu'un code est parfait s'il existe un entier $t \geq 1$ tel que les boules fermées de rayon t et de centre dans C constituent une partition de E.

Exercice

Montrer que dans ce cas

$$d = 2t + 1$$

Indication

Utiliser un mot x de poids $t + 1$.

Codes parfaits avec $t = 1$ et $d = 3$

Cardinal d'une boule : $n + 1$; nombre de boules : 2^k ; nécessairement :

$$2^n = (n + 1)\, 2^k$$

Donc $n = 2^{n-k} - 1$, soit 3, 7, 15...

Pour $n = 3$, on retrouve le code de répétition.

Part II
Le code de Hamming$(7, 4, 3)$

9 La matrice de contrôle

On cherche à décrire le code à l'aide de la matrice de contrôle H.

Les paramètres : $n = 7$, $n - k = 3$, $k = 4$, $d = 3$, $R = \frac{4}{7}$

On cherche donc une matrice à 3 lignes et 7 colonnes non nulles *distinctes* ; on n'a guère le choix :

$$H = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Le code obtenu est appelé code de Hamming de type $(7, 4, 3)$.

On obtient des codes équivalents en permutant les colonnes.

$$\begin{cases} x_1 + x_3 + x_5 + x_7 = 0 \\ x_2 + x_3 + x_6 + x_7 = 0 \\ x_4 + x_5 + x_6 + x_7 = 0 \end{cases}$$

10 La correction

Dans le cas de la matrice

$$H = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

on va voir que la correction est particulièrement simple.

10.1 Un rappel

Soit M une matrice à p lignes et q colonnes : $M \in M_{p,q}(K)$.

Soit $E_j = (0, ..., 0, 1, 0...)^T$ la $j-$ième colonne canonique de taille q.

Que représente $M.E_j$?

Réponse

La $j-$ième colonne de M.

10.2 Le syndrome

Notons $e = c' - c$: l'erreur ; on suppose $w(e) \leq 1$.

On appelle syndrome :

$$s = H.e^T$$

En détail :

$$s = \begin{bmatrix} e_1 + e_3 + e_5 + e_7 \\ e_2 + e_3 + e_6 + e_7 \\ e_4 + e_5 + e_6 + e_7 \end{bmatrix}$$

On ne connaît pas e, mais on connaît s, car $H.c^T = 0$, donc

$$s = H.(c' - c)^T = H.(c')^T$$

La correction

$w(e) \leq 1$; par exemple, si $e = (0, 0, 1, 0, 0, 0, 0)$:

$$s = H.e^T = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

Plus généralement :

Si l'erreur porte sur le $j^{ième}$ bit : $e^T = E_j$.

Donc $s = H.E_j = H_j$: c'est la $j^{ième}$ colonne de H.

En résumé, le syndrome est l'écriture en base 2 de la position de l'erreur.

Et s'il n'y a pas d'erreur ?

Dans ce cas, $s = 0$.

10.3 Que représente le syndrome ?

Il existe une base de E $(u_1, ..., u_n)$ dont les k premiers vecteurs constituent une base de C, dans laquelle le syndrome est constitué des $n - k$ dernières coordonnées de c', ou de e.

C est défini dans cette base par

$$y_{k+1} = ... = y_{n-1} = y_n = 0$$

11 La matrice génératrice

La matrice H définit le code C. Evidemment, C ne possède pas une base unique. On peut choisir

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

12 Les codes cycliques

12.1 Définition

Soit

$$T : (x_0, ..., x_{n-1}) \to (x_{n-1}, x_0, ..., x_{n-2})$$

T est évidemment un endomorphisme de $E = K^n$; un code C est dit cyclique s'il est invariant par T. Exemples : les codes de répétition.

Représentation à l'aide de polynômes

Il est commode d'identifier $E = K^n$ à $K_{n-1}[X]$ et $(p_0, ..., p_{n-1})$ à

$$P = p_0 + p_1 X + ... + p_{n-1} X^{n-1}$$

Avec ces notations :

$$T(P) = XP - p_{n-1}(X^n - 1)$$

C'est donc le reste de la division euclidienne de XP par $X^n - 1$.

Passage à l'anneau quotient

les notations se simplifient si on remplace $K[X]$ par l'anneau quotient

$$A = K[X] / (X^n - 1)$$

T s'écrit alors $T : P \to X.P$; un SEV est stable par T si et seulement si c'est un idéal de A ; c'est l'ensemble des multiples d'un polynôme g, appelé générateur, qui divise $X^n - 1$.

Exemple : le code de répétition

$E = K^3$; un diviseur de $X^3 - 1$ est $g = X^2 + X + 1$; rappelons que

$$G = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}, H = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

Un codage systématique

On cherche une matrice génératrice de le forme $G = [A, I_k]$; la ligne i représente un polynôme P_i qui est multiple de g, et qui est de la forme

$$X^{n-k+i-1} - R_i$$

avec $\deg(R_i) < n-k$; R_i est donc le reste de la division euclidienne de X^{n-k+i} par g ; le codage d'un polynôme P est donc :

$$P \to (-R, P)$$

où R est le reste de la division euclidienne de $X^{n-k}.P$ par g.

13 Retour au code de Hamming

Dans $\mathbb{F}_2[X]$:

$$X^7-1 = (X+1)\left(X^3+X+1\right)\left(X^3+X^2+1\right) = \left(X^3+X+1\right)\left(X^4+X^2+X+1\right)$$

Choisissons $g = 1+X+X^3$; on obtient

$$G = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}$$

Comment trouver une matrice de contrôle ?

La matrice de contrôle

Soit C un code cyclique, de polynôme générateur g ; soit h tel que

$$X^n - 1 = g.h$$

Une matrice de contrôle H est obtenue à partir du polynôme réciproque $\widetilde{h}$ de h ; ici :

$$h = 1+X+X^2+X^4,\ \widetilde{h} = 1+X^2+X^3+X^4$$

$$H = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

On retrouve un code de Hamming.

14 La majoration de Singleton

$$k + d \leq n + 1$$

On dit que le code est MDS (maximum distance separable) s'il y a égalité ; pour les codes de répétition, $d = 3$, $n = 3k$; donc MDS si et seulement si $k = 1$ et $n = 3$.

Démonstration

On introduit E', l'ensemble des éléments x de E tels que

$$x_d = x_{d+1} = ... = x_n = 0$$

On constate que $C \cap E' = \{0\}$.

Autre forme

$$d \leq 1 + n\,(1 - R)$$

Donc pour augmenter la capacité de correction, à taux de transmission donné, il faut nécessairement augmenter n.

Remarque

Il n'y a guère de codes linéaires MDS intéressants si $K = \mathbb{F}_2$; voir les codes de Reed-Solomon.

15 Le code de Golay

On peut remarquer que

$$\binom{23}{0} + \binom{23}{1} + \binom{23}{2} + \binom{23}{3} = 2^{11} = 2048$$

On peut donc conjecturer l'existence d'un code parfait de type $(23, 12, 7)$:

$n = 23$, $k = 12$, $d = 7$, $t = 3$, $R = \frac{12}{23}$; un tel code existe : le code binaire de Golay (voir codes cycliques).

Comparaison avec le code de répétition

On suppose que l'on veut transmettre $12N$ bits ; pour chaque paquet de 3 bits, la probabilité d'une erreur non corrigée est d'environ $3p^2$.

Espérance du nombre d'erreurs : $36N.p^2$

Avec le code de Golay

Pour chaque paquet de 23 bits, la probabilité d'une erreur non corrigée est d'environ $8855.p^4$

Espérance du nombre de blocs incorrects :

$$8855.N.p^4$$

Moins d'erreurs avec un taux de transmission supérieur ; remarquons que ce n'est vrai que si $p \ll 1$.

Le code de Golay ternaire

$$1 + 2\binom{11}{1} + 4\binom{11}{2} = 3^5 = 243$$

Il existe un code de type $(11, 6, 5)$ sur le corps $\mathbb{F}_3$ (code ternaire).