Codes correcteurs d'erreurs

Christophe Ritzenthaler

Sources. Le cours Codes correcteurs de Coste, Paugam et Quarez disponible sur le web à l'adresse `http://agreg-maths.univ-rennes1.fr/documentation/docs/codes.pdf`.
Les livres 'An introduction to error correcting codes with applications' de Vanstone et van Oorschot et 'Error-correcting codes and finite fields' de Pretzel.

Les codes correcteurs ont été introduits pour corriger les erreurs de transmission ou de lecture de données numériques, ou les erreurs survenant au cours de leur inscription sur un support physique (bande, CD) ou encore lorsque les données subissent une altération sur le support de stockage. Voici quelques domaines où ils sont appliqués :

- transmissions spatiales ;
- minitel ;
- codes barres ;
- disque compact et DVD ;
- communications par internet.

1 Codes et distance de Hamming

Les messages transmis sont supposés d´ecoupés en blocs (ou mots) de longueur n écrits avec l'alphabet $\{0,1\}$. Un *code* (binaire) est un sous-ensemble C de l'ensemble $\{0,1\}^n$ de tous les mots possibles. On dit que n est la longueur de C.
La *distance de Hamming* entre deux mots $x = (x_1, ..., x_n)$ et $y = (y_1, ..., y_n)$, que l'on notera $d(x, y)$, est le nombre d'indices i tels que $x_i \neq y_i$. C'est bien une distance sur $\{0,1\}^n$. La *distance minimum* du code C est le minimum des $d(x, y)$ pour x et y des mots différents de C (on suppose que C a au moins 2 mots !). On la notera toujours d.

Exemple 1. *Considère* $C = \{c_0, c_1, c_2, c_3\}$ *avec*

$$c_0 = (00000), \quad c_1 = (10110), \quad c_2 = (01011), \quad c_3 = (11101).$$

C'est un code de longueur 5 *et de distance* $d = 3$.

Le mot $c \in C$ est émis et, après d'éventuelles erreurs de transmission, le mot $r \in \{0,1\}^n$ est reçu. On décode le mot r selon le principe du maximum de vraisemblance, c.-à-d. qu'on le décode comme un mot de C à distance minimum de r. On dit que C est t-correcteur (ou corrige t erreurs) quand toute erreur portant sur au plus t bits est corrigée correctement. On voit donc que le code C est t-correcteur si et seulement si les boules fermées (dans $\{0,1\}^n$ muni de la distance de Hamming) de centres les éléments de C et de rayon t sont disjointes, ou encore si et seulement si la distance minimum d de C vérifie $d \geq 2t + 1$.

Il est souvent difficile de calculer la distance minimale et plus encore de décoder un mot sans structure additionnel c'est pourquoi on préfère travailler avec des codes linéaires.

Remarque 1. *Supposons que $\#C = 2^k$. Le rapport k/n s'appelle le taux de transmission du code. Soit*

$$C(p) = 1 + p\log_2 p + (1-p)\log_2(1-p).$$

Un théorème de Shannon montre que pour une probabilité d'erreur $p < .5$ à chaque bit, il existe un code dont le taux de transfert est inférieur mais arbitrairement proche de $C(p)$ qui peut corriger tout message avec une probabilité arbitrairement proche de 1. Par exemple si $p = .999$, on a $C(p) = .986$ donc en ajoutant à peine 15 bits pour 1000 bits transmis on peut arriver à une correction arbitrairement proche de 1. Bien sûr le théorème de Shannon ne dit pas comment construire de tels codes.

2 Codes linéaires

2.1 Définitions

On note $\mathbb{F}_2$ le corps à deux éléments 0 et 1. Les mots de longueurs n sont les éléments de $\mathbb{F}_2^n$, que l'on écrira comme des vecteurs lignes. Un code linéaire de longueur n est un sous-espace vectoriel $C \subset \mathbb{F}_2^n$. La lettre k désignera toujours la dimension de C (comme espace vectoriel). Le nombre de mots du code C est 2^k. Le *poids* d'un mot $x = (x_1, \ldots, x_n) \in \mathbb{F}_2^n$, noté $w(x)$, est le nombre d'indices i tels que $x_i \neq 0$. Comme $d(x,y) = w(x-y)$, la distance minimum d d'un code linéaire C est le minimum des poids $w(x)$ pour $x \in C$ non nul. (On suppose que C n'est pas le code nul.) On regroupe les trois paramètres n, k et d d'un code linéaire C en disant que C est de type (n,k,d).

Lemme 2.1 ([Dem, Prop.9.1],[Pre, 18.4])**.** *On a toujours $d+k \leq n+1$ (borne de Singleton).*

Démonstration. Soit V le sous espace de $\mathbb{F}_2^n$ engendré par les vecteurs donc les $d-1$ premières coordonnées sont nulles. Considérons le code $C \cap V$. Sa longueur est $n-d+1$. Montrons que sa dimension est toujours k. En effet sinon, cela veut dire qu'il existe une relation linéaire entre les k vecteurs engendrant C à valeur dans un supplémentaire de V. Mais ceci veut dire qu'il existe un élément de C de poids $\leq d-1$, ce qui est exclu. Ainsi on a $k \leq n-d+1$, ce qu'on voulait. □

La borne de Singleton quantifie le fait qu'on ne peut pas avoir à la fois le beurre (une capacité de correction importante) et l'argent du beurre (un nombre de mots de code important), pour une longueur n fixée.

Remarque 2. *Il existe d'autres bornes. Par exemple la borne de Hamming. Si B est la boule de centre 0 et de rayon r dans $\mathbb{F}_2^n$, elle contient $R = \sum_{k=0}^{r} \binom{n}{k}$ mots. Ainsi si C est un code de longueur n et de distance minimale $d = 2r+1$ alors C a au plus $2^n/R$ mots de code.*

Lemme 2.2. *Si C est un code linéaire de type (n,k,d), on définit le code étendu $\tilde{C}$ comme le code formé des mots $(x_1, \ldots, x_{n+1}) \in \mathbb{F}_2^{n+1}$ tels que $(x_1, \ldots, x_n) \in C$ et $\sum_{i=1}^{n+1} x_i = 0$. Le type de $\tilde{C}$ est $(n+1, k, d+1)$ si d est impair et $(n+1, k, d)$ si d est pair.*

Démonstration. Le seul paramètre a déterminé est la distance minimale. Soit x un mot de poids d de C. Si d est pair alors $y = (x, 0) \in \tilde{C}$ et $w(y) = d$. Si d est impair alors $y = (x, 1) \in \tilde{C}$ et $w(y) = d+1$. □

2.2 Matrice génératrice

On peut se donner un sous-espace vectoriel (et donc un code) par une base. Soit C un code linéaire. Une *matrice génératrice* de C est une matrice dont les lignes forment une base de C. Une matrice génératrice G est donc de taille $k \times n$ et de rang k. Si m est un vecteur ligne de $\mathbb{F}_2^k$, le produit mG est un mot du code C et l'application $m \mapsto mG$ est un isomorphisme de $\mathbb{F}_2^k$ sur C (que l'on peut voir comme une opération de codage). Si la matrice G est de la forme (I, P), on dit que le codage est *systématique*. Les k premiers bits d'un mot de code portent l'information (on y recopie le vecteur de $\mathbb{F}_2^k$), les $n-k$ suivants sont de la redondance.

Exemple 2. *La matrice*

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

est une matrice génératrice pour le code

$$(00000), (10000), (11010), (11101), (01010), (01101), (00111), (10111).$$

On dit que deux codes linéaires de même longueur sont *équivalents* si l'un s'obtient à partir de l'autre par une permutation des coordonnées. On peut vérifier que deux codes équivalents ont même type. De plus tout code est équivalent à un code donné par un codage systématique.

Exemple 3. *Soit G la matrice génératrice*

$$\begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix}.$$

On additionne la ligne 1 *aux lignes* 2 *et* 3 *(ceci correspond à un changement de base). On obtient*

$$\begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

Il ne reste plus qu'à permuter les trois premières colonnes.

2.3 Matrice de contrôle

On peut ausi se donner un sous-espace vectoriel par un système d'équations indépendantes. Soit C un code linéaire. Une *matrice de contrôle* de C est la matrice d'un système d'équations linéaires homogènes indépendantes dont l'espace des solutions est C. Autrement dit, une matrice de contrôle H est de taille $(n-k) \times n$ et de rang $n-k$, et $C = \{x \in \mathbb{F}_2^n, H^t x = 0\}$. Si C est donné sous forme systématique par la matrice génératrice $G = (I_k, P)$, alors on peut prendre comme matrice de contrôle $H = (-{}^tP, I_{n-k})$ (le signe $-$ est superflu en caractéristique 2). Supposons que $c \in C$ est le mot du code envoyé et $r \in \mathbb{F}_2^n$ le mot reçu. La différence $e = r - c$ est le vecteur d'erreur. Son poids $w(e)$ est le nombre de bits erronés dans le mot reçu. Soit H une matrice de contrôle de C. Le *syndrome* du mot reçu r est le vecteur $s \in \mathbb{F}_2^{n-k}$ défini par ${}^ts = H^t r = H^t e$. Le syndrome est nul si et seulement si $r \in C$. Le syndrome définit un isomorphisme du quotient $\mathbb{F}_2^n / C$ sur $\mathbb{F}_2^{n-k}$. Si le syndrome est non nul, on corrige le mot reçu r en appliquant le principe du maximum de vraisemblance : on soustrait à r un mot de poids minimum dans sa classe modulo C, c.-à-d. un mot de poids minimum parmi ceux

ayant même syndrome que r. Dans le cas où $w(e)$ est strictement inférieur à $d/2$, alors e est l'unique mot de poids minimum dans la classe de r modulo C et on récupère bien le mot de code émis.

Remarque 3. *La matrice de contrôle peut être vue comme la matrice génératrice du code dual*

$$C^{\perp} = \{y \in \mathbb{F}_2^n, \forall c \in C, y.c = 0\}$$

où . est le produit scalaire usuel.

Proposition 2.1. *Soit H une matrice de contrôle du code C. La distance minimum d de C est caractérisée par les propriétés suivantes :*

- *$d-1$ colonnes de H sont toujours linéairement indépendantes.*
- *Il y a un système de d colonnes de H qui est lié.*

Démonstration. Supposons que $d-1$ colonnes de H sont toujours linéairement indépendantes. Soit $c = (c_1, \ldots, c_n)$ un mot de code. On a $H^t c = 0$. Si $w(c) < d$ alors on a relation entre moins de d colonnes de H. Inversement un mot c de poids d donne une relation entre d colonnes de H. □

Exemple 4. *Donnons un exemple de décodage par syndrome. Soit C le code donné par la matrice de contrôle*

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

On calcule tout d'abord des représentants de poids minimal (appelé leader*) pour chacune des classes $\mathbb{F}_2^6/C$ ainsi que le syndrome associé.*

Leader	*Syndrome*
000000	000
100000	110
010000	101
001000	011
000100	100
000010	010
000001	001
100001	111

Supposons que $u = 100011$ est reçu. Son syndrome est $H^t u = 101$. Pour décoder u il faut donc lui soustraire 010000*.*

3 Quelques codes linéaires

3.1 Code de Hamming

voir TD.

3.2 Codes cycliques

Un code linéaire $C \subset \mathbb{F}_2^n$ est dit *cyclique* quand il est stable par l'automorphisme de décalage cyclique

$$\begin{array}{rcl} T : \mathbb{F}_2^n & \to & \mathbb{F}_2^n \\ (x_1, \ldots, x_n) & \mapsto & (x_2, \ldots, x_n, x_1) \end{array}$$

On identifie $\mathbb{F}_2^n$ à l'algèbre $\mathbb{F}_2[X]/(X^n - 1)$ par

$$(x_1, \ldots, x_n) \mapsto x_1 X^{n-1} + \cdots + x_{n-1} X + x_n.$$

On désigne ici par la même lettre l'indéterminée X et son image dans le quotient. L'endomorphisme T, modulo cette identification, est l'endomorphisme de multiplication par X. Par définition, un code cyclique est un sous-espace vectoriel stable par multiplication par X, et donc par n'importe quel polynôme en X. Donc, un code linéaire C de longueur n est cyclique si et seulement si C est un idéal de $\mathbb{F}_2[X]/(X^n - 1)$.
L'homorphisme de passage au quotient induit une bijection entre l'ensemble des idéaux de $\mathbb{F}_2[X]/(X^n - 1)$ et l'ensemble des idéaux de $\mathbb{F}_2[X]$ qui contiennent $(X^n - 1)$. Puisque $\mathbb{F}_2[X]$ est principal, ce sont exactement les idéaux engendrés par les diviseurs (que l'on prend unitaires pour assurer l'unicité) de $X^n - 1$ dans $\mathbb{F}_2[X]$. Le diviseur unitaire g de $X^n - 1$ ainsi associé à un code cyclique C s'appelle le *polynôme générateur* de C. Si $g \neq X^n - 1$ (dans le cas contraire C est nul), le code C est engendré (comme espace vectoriel sur $\mathbb{F}_2$) par $g, Xg, ..., X^{n-1-\deg(g)}g$. La dimension de C est dans tous les cas $k = n - \deg(g)$.
Le procédé de codage systématique $\mathbb{F}_2^k \to C$ d'un code cyclique de polynôme générateur g est donné par la division euclidienne par g : le vecteur $(x_1, \ldots, x_k) \in \mathbb{F}_2^k$ est codé par le polynôme $c = c_I - c_R$, où $c_I = x_1 X^{n-1} + \ldots + x_k X^{n-k}$, et c_R (de degré $< n - k$) est le reste de la division euclidienne de c_I par g. (Le polynôme c_I porte l'information, et c_R la redondance).
On suppose à partir de maintenant que n est premier avec la caractéristique de $\mathbb{F}_2$, c.-à-d. impair. Cette hypothèse entraîne que le polynôme $X^n - 1$ a n racines distinctes dans son corps de décomposition sur $\mathbb{F}_2$. Notons K ce corps de décomposition, c'est-à-dire le corps engendré par les racines n-ièmes de l'unité sur $\mathbb{F}_2$. On fait le choix d'une racine primitive n-ième de l'unité dans K, que nous noterons α. Le polynôme minimal P de α sur $\mathbb{F}_2$ a pour degré l'ordre r de 2 dans le groupe multiplicatif $(\mathbb{Z}/n\mathbb{Z})^*$, et ses racines sont $\alpha, \alpha^2, \alpha^4, \ldots, \alpha^{2^{r-1}}$. On a $K = \mathbb{F}_2[\alpha] = \mathbb{F}_2[X]/P = \mathbb{F}_2^r$. Le polynôme cyclotomique ϕ_n factorise sur $\mathbb{F}_2$ en produit de facteurs irréductibles de degré r (par exemple $\phi_{15} = X^8 - X^7 + X^5 - X^4 + X^3 - X + 1$ se factorise en $(X^4 + X + 1)(X^4 + X^3 + 1)$ sur $\mathbb{F}_2$), et P est un de ces facteurs. Le polynôme générateur g va être déterminé par ses racines dans K, qui forment un sous-ensemble de l'ensemble $\{1, \alpha, \alpha^2, \ldots, \alpha^{n-1}\}$ des racines n-ièmes de l'unité (les zéros du code). Soit Σ un sous-ensemble de $\mathbb{Z}/n\mathbb{Z}$. Le polynôme $g_\Sigma = \prod_{i \in \Sigma}(X - \alpha^i)$ est un diviseur de $X^n - 1$ à coefficients dans $\mathbb{F}_2$ si et seulement si Σ est stable par multiplication par 2 (se souvenir que $\mathbb{F}_2$ est l'ensemble des éléments de K laissés fixes par l'élévation au carré). En conclusion, on a une bijection entre les codes cycliques de longueur n et les sous-ensembles de $\mathbb{Z}/n\mathbb{Z}$ stables par multiplication par 2. La configuration des racines du polynôme générateur nous renseigne sur la distance minimale du code cyclique.

Proposition 3.1 ([Dem, Prop.9.4])**.** *Si Σ contient s entiers consécutifs $a+1, a+2, \ldots, a+s$ modulo n, alors le code cyclique de polynôme générateur g_Σ est nul ou a une distance minimum supérieure ou égale à $s+1$.*

La démonstration est une application des propriétés du déterminant de Vandermonde.

3.3 Codes BCH

Les codes BCH (Bose-Chaudhuri-Hocquenghem) sont des codes cycliques particuliers. La famille des codes BCH contient les codes de Reed-Solomon qui servent pour la lecture des CD (voir [Dem, p.238]). Nous ne considérerons ici que des codes BCH binaires primitifs. Leur longueur n est de la forme $n = 2^r - 1$. Alors r est l'ordre de 2 dans le groupe multiplicatif $(\mathbb{Z}/n\mathbb{Z})^*$, et le corps K engendré par les racines n-ièmes de l'unité sur $\mathbb{F}_2$ est $\mathbb{F}_2^r$. Tous les calculs de décodage vont se faire sur ce corps K. Choisissons une racine primitive n-ième de l'unité α dans K. Concrètement, on se donne α par son polynôme minimal P sur $\mathbb{F}_2$. Tout élément du groupe multiplicatif K^* s'écrit de manière unique sous la forme α^i avec $0 \leq i < n$, et il s'écrit aussi de manière unique comme combinaison linéaire à coefficients dans $\mathbb{F}_2$ de $1, \alpha, \ldots, \alpha^r - 1$. On peut voir la table de correspondance entre ces deux représentations pour $K = \mathbb{F}_{16}$, avec α vérifiant $\alpha^4 + \alpha + 1 = 0$ dans [Dem, p.213].
On appelle *code BCH* de longueur $n = 2r - 1$ et de distance prescrite δ (δ entier tel que $0 < \delta \leq n$) le code cyclique de polynôme générateur g_Σ, où Σ est le plus petit sous-ensemble de $\mathbb{Z}/n\mathbb{Z}$ contenant $1, 2, \ldots, \delta - 1$ et stable par multiplication par 2. Autrement dit, un polynôme $c = x_1 X^{n-1} + \cdots + x_n \in \mathbb{F}_2[X]$ appartient à ce code si et seulement si

$$c(\alpha) = c(\alpha^2) = ... = c(\alpha^{\delta-1}) = 0.$$

On peut trouver des exemples de codes BCH explicites pour $n = 15$ dans [Dem, p.240].

Références

[Dem] M. Demazure, Cours d'algèbre, Cassini 1997.

[Pre] O. Pretzel, Error-correcting codes and finite fields, Clarendon Press, 1992.