Théorie des codes détecteurs et correcteurs d'erreurs

par Justin Vast

Classification des corps finis

Théorème. *Le cardinal d'un corps fini est une puissance d'un nombre premier. Réciproquement, étant donné p^k une puissance d'un nombre premier, il existe un unique corps, à isomorphisme près, de cardinalité p^k.*

Notation. $\mathbb{F}_{p^k}$ *désigne le corps de décomposition de* $X^{p^k} - X$ *sur* $\mathbb{F}_p$*, et possède* p^k *éléments.*

Nous utiliserons principalement le corps $\mathbb{F}_2$ à deux éléments (notés 0 et 1), dont les tables d'addition et de multiplication sont les suivantes :

$+$	0	1
0	0	1
1	1	0

$\cdot$	0	1
0	0	0
1	0	1

Groupe des inversibles d'un corps fini

Théorème. *L'ensemble des inversibles du corps* $\mathbb{F}_q$*, noté* $\mathbb{F}_q^\times$*, muni de la multiplication est un groupe cyclique ; cela signifie qu'il existe* $\alpha \in \mathbb{F}_q$ *tel que*

$$\langle \alpha \rangle := \{\alpha^n | n \in \mathbb{Z}\} = \mathbb{F}_q^\times$$

Remarque. Dans un corps, être inversible = être non nul.

Exemples : $\mathbb{F}_2^\times = \{1\} = \langle 1 \rangle$, $\mathbb{F}_4^\times = \langle \alpha \rangle$ où $\alpha \in \mathbb{F}_4^\times \backslash \{1\}$, $\mathbb{F}_8^\times = \langle \beta \rangle$ où $\beta \in \mathbb{F}_8^\times \backslash \{1\}$,

$\mathbb{F}_{16}^\times = \langle \gamma \rangle$ où $\gamma \in \mathbb{F}_{16}^\times$ est tel que $\gamma^4 + \gamma + 1 = 0$ ou $\gamma^4 + \gamma^3 + 1 = 0$,

$\mathbb{F}_7^\times = \langle 3 \rangle$,...

Vocabulaire

Définition.

- *Un **alphabet** $\mathcal{A}$ est un ensemble non-vide de symboles sans signification individuelle. Dans la pratique, on exige que l'alphabet soit fini.*
- *Un **code** $\mathcal{C}$ de **longueur** $n \in \mathbb{N}^{>0}$ est un sous-ensemble non-vide de $\mathcal{A}^n$ dont les éléments sont appelés **mots**.*
- *Le cardinal d'un code $\mathcal{C} \subseteq \mathcal{A}^n$ est appelé **taille du code**.*
- *Si $\mathcal{A} = \mathbb{F}_q$, où $q = p^k$, un sous-espace vectoriel de $\mathbb{F}_q^n$ est appelé **code linéaire** de longueur n.*

Code barre : exemple avec EAN-13

EAN : *European Article Number*

Ce code permet notamment d'identifier les livres via leur ISBN (*International Standard Book Number*).

EAN-13 est un code barre à 13 chiffres (décimaux).

Le dernier chiffre est un *chiffre de contrôle*, et permet de s'assurer dans la pratique que le code a été bien scanné. Un mot $a_1a_2...a_{13}$ appartient au code si et seulement si

$$a_1 + 3a_2 + a_3 + 3a_4 + \cdots + 3a_{12} + a_{13} \equiv 0 \pmod{10}$$

(Exemple : ISBN 9 781441 998538)

Le code ASCII

ASCII : *American Standard Code for Information Interchange*

C'est une norme informatique permettant de coder $128 (= 2^7)$ caractères, numérotés de 0 à 127.

7 bits nécessaires pour représenter ces caractères.

Ex: 0000110, 0101010

On ajoute généralement un huitième bit, appelé bit de parité.

Ex: 00001100, 01010101

$\mathcal{A} = \{0, 1\} = \mathbb{F}_2$, et $\mathcal{C}$ est un code linéaire, car c'est un $\mathbb{F}_2$-sous-espace vectoriel de $\mathbb{F}_2^8$ de dimension 7.

Plus précisément, c'est le noyau de l'application linéaire

$$\mathbb{F}_2^8 \longrightarrow \mathbb{F}_2 \colon (a_i)_{1 \le i \le 8} \longmapsto \sum_{i=1}^{8} a_i$$

Avantage du bit de parité

Le bit de parité donne de l'information redondante.

Supposons qu'Alice veuille envoyer le caractère ASCII 10010101 à Bob.

Malheureusement, l'information est mal transmise et Bob reçoit le mot 10011101.

Puisque la somme des bits (dans $\mathbb{F}_2$) vaut 1, Bob peut *détecter* une erreur.

C'est un exemple de *code détecteur d'erreurs*.

Limites pratiques du bit de parité

Bob ne peut pas identifier la position du bit erroné.

Si lors de l'envoi, deux bits sont erronés (ex: 00011101), alors Bob ne détectera aucune erreur et interprétera mal le message d'Alice.

Ex:

Alice envoie : Hey Boby!

Bob reçoit : Hey Baby!

Autre exemple de code

Un autre exemple est le dénommé code de répétitions $\mathcal{C} = \{(a, a, \cdots, a) \mid a \in \mathcal{A}\} \subseteq \mathcal{A}^n$, pour $\mathcal{A}$ un alphabet.

Si $\mathcal{A} = \mathbb{K}$ est un corps (fini), alors $\mathcal{C}$ est un code linéaire, car c'est un sous-espace vectoriel de $\mathbb{K}^n$ de dimension 1. Le code $\mathcal{C}$ est le noyau de l'application linéaire

$$\mathbb{K}^n \longrightarrow \mathbb{K}^{n-1} \colon (a_i)_{1 \leq i \leq n} \longmapsto (a_n - a_i)_{1 \leq i \leq n-1}$$

Prenons $n=3$ et $\mathcal{A}=\mathbb{F}_{11}$ pour l'exemple.

No envoie le mot $000\in\mathcal{C}$ à Bond, et Bond reçoit 007. Bond détecte une erreur car $007\notin\mathcal{C}$. De plus, en supposant qu'au plus une erreur ait été commise, il peut récupérer le message original. On dit que le code corrige 1 erreur.

C'est un premier exemple de code correcteur d'erreurs.

Pour n et $\mathcal{A}$ arbitraires, ce code $\mathcal{C}$ corrige $\left\lfloor \frac{n-1}{2} \right\rfloor$ erreurs, voyez-vous pourquoi ?

Autres codes

Code QR (Quick Response Code)/Code Aztec/Data Matrix/PDF417/...

Codes de Reed-Solomon (codage des CDs) (code sous-jacent aux codes QR)

Codes de Hamming

Codes BCH

Codes de Goppa (codes cycliques)

Codes de Golay

Codes de Hadamard

Codes LDPC

...

Distance de Hamming

Définition. *On définit une distance sur l'ensemble $\mathcal{A}^n$, appelée* ***distance de Hamming****, telle que pour $x = (x_i), y = (y_i) \in \mathcal{A}^n$*

$$d(x, y) := \#\{i \in \{1, \cdots, n\} \,|\, x_i \neq y_i\}$$

Si $\mathcal{A} = \mathbb{F}_q$, on définit le ***poids*** *de $x \in \mathbb{F}_q^n$ par $w(x) := d(x, 0_{\mathbb{F}_q^n})$.*

En d'autres termes, $d(x, y)$ mesure le nombre de coordonnées distinctes entre x et y.

Exemple pour $\mathcal{A} = \mathbb{F}_3$ et $\mathcal{C} = \mathbb{F}_3^2$:

$$d(01, 21) = 1 \text{ et } B[00, 1] = \{00, 01, 02, 10, 20\}$$

Décodage par vraisemblance maximale

Soit un mot $z \in \mathcal{A}^n$ (reçu). On définit son décodage par vraisemblance maximale comme étant l'unique mot $x \in \mathcal{C}$ (s'il existe) qui maximise la probabilité

$$P(z \,\text{reçu} \,|\, x \,\text{envoyé})$$

Méthodes de décodage

Décodage par distance minimale

Soit un mot $z \in \mathcal{A}^n$ (reçu). On définit son décodage par distance minimale comme étant l'unique mot $x \in \mathcal{C}$ (s'il existe) qui minimise

$$d(x\,,\,z)$$

Dans la plupart des cas, les deux stratégies de décodage présentées ici coïncident.

Exercice.

Soient $\mathcal{C} \subseteq \mathcal{A}^n$ un code, $z \in \mathcal{A}^n$ un mot reçu, et $x, y \in \mathcal{C}$. Supposons que la probabilité qu'il se produise une erreur de transmission à la coordonnée i d'un mot soit $p < \frac{1}{2}$ indépendamment de i, et que ces évènements soient indépendants. Prouver que

$$d(x, z) < d(y, z) \Longleftrightarrow P(z \text{ reçu} \mid x \text{ envoyé}) > P(z \text{ reçu} \mid y \text{ envoyé})$$

Vocabulaire et lemme

Définition. *Soit un code $\mathcal{C} \subseteq \mathcal{A}^n$.*

- *On appelle* ***distance minimale du code*** *le nombre*

$$d(\mathcal{C}) := \min \{d(x, y) \mid x, y \in \mathcal{C}, x \neq y\}$$

- *On dit que le code est de type $(n, m, d)_q$, ou simplement (n, m, d), s'il est de longueur n, de taille $m = |\mathcal{C}|$, de distance minimale d, avec $|\mathcal{A}| = q$.*

Lemme. *Si $\mathcal{C} \subseteq \mathbb{F}_q^n$ est un code linéaire, alors*

$$d(\mathcal{C}) = \min \{w(x) | x \in \mathcal{C} \setminus \{0\}\}$$

Lemmes de détection et correction d'erreurs

Lemme. (de détection) *Soit* $\mathcal{C}$ *un code de type* (n, m, d). *Pour tous* $x \in \mathcal{C}$ *et* $y \in \mathcal{A}^n$, *si* $0 < d(x, y) < d$, *alors* $y \notin \mathcal{C}$. *On dit que le code détecte* $d - 1$ *erreurs.*

Lemme. (de correction) *Soit* $\mathcal{C}$ *un code de type* (n, m, d). *Pour tout* $y \in \mathcal{A}^n$, *s'il existe un* $x \in \mathcal{C}$ *tel que*

$$d(x, y) \leq \left\lfloor \frac{d-1}{2} \right\rfloor$$

alors x *est l'unique mot du code* $\mathcal{C}$ *à posséder cette propriété. On dit que le code corrige* $\left\lfloor \frac{d-1}{2} \right\rfloor$ *erreurs.*

Définition. *Soit* $\mathcal{C} \subseteq \mathbb{F}_q^n$ *un code linéaire de dimension* k*.*

On dit que $M \in \mathcal{M}(n \times k, \mathbb{F}_q)$ *est une* ***matrice génératrice*** *du code* $\mathcal{C}$ *si ses colonnes forment une base de* $\mathcal{C}$ *en tant que* $\mathbb{F}_q$*-espace vectoriel.*

On dit que $A \in \mathcal{M}((n-k) \times n, \mathbb{F}_q)$ *est une* ***matrice de contrôle*** *du code* $\mathcal{C}$ *si*

$$\mathcal{C} = \operatorname{Ker} A$$

Proposition. *Soit $\mathcal{C}$ un code linéaire, et $A \in \mathcal{M}((n-k) \times n, \mathbb{F}_q)$ une matrice de contrôle de $\mathcal{C}$.*

La distance minimale d'un code linéaire $\mathcal{C}$ est la plus petite quantité de colonnes nécessaires pour former un ensemble linéairement dépendant.

Exemple. Soit $q = p^k > 3$, et soit $\alpha \in \mathbb{F}_q^\times$ un générateur du groupe cyclique $\mathbb{F}_q^\times$. Alors le code linéaire dont une matrice de contrôle est A a une distance minimale valant 3 :

$$A = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \alpha & \alpha^2 & \cdots & \alpha^{q-2} \end{pmatrix}$$

Exemple plus sophistiqué : Codes de Reed-Solomon

Soient $q = p^k > 3$, et $r \in \mathbb{N}^{>0}$ tels que $r < q - 2$, et soit α un générateur de $\mathbb{F}_q^\times$. La matrice

$$A = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 & \cdots & 1 \\ 1 & \alpha & \alpha^2 & \cdots & \alpha^r & \cdots & \alpha^{q-2} \\ 1 & \alpha^2 & \alpha^4 & \cdots & \alpha^{2r} & \cdots & \alpha^{2(q-2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & \alpha^r & \alpha^{2r} & \cdots & \alpha^{r^2} & \cdots & \alpha^{r(q-2)} \end{pmatrix}$$

est une matrice de contrôle d'un code linéaire dont la distance minimale vaut $r + 2$.

Il s'agit d'un **code de Reed-Solomon**.

Prenons $q = 7$, $r = 3$, et $\alpha = 3 \in \mathbb{F}_7^\times$ (qui est un générateur de $\mathbb{F}_7^\times$). Alors

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 3^2 & 3^3 & 3^4 & 3^5 \\ 1 & 3^2 & 3^4 & 3^6 & 3^8 & 3^{10} \\ 1 & 3^3 & 3^6 & 3^9 & 3^{12} & 3^{15} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 2 & 6 & 4 & 5 \\ 1 & 2 & 4 & 1 & 2 & 4 \\ 1 & 6 & 1 & 6 & 1 & 6 \end{pmatrix}$$

est une matrice de contrôle d'un code de Reed-Solomon (un $\mathbb{F}_7$-espace vectoriel) dont la distance minimale vaut 5 ; celui-ci corrige donc 2 erreurs.

Exercice

Prouver que la distance minimale du code linéaire dont une matrice de contrôle est A vaut 4. Prouver également que la matrice M est une matrice génératrice du code.

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{pmatrix} \quad M = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Algorithme

Entrée : Un code linéaire $\mathcal{C} \subseteq \mathbb{F}_q^n$ du type (n, q^k, d) et un mot $z \in \mathbb{F}_q^n$
Sortie : Correction de z par distance minimale

$\min := n + 1$
FOR $w \in \mathcal{C}$ DO
  IF $d(w, z) < \min$ DO
    $x := w$
    $\min := d(w, z)$
RETURN x

La complexité de cet algorithme est en $O(n q^k)$ (assez mauvais).

Syndrome

Définition. *Pour $v \in \mathbb{F}_q^n$, le vecteur $s(v) := Av \in \mathbb{F}_q^{n-k}$ est appelé **syndrome** de v, où A est une matrice de contrôle de $\mathcal{C}$.*

Remarque. L'application $s\colon \mathbb{F}_q^n \longrightarrow \mathbb{F}_q^{n-k}$ est linéaire, et $\operatorname{Ker} s = \mathcal{C}$.

Si $z = x + e$ (avec $x \in \mathcal{C}$), alors $s(z) = s(e)$, donc le syndrome du message reçu est le syndrome des erreurs commises.

Exemple. Considérons le code de Hamming (de type $(7, 2^4, 3)_2$), dont une matrice de contrôle est

$$A = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix} \in \mathcal{M}(3 \times 7, \mathbb{F}_2)$$

et supposons que nous ayons reçu le mot $z = 0111010$. Son syndrome vaut $s(z) = Az = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$. On en déduit que ce mot n'appartient pas au code de Hamming. Si au plus une erreur a été commise, c'est le 3e bit qui est erroné, donc la correction de z est 0101010.

Prouver que la matrice M suivante est une matrice génératrice du code de Hamming (pour rappel, la matrice A est une matrice de contrôle de ce code) :

$$A = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix} \quad M = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

On considère de nouveau le code $\mathcal{C}$ dont une matrice de contrôle est $\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}$. Corriger le message $m = 10110011$ (voir m comme un vecteur colonne) si l'on suppose qu'au plus une erreur ait été commise.

On a $Am = 1000$, qui est exactement la 1e colonne de A ; autrement dit, l'erreur se trouve en première position et le message original était 00110011. Qu'en est-il si l'on suppose qu'au plus deux erreurs aient été commises ? Et pour trois erreurs ?

On reçoit cette fois-ci le message $m = 11001010$. Est-il possible de le corriger ?