Codes linéaires

Codes correcteurs d'erreurs

La théorie d'information concerne les transformations permettant la transmission efficace, fiable et sécurisée de l'information. Les codages correcteurs d'erreurs sont les transformations avec rédondance qui protègent l'information contre pertes lors d'une transmission sur un canal bruyant. Le canal de transmission peut être l'atmosphère lors d'une émmission par un antenne radio ou un satellite, le lecteur d'un CD, DVD ou disque dur, ou le disque ou matériel de stockage de mémoire lui-même — transmettant l'information à une date ultérieure avec dégradation pendant le temps.

On va s'intéresser particulièrement aux codes linéaires, qui sont des sous-espaces vectoriels d'un espace vectoriel $\mathbb{F}_q^n$, dont le corps fini $\mathbb{F}_q$ est l'alphabet du code.

Représentation de l'information. Avant de présenter les codes linéaires, il faut introduire de la terminologie pour la représentation de l'information.

Définition. Un *alphabet* $\mathcal{A}$ est un ensemble fini. On appelle leurs éléments des lettres (ou caractères ou symboles). Un *mot* est une suite finie de lettres d'un alphabet. L'ensemble de mots de longueur n s'écrit $\mathcal{A}^n$, et l'union disjointe

$$\mathcal{A}^* = \bigcup_{n \geq 0} \mathcal{A}^n.$$

Un *codage* est une application

$$C : \mathscr{M} \longrightarrow \mathcal{A}^*,$$

d'un ensemble $\mathscr{M}$, permettant d'écrire des éléments de $\mathscr{M}$ en mots de l'alphabet $\mathcal{A}$. On dit que le codage est *sans pertes* s'il est injectif (terminologie spécifique à la théorie des codes). Un *code* est l'image $\mathcal{C} = C(\mathscr{M})$ d'un codage, et un mot de code de $\mathcal{C}$ est un élément.

Exemples. L'alphabet classique $\{\mathtt{A}, \ldots, \mathtt{Z}\}$ réduit est souvant utilisé en cryptographie classique, mais en ajoutant les lettres miniscules, des espaces et pontuation, etc., on arrive à un alphabet $\mathcal{A}$ permettant le codage sans pertes de la plupart des langues européenes occidentales dans $\mathcal{A}^*$. L'alphabet binaire $\{0, 1\}$ des *bits* est l'alphabet de base pour les ordinateurs modernes. Les octets $\{0, 1\}^8$ constitue un autre alphabet. Les codages ASCII ou ISO-8859-1, permettent d'encoder un alphabet occidental dans les 256 octets. Dans la théorie des codes correcteurs d'erreurs va s'orienter vers $\mathcal{A} = \mathbb{F}_q$, où souvant $q = 2$.

Remarque 1. Une *langue* est un sous-ensemble $\mathscr{M}$ de $\mathcal{A}^*$, dont des éléments s'appellent des *messages* ou *textes*, qui peut être équipée d'une fonction de probabilité, ce qui permet de différencier plusieurs langues dans le même ensemble de mots $\mathcal{A}^*$. Les messages d'une

langue sont justes les mots avec probabilité non nul. Un message d'une langue, peut être le codage d'un mot (dans le sens familier), d'une phrase, ou d'un roman tout entier d'une langue naturelle.

Remarque 2. En cryptographie on dit *chiffrement* pour un codage qui transforme les messages en forme dont leur information est cachée. Le domaine d'un chiffrement est l'espace $\mathscr{M}$ de messages ou de textes clairs et le codomaine $\mathcal{C}$ l'espace des textes chiffrés pour l'ensemble des mots de code. L'utilisation du mot *espace* évoque le fait que $\mathscr{M}$ et $\mathcal{C}$, qui sont de fois égaux en tant qu'ensembles, mais se différencient par la distribution de leur éléments, portant des structures d'espaces de probabilité.

Exemples des codes correcteurs d'erreurs

Avant d'introduire des codes formellement, on donne quelques exemples de codages permettant d'ajouter de la redondance lors d'une transmission. Dans ces exemples, tous linéaires, l'alphabet est binaire : $\mathcal{A} = \mathbb{F}_2$. Pour concision, on écrit un vecteur de l'espace vectoriel $\mathbb{F}_2^k$ par $x_1x_2\dots x_k$ et son image $c = C(x)$, par $c_1c_2\dots c_n$, au lieu d'un forme de vecteur $(x_1, x_2, \dots, x_k)$ et $(c_1, c_2, \dots, c_n)$. On note le *rendement* $R = k/n$ du code

$$C : \mathbb{F}_2^k \longrightarrow \mathbb{F}_2^n,$$

qui est le nombre k de bits d'information par rapport au nombre de bits n utilisés pour l'écriture d'un mot de code. Dans les exemples suivants on a $k = 8$, et on donne l'image d'un message type

$$m = 01101110 \in \mathbb{F}_2^8.$$

Exemple 1. Un codage de répétition est l'application qui double le nombre de bits par répétition de chaque bit. La répétition peut être par bit 0011110011111100 ou de l'octet tout entier 0110111001101110. Le rendement est donc $R = 1/2$.

Exemple 2. Au lieu de répéter tous les bits on peut garder un rendement plus élévé en ajoutant un bit de parité (= somme des bits dans $\mathbb{F}_2$) après chaque couple de bits : $C(01101110) = \underline{01}1|\underline{10}1|\underline{11}0|\underline{10}1 = 011101110101 \in \mathbb{F}_2^{12}$. Le rendement est donc $R = 2/3$.

Exemple 3. On peut construire un codage en combinant un bit de parité (par couple) avec répétition du deuxième bit :

$$C(01101110) = \underline{01}11|\underline{10}10|\underline{11}01|\underline{10}10 = 0111101011011010 \in \mathbb{F}_2^{16}.$$

Le rendement est de nouveau $R = 1/2$.

En quel sens peut-on dire qu'on obtient une meilleure séparation des mots de codes ?

Exemple 4. Pour faire mieux, on regarde des blocs de quatre bits, ajoutant un bit de parité pour la somme de chaque trois sur quatre entre eux :

$$C(01101110) = \underline{0110}0110|\underline{1110}1000 = 0110011011101000 \in \mathbb{F}_2^{16}$$

Exemple 5. Est-ce qu'on peut faire mieux ? Les paramétres n (longueur), k (dimension) et d (distance minimum) sont des invariants importants d'un code. Pour des valeurs $[n, k]$

données on veut maximiser la valeur la distance minimum. On peut montrer que les paramétres $[n, k, d]$ sont respectivement

$$\textbf{1.}\ [16, 8, 2], \quad \textbf{2.}\ [16, 12, 2], \quad \textbf{3.}\ [16, 8, 2], \quad \text{et} \quad \textbf{4.}\ [16, 8, 4].$$

Néanmoins, on peut montrer qu'il existe un code linéaire avec paramétres $[16, 8, 5]$.

Distance de Hamming

Définition. La *distance de Hamming* est une fonction $d : \mathcal{A}^n \times \mathcal{A}^n \longrightarrow \mathbb{N}$, définie par

$$d(x, y_) = \left|\{i \ : \ 1 \leq i \leq n, x_i \neq y_i\}\right|$$

où $x = x_1 x_2 \dots x_n$ et $y = y_1 y_2 \dots y_n$. La *distance minimum* d'un codage $C : \mathscr{M} \to \mathcal{A}^n$ est

$$d(C) = \min\left(\{d(x, y) \ : \ x, y \in \mathscr{M}, x \neq y\}\right).$$

et la *distance minimum* d'un code $\mathcal{C} \subseteq \mathcal{A}^n$ est $d(\mathcal{C}) = \min\left(\{d(x, y) \ : \ x, y \in \mathcal{C}, x \neq y\}\right)$.

Remarque. Si un codage est sans pertes on a donc $d(C) = d(C(\mathscr{M}))$, sinon $d(C) = 0$.

Propriétés.

1. La distance de Hamming satisfait l'inégalité triangulaire $d(x, y) \leq d(x, z) + d(z, y)$ pour tout x, y, et $z \in \mathcal{A}^n$.
2. Ensemble avec les autres propriétés évidentes : de symétrie ($d(x, y) = d(y, x)$) et de séparation ($d(x, y) = 0$ si et seulement si $x = y$), la distance de Hamming est, en effet, une distance sur $\mathcal{A}^n$.
3. Si $\mathcal{A}$ est un anneau (e.g. $\mathcal{A} = \mathbb{F}_q$) alors la distance est linéaire :

   $$d(x + z, y + z) = d(x, y).$$

   Dans ce cas, il existe un élémement distingué 0 de $\mathcal{A}$, et on définit le *support* d'un élément $x \in \mathcal{A}^n$ d'être les indices des coordonnées non nuls :

   $$\mathrm{Supp}(x) = \{i \ : \ 1 \leq i \leq n, x_i \neq 0\}$$

   et le *poids* d'être son cardinal : $\|x\| = d(x, \mathbf{0}) = |\mathrm{Supp}(x)|$ où $\mathbf{0} \in \mathcal{A}^n$ est le mot des zéros de longueur n.
4. Par conséquent de linéarité, si $\mathcal{C}$ est un code linéaire (satisfaisant $ax + by \in \mathcal{C}$ pour tous $a, b \in \mathcal{A}$ et $x, y \in \mathcal{C}$), on a $d(\mathcal{C}) = \min\left(\{\|x\| \ : \ x \in \mathcal{C}\}\right)$.

La distance de Hamming permet de définir des boules et sphères comme pour tout espace métrique.

Définition. La *boule de rayon t* autour de $x \in \mathcal{A}^n$ est

$$B(x, t) = \{y \in \mathcal{A}^n \ : \ d(x, y) \leq t\},$$

et la sphère de rayon t est

$$S(x,t) = \{y \in \mathcal{A}^n \,:\, d(x,y) = t\}.$$

Par conséquent, la boule $B(x,t)$ est l'union disjointe des sphères $S(x,i)$ pour $0 \leq i \leq t$. On peut également définir une boule (ou voisinage) de rayon t autour de $\mathcal{C} \subset \mathcal{A}^n$:

$$B(\mathcal{C},t) = \bigcup_{x \in \mathcal{C}} B(x,t).$$

Attention. Malgré le fait que $d(x,y)$ est une vraie distance, la structure des boules est potentiellement contre-intuitive : Si $A = \mathbb{F}_q$, l'espace $\mathcal{A}^n$ est un espace vectoriel, dans lequel la boule $B(x,1)$ est l'union des droites affines parallèles aux axes de $\mathbb{F}_q^n$ passant par x, la boule $B(x,2)$ est une union de plans, etc.

Exemple. Soit $\mathcal{C} = \{c_0, c_1, c_2, c_3\} = \{0000, 10110, 01011, 11101\} \subset \mathbb{F}_2^5$. En tabulant les distances entre mots de code :

$d(x,y)$	c_0	c_1	c_2	c_3
c_0	0	3	3	4
c_1	3	0	4	3
c_2	3	4	0	3
c_3	4	3	3	0

on trouve que la distance minimum est 3. Or, en vérifiant que le code est linéaire (du fait que $c_3 = c_1 + c_2$) il suffit de calculer les poids des trois éléments non nuls :

$$\big(\|c_1\|, \|c_2\|, \|c_3\|\big) = (3,3,4).$$

Codes linéaires

Si $\mathcal{A} = \mathbb{F}_q$, on dit que $\mathcal{C} \subseteq \mathcal{A}^n = \mathbb{F}_q^n$ est un *code linéaire* si $\mathcal{C}$ est un sous-espace vectoriel. La *dimension* de $\mathcal{C}$ est sa dimension en tant qu'espace vectoriel, déterminée par la relation $|\mathcal{C}| = q^k$. Pour un code en bloc dans $\mathcal{A}^n$ pour le préciser en général il faut l'énumération des ces éléments, mais un code linéaire est déterminé par une base de k éléments. La détermination de la distance minimum d'un code coût plus cher, mais on a vu que l'algorithme naïf se réduit au calcul de $|\mathcal{C}|(|\mathcal{C}|-1)/2$ distances à $|\mathcal{C}|-1$ poids pour un code linéaire.

Par un choix de base de $\mathcal{C}$, un code linéaire est donc l'image d'un codage $C : \mathbb{F}_q^k \to \mathbb{F}_q^n$. Une matrice G dont les lignes forment une base s'appelle une *matrice génératrice* de $\mathcal{C}$. Le codage est donné par $C(x) = x.G$.

On peut préciser $\mathcal{C}$ comme noyau d'une transformation linéaire : $\pi : \mathbb{F}_q^n \to \mathbb{F}_q^{n-k}$. La matrice H d'une telle transformation, telle que $\pi(c) = c.H^t$, s'appelle une *matrice de contrôle* (ou matrice de parité). Les lignes de H forment une base du *code dual* $\mathcal{C}^\perp$ de $\mathcal{C}$, de dimension $n-k$, tel que $\langle \mathcal{C}, \mathcal{C}^\perp \rangle = 0$. On a donc une suite exacte

$$0 \longrightarrow \mathbb{F}_q^k \xrightarrow{C} \mathbb{F}_q^n \xrightarrow{\pi} \mathbb{F}_q^{n-k} \longrightarrow 0.$$

Un élément de l'image de π s'appelle un *syndrôme*, qui measure l'erreur par rapport à un mot de code : si $r = c + e$ est un mot de $\mathbb{F}_q^n$, avec $c \in \mathcal{C}$, le vecteur e est un vecteur d'erreur par rapport à c et le syndrôme $s = \pi(r) = \pi(e)$, ne dépend que de e.

La matrice génératrice G est en forme *systématique* si $G = [\, I_k \,|\, P \,]$, pour une matrice P de $k \times (n-k)$, telle que l'application $\mathbb{F}_q^k \to \mathbb{F}_q^n$, est donnée par l'identité sur les k premières coordonnées :

$$x_1 x_2 \dots x_k \longmapsto x_1 x_2 \dots x_k \dots n_n.$$

On peut construire une matrice de contrôle $H = [\, -P^t \,|\, I_{n-k} \,]$, telle que

$$G.H^t = I_k.(-P) + P.I_{n-k} = -P + P = 0.$$

Pour un codage linéaire en forme systématique, l'opération de codage comporte le prolongement d'un vecteur par l'ajout de caractères nouveaux. Le décodage d'un mot de code (sans erreur) est particulièrement façile : il suffit de projécter sur les premiers k caracterères.

Néanmoins, en générale l'inversion du codage (sans pertes) $C : \mathscr{M} \to \mathcal{A}^n$, restreint à son image $\mathcal{C}$, est une application $D : \mathcal{C} \to \mathscr{M}$ bijective, efficace et transparente. Par *décodage*, on cherche à prolonger D à un voisinage $B(\mathcal{C}, t)$ de $\mathcal{C}$, quitte à le prolonger à l'espace ambient $\mathcal{A}^n$ entier :

$$\mathcal{A}^n \dashrightarrow \mathcal{C} \xrightarrow{D} \mathscr{M}.$$

Comme l'application D est une bijection, on se focalise sur la fonction $\mathcal{A}^n \dashrightarrow \mathcal{C}$.

Décodage : correction et détection d'erreurs

Soit $\mathcal{C} \subset \mathcal{A}^n$ un code, $c \in \mathcal{C}$ un mot transmis lors d'une communication sur un canal bruyant et soit $r \in \mathcal{A}^n$ le mot reçu. On veut dévélopper des conditions pour lesquelles on peut identifier c comme le mot de code plus proche et conditions pour lesquelles on peut détecter que r contient trop d'erreurs pour être corrigé.

Supposer que $|\mathcal{A}| = q$ et $x \in \mathcal{A}^n$. On note que le cardinal $|B(x,t)|$ ne dépend que de q, n et t, et on ne note $V(q,n,t)$. En particulier, on a l'égalité

$$V(q,n,t) = \sum_{i=0}^{t} |S(x,i)| = \sum_{i=0}^{t} \binom{n}{i} (q-1)^i.$$

Définition. On dit qu'un code $\mathcal{C}$ est *s-détecteur* si $B(c,s) \cap \mathcal{C} = \{c\}$ pour tout $c \in \mathcal{C}$, et que $\mathcal{C}$ est *t-correcteur* (d'erreurs) si les boules de rayon t autour de mots distincts de $\mathcal{C}$ sont disjointes.

Remarque. On en déduit qu'un code est s-détecteur si est seulement si $d(\mathcal{C}) \geq s+1$, et que $\mathcal{C}$ est t-correcteur d'erreurs, si est seulement si $d(\mathcal{C}) \geq 2t+1$, si est seulement si

$$|B(\mathcal{C}, t)| = |\mathcal{C}| V(q,n,t).$$

Cette dernière relation est l'expression que $B(\mathcal{C}, t)$ est l'union disjointes des boules $B(c, t)$.

Stratégies de décodage

On suppose que $\mathcal{C} \subset \mathcal{A}^n$ est un code, $d(\mathcal{C})$ sa distance minimum et $s, t \in \mathbb{N}$ tels que $d(\mathcal{C}) = 2t + s + 1$. Il suit que $\mathcal{C}$ est t-correcteur.

Stratégie détection maximale. Cette strategie consiste à identifier l'existence d'erreurs de transmission si le mot reçu $r \notin \mathcal{C}$, sans tentative de correction d'erreurs. En cas d'erreur détectée, cette stratégie peut être couplé avec une demande de retransmission.

Analyse. On suppose que $\mathcal{C}$ est maximal s-détecteur, $d(\mathcal{C}) = s + 1$, avec $t = 0$. Si $d(c, r) \leq s$, le mot erroné $r \neq c$ est détecté, mais si $d(c, r) > s$ on peut recevoir un mot erroné $r \in \mathcal{C}$ différent de c, sans détection d'erreur.

Stratégie correction maximale. On pose $t = \lfloor (d(\mathcal{C}) - 1)/2 \rfloor$, la valeur maximale telle que $\mathcal{C}$ est t-correcteur. On associe l'unique mot de code c' le plus proche à r, dans une boule de rayon t, sinon on signale une erreur.

Analyse. On a $d(\mathcal{C}) = 2t + s + 1$ avec $s = 0$ si $d(\mathcal{C})$ est impair et $s = 1$ si $d(\mathcal{C})$ est pair. S'il y a au maximum t erreurs, on corrige r à $c' = c$ car $B(c, t)$ est la seule boule de rayon t qui contient r. Si $d(\mathcal{C}) = 2t + 2$, et il y a exactement $t + 1 = s + t$ erreurs, on peut détecter l'existence d'au moins $t + 1$ erreurs, mais il peut y avoir plusieurs mots de code c' à distance $t + 1$ de r.

Stratégie t-correction (hybride). S'il existe $c' \in \mathcal{C}$ avec $d(c', r) \leq t$, on retourne c'. Sinon, on signale une erreur.

Analyse. On suppose que $d(\mathcal{C}) = 2t + s + 1$; si $d(c, r) \leq t$, le résultat du décodage est le mot $c' = c$, par l'unicité de la boule $B(c, t)$ qui contient r. Si le nombre d'erreurs est entre $t + 1$ et $s + t$, la transmission erroneé est détectée. Si $B(\mathcal{C}, s + t) \subsetneq \mathcal{A}^n$, il existe des erreurs supplémentaires, avec $d(c, r) > s + t$, qui peuvent être détectées. Or, en depassant $s + t$ certains mots reçus r passent dans $B(c', t)$ avec $c \neq c' \in \mathcal{C}$.

Remarque. La stratégié hybride permet une interpolation entre la stratégie de détection maximale et de correction maximale, avec le choix de t satisfaisant $d(\mathcal{C}) = 2t + s + 1$.

Bornes de Singleton et Hamming

On suppose dans la suite que $\mathcal{C}$ est un code en bloc dans $\mathcal{A}^n$, avec $q = \mathcal{A}$, mais sans hypothèse que $\mathcal{A}$ soit un anneau, ni que $\mathcal{C}$ soit linéare. Si le code est linéaire, l'invariant k signifie sa dimension, un entier, sinon on pose $k = \log_q(|\mathcal{C}|)$. L'objectif des bornes sur les paramétres est de déterminer des contraintes entre les paramétres $[n, k, d]$ dans $\mathcal{A}^n$.

Théorème (Singleton)**.** *Si $\mathcal{C} \subseteq \mathcal{A}^n$ est un code avec $d = d(\mathcal{C})$, alors $|\mathcal{C}| \leq q^{n-d+1}$.*

Définition. Si $k = \log_q(|\mathcal{C}|) = n - d + 1$, on dit que $\mathcal{C}$ est MDS (pour *maximum distance separable* en anglais).

Preuve. Soit $\pi : \mathcal{A}^n \to \mathcal{A}^{n-d+1}$ la projection $\pi(x_1 x_2 \dots x_n) = x_1 x_2 \dots x_{n-d+1}$, sur les premiers $n - d + 1$ coordonnées. La restriction à $\mathcal{C}$ est injective, car $\pi(x) = \pi(y)$ si et

seulement si $x = y$ car $d(x, y) \leq d - 1$. Il suit que $|\mathcal{C}| = |\pi(\mathcal{C})| \leq |\mathcal{A}^{n-d+1}| = q^{n-d+1}$. □

Exemple. Soit $\mathcal{C} = \langle 1001, 0101, 0011 \rangle$ le code linéaire dans $\mathbb{F}_2^4$. On observe que tout mot de code est de poids pair — c'est le code définit par $x_1 + x_2 + x_3 + x_4 = 0$ – donc la distance minimum d est pair et majoré par 2, donc $d = 2$. La dimension est $k = 3$, alors $[n, k, d] = [4, 3, 2]$. Il suit que $\mathcal{C}$ est MDS car $k = n - d + 1 = 3$.

On rappelle la forme du cardinal d'une boule autoru de $x \in \mathcal{A}^n$.

$$V(q, n, t) = |B(x, t)| = \sum_{i=0}^{t} \binom{n}{i} (q-1)^i.$$

Le théorème suivant énonce une majoration impliquée par la couverture disjointes de $\mathcal{C}$ par des boules de rayon t.

Théorème (Hamming)**.** *Si $\mathcal{C} \subseteq \mathcal{A}^n$ est un code t-correcteur avec $d = d(\mathcal{C})$, alors*

$$|\mathcal{C}|V(q, n, t) \leq q^n.$$

Preuve. Les boules de rayon t sont disjointes, alors

$$q^n = |\mathcal{A}^n| \geq |B(\mathcal{C}, t)| = \sum_{c \in \mathcal{C}} |B(c, t)| = |\mathcal{C}|V(q, n, t). \quad \square$$

Définition. Un code qui satisfait $|\mathcal{C}| = q^n / V(q, n, t)$ s'appelle *parfait*.

Corollaire. *Le taux de transmission de $\mathcal{C}$ est au maximum* $1 - \dfrac{\log_q(V(q, n, t))}{n}$.

Preuve. Par définition $R = \dfrac{\log_q(\mathcal{C})}{n} \leq \dfrac{\log_q(q^n / V(q, n, t))}{n}$. □

Canal binaire symétrique

Définition. Un *canal binaire symétrique* est un modèle probabiliste pour un canal discret de transmission avec bruit. On suppose que le canal transmet des suites de bits (des 0 et 1) avec probabilité ε qu'un bit soit changé – de 0 en 1 ou 1 en 0 avec la même probabilité (propriété symétrique). On dit que le canal est *sans mémoire* si le probabilité de changement est indépendant en position (ou temps) i et $j \neq i$.

On étudie un exemple de code linéaire $\mathcal{C}$ dans $\mathbb{F}_2^n$, en regardant les propriétés de couverture par des boules, et le lien avec la correction et détection d'erreurs après transmission par un canal binaire symétrique sans mémoire.

Exemple. Soit $\mathcal{C}$ le code linéaire de dimension 2 dans $\mathbb{F}_2^6$:

$$\langle 101010, 010101 \rangle = \{000000, 101010, 010101, 111111\}.$$

En regardant les poids des éléments, on détermine que le code est de distance minimum 3, alors 1-correcteur. On a alors

$$B(\mathcal{C}, 1) = 4(1 + 6) = 28 < |\mathbb{F}_2^6| = 64,$$

et on peut montrer que $B(\mathcal{C}, 2) = 64$.

Supposons qu'on utilise ce code pour transmission sur un canal binaire symétrique sans mémoire, avec probabilité d'erreur ε. Si c est le mot de code transmis, et r le mot reçu, on peut tabuler les probabilités des distances $d(c, r)$.

$d(c,r)$	num	prob	notes
0	1	$(1-\varepsilon)^6$	$\}$ $r \in B(c,1)$
1	6	$6(1-\varepsilon)^5\varepsilon^1$	
2	15	$15(1-\varepsilon)^4\varepsilon^2$	dont 6 dans $B(\mathcal{C},1)$
3	20	$20(1-\varepsilon)^3\varepsilon^3$	dont 2 dans $\mathcal{C}$
4	15	$15(1-\varepsilon)^2\varepsilon^4$	dont 6 dans $B(\mathcal{C},1)$
5	6	$6(1-\varepsilon)^1\varepsilon^5$	$\}$ $r \in B(c+111111,1)$
6	1	ε^6	
Total :	64	1	

Analyse. Avec une stratégie de 1-correction, la probabilité de transmission correcte est

$$(1-\varepsilon)^6 + 6(1-\varepsilon)^5\varepsilon \sim 1 - 15\varepsilon^2.$$

Or il y a 28 mots dans $B(\mathcal{C}, 1)$, dont seulement 7 sont correctement corrigés. Les autres 21 seront corrigés vers le mauvais mot de code, avec probabilité totale

$$6(1-\varepsilon)^4\varepsilon^2 + 2(1-\varepsilon)^3\varepsilon^3 + 6(1-\varepsilon)^2\varepsilon^4 + 6(1-\varepsilon)^1\varepsilon^5 + \varepsilon^6 \sim 6\varepsilon^2.$$

Les 36 mots restants sont détectables comme erronés, constituant une probabilité totale :

$$9(1-\varepsilon)^4\varepsilon^2 + 18(1-\varepsilon)^3\varepsilon^3 + 9(1-\varepsilon)^2\varepsilon^4 \sim 9\varepsilon^2.$$

Les équivalences de la forme $f(\varepsilon) \sim g(\varepsilon)$ sont dans le voisinage de $\varepsilon = 0$. On aperçoit qu'environs 60% des erreurs non corrigés sont des erreurs détectés et et 40% sont mal corrigés.

Détection seule. Avec une stratégie de détection (0-correction), on obtient les probabilités respectives de bonne correction :

$$(1-\varepsilon)^6 \sim 1 - 6\varepsilon,$$

de détection :

$$6(1-\varepsilon)^5\varepsilon^1 + 15(1-\varepsilon)^4\varepsilon^2 + 18(1-\varepsilon)^3\varepsilon^3 + 15(1-\varepsilon)^2\varepsilon^4 \sim 6\varepsilon,$$

et de correction erronée :

$$2(1-\varepsilon)^3\varepsilon^3 + \varepsilon^6 \sim 2\varepsilon^3.$$

En particulier la probabilité d'une correction erronée dévient négligeable, $O(\varepsilon^3)$, par rapport à détection $O(\varepsilon)$ dans le voisinage de $\varepsilon = 0$.

Codes de Hamming

On rappelle que le code de Hamming de paramétre $r \geq 2$ est un code linéaire $\mathcal{H}_r \subset \mathbb{F}_2^n$ de longeur $n = 2^r - 1$ et dimension $k = n - r$, ayant pour matrice de contrôle une matrice $H(r)$ de r lignes et n colonnes dont toutes les colonnes sont distinctes et non nulles. Autrement dit, les lignes de $H(r)^t$ parcourrent les éléments non nuls de $\mathbb{F}_2^r$. À équivalence près, on peut supposer que le i-ième ligne de $H(r)^t$ représente l'écriture de l'entier i en binaire.

Exemple. Les premières matrices de contrôle d'un code de Hamming sont :

$$H(2) = \begin{matrix} \\ \end{matrix} \begin{matrix} 1 & 2 & 3 \end{matrix} \\ \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \text{ et } H(3) = \begin{matrix} \\ \end{matrix} \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{matrix} \\ \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

et on peut contruire les matrices de contrôle d'ordre supérieur de manière récursive de la forme :

$$H(r+1) = \left[\begin{array}{ccc|c|ccc} & & & 0 & & & \\ & H(r) & & \vdots & & H(r) & \\ & & & 0 & & & \\ \hline 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \end{array}\right].$$

D'autres présentations des codes de Hamming sont possibles, en écrivant la matrice génétrice et matrice de contrôle en forme systématique. Or, cette formulation permet un décodage intuitif par syndrôme (voir la suite).

Les paramétres du code de Hamming sont $[n, n-r, 3]$, où $n = 2^r - 1$, par conséquent du lemme suivant, les premiers paramétres sont $[3, 1, 3]$, $[7, 4, 3]$, et $[15, 11, 3]$.

Lemme. *Un code de Hamming est de distance minimum* 3.

Preuve. On note qu'un code binaire avec matrice de contrôle H est l'ensemble de vectors

$$\mathcal{C} = \{c \in \mathbb{F}_2^n \,:\, c.H^t = 0\}.$$

On fait quelques observations concernant le rapport entre la distance minimum et la matrice de contrôle.

- Un code $\mathcal{C}$ est de distance 1 si et seulement s'il existe une colonne toute nulle dans la matrice de contrôle, car $d(\mathcal{C}) = 1$ si est seulement un des vecteurs de la base canonique,

  $$e_i = 0 \cdots 010 \cdots 0,$$

  est dans $\mathcal{C}$. Or, $e_i.H^t \in \mathbb{F}_2^r \backslash \{0\}$ est la i-ième ligne de H^t.
- Un code $\mathcal{C}$ a distance minimum 2 si $d(\mathcal{C}) \neq 1$ et la somme de deux colonnes de H est nulle, et, sur $\mathbb{F}_2$, si est seulement si deux colonnes sont égaux. On utilise le même argument : la distance minimum est 2 si est seulement si $c = e_i + e_j \in \mathcal{C}$, et, $c.H^t$ est la sommes des i-ième et j-ième lignes de H^t.

- Par la suite, on voit que la distance minimum est au maximum d si et seulement si un combinaison linéaire (une somme sur $\mathbb{F}_2$) de d colonnes de H est le vecteur nul.

On rappelle que les lignes de $H(r)^t$ parcourrent les vecteurs de $\mathbb{F}_2^r \backslash \{0\}$. Alors par définition de $H(r)$, aucune colonne est nulle et aucun pair de colonnes sont égalles. Or, pour tout pair de colonnes leur somme est encore une colonne de $H(r)$ (par clôture d'addition sur l'espace vectoriel $\mathbb{F}_q^r$). Il suit que la distance minimum est 3. □

Les codes de Hamming ont asymptotiquement un très haut rendement (après les premiers petits exemples), $R = (n-r)/n$ où $n = 2^r - 1$, même si les et faible correction, étant 1-correcteurs. Les paramètres des premiers codes de Hamming sont

r	n	k	d	R
2	3	1	3	$1/3 = 0{,}3333\ldots$
3	7	4	3	$4/7 = 0{,}5714\ldots$
4	15	11	3	$11/15 = 0{,}7333\ldots$

r	n	k	d	R
5	31	26	3	$26/31 = 0{,}8387\ldots$
6	63	57	3	$19/21 = 0{,}9047\ldots$
7	127	120	3	$120/127 = 0{,}9448\ldots$

On peut construire des matrices génératrices pour ces premier codes. La matrice $H(2)$ est

$$H(2) = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

alors, on trouve une matrice génératrice $G(2) = \begin{bmatrix} 1\ 1\ 1 \end{bmatrix}$, engendré par le mot de code de triple répétition. De même, à partir de la matrice $H(3)$ on trouve une matrice $G(3)$

$$H(3) = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \text{ on trouve } G(3) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

une matrice génératrice en forme systématique pour le $[7, 4, 3]$-code de Hamming.

Finalement, on conclure cette section avec le résultat qui montre un des intérêts principaux pour les codes de Hamming, à la fois théorique et pratique.

Théorème. *Les codes de Hamming sont parfaits.*

Preuve. Le code $\mathcal{H}_r$ de Hamming, avec paramétres $[n, k, d] = [2^r - 1, n - r, 3]$ est un code 1-correcteur, donc

$$|B(\mathcal{H}_r, 1)| = |\mathcal{H}_r| V(q, n, 1) = 2^{n-r}(1 + n) = 2^{n-r} 2^r = 2^n,$$

alors $\mathcal{H}_r$ est parfait. □

Il suit qu'il y a un algorithme efficace pour décodage $\mathcal{A}^n \to \mathcal{C}$, qui permet de calculer le mot de code unique à distance 1 maximum associé à tout mot reçu $r \in \mathcal{A}^n$.

Décodage par syndrôme

Les codes de Hamming sont particulièrement bien adaptés à l'algorithme de décodage nommé décodage par syndrôme. Il y a deux étapes : la construction d'une table des syndrômes associés aux vecteurs d'erreurs de petits poids, et décodage par recherche de syndrôme dans la table pour trouver le vecteur d'erreur.

Le décodage par syndrôme d'un code linéaire $\mathcal{C}$ à paramétres $[n, k, d]$ est une méthode efficace 1) quand $n-k$ est petit — tel qu'on peut énumérer les q^{n-k} syndrômes de $\mathbb{F}_q^{n-k}$ – ou 2) quand on veut décoder à faible distance t — tel qu'on peut énumérer les $V(q, n, t)$ éléments de la boule $B(0, t)$. On effectue le décodage en utisant une table précalculée de couples (s, e) d'un syndrôme $s = \pi(e)$ avec vector d'erreur e, de poids $\|e\| \leq t$ dans $B(0, t)$.

On rappelle que le homomorphisme $\pi : \mathbb{F}_q^n \to \mathbb{F}_q^{n-k}$ associe le syndrôme $s = \pi(r)$ à un mot reçu $r \in \mathbb{F}_q^n$, et que $\pi(c) = 0$ pour tout mot de code $c \in \mathcal{C}$. Si le mot reçu est de la forme $r = c + e$, pour un mot trans $c \in \mathcal{C}$ et erreur $e \in \mathbb{F}_q^n$, on observe que le syndrôme $\pi(r)$ ne dépend que de la classe d'équivalence $\mathbb{F}_q^n/\mathcal{C}$, et en particulier $\pi(r) = \pi(e)$. Un *leader de classe* est un mot e de poids minimal dans la class $r + \mathcal{C}$. La correction du mot reçu est effectuée en mettant $c' = r - e$ où e est un leader de classe de $r + \mathcal{C}$.

Si $2t+1 \leq d(\mathcal{C})$ et $\|e\| \leq t$, alors e est un leader de classe. L'idée est donc de précalculer chaque syndrôme $s = \pi(e)$ des vecteurs d'erreurs e de poids 1 jusqu'à t, en mettant le pair (s, e) dans un *tableau standard*, une dictionnaire de recherche de leader de classe par syndrôme.

Exemple. Le tableau standard pour le $[7, 4, 3]$-code $\mathcal{H}_3$ de Hamming, est la suivante :

$$\{(001, e_1), (010, e_2), (011, e_3), (010, e_4), (101, e_5), (110, e_6), (111, e_7)\},$$

où e_i est le i-ième vecteur de la base canonique de $\mathbb{F}_2^7$. Par construction de l'ordre des colonnes de $H(r)$, le syndrôme du vecteur canonique e_i est l'écriture d'entier i en binaire.

Exercices

1. Soit $\mathcal{C}$ le code engendré par la matrice

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

- **a.** Déterminer le nombre de mots de code de $\mathcal{C}$.
- **b.** Calculer une matrice de contrôle.
- **c.** Calculer la distance minimum de $\mathcal{C}$.
- **d.** Déterminer le nombre d'erreurs que $\mathcal{C}$ peut détecter/corriger.

2. Soit $\mathcal{C}$ le code avec matrice de contrôle

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

a. Donner une matrice génératrice pour $\mathcal{C}$.

b. Décoder par syndrome $r = 11101$ et $r' = 11011$.

3. On appelle *code de Hamming* de paramètre $r \geq 2$ un code binaire de longueur $2^r - 1$ et dimension $2^r - r - 1$ ayant pour matrice de contrôle une matrice $H(r)$ de r lignes et $2^r - 1$ colonnes dont toutes les colonnes sont distinctes et non nulle. A équivalence (isomorphisme) près on peut supposer que la i-ème colonne de $H(r)$ représente l'écriture binaire de l'entier i.

a. Construire $H(2)$ et $H(3)$.

b. Donner une matrice génératrice pour ces codes.

c. Montrer que les codes de Hamming sont de distance 3.

d. Montrer que ce sont des codes parfaits, en particulier l'union des boules de rayon $t = 1$ centré sur un mot du code est égale à $\mathbb{F}_2^n$.

e. Montrer qu'un code de Hamming est MDS si et seulement si $r = 2$.

f. Ces codes sont très faciles à décoder : montrer qu'on peut choisir pour leader de classe un mot ayant un seul 1 à la place i pour les $2^r - 1$ classes non triviales.

4. Montrer que le code binaire de répétition de longueur n (code linéaire avec $G = [1, 1, \ldots, 1]$) est un code MDS, et si $n = d = 2t + 1$, il est parfait.

5. Remplir les valeurs de $V(2, n, t)$ dans le tableau suivant :

t	0	1	2	3	4	5	6	7
$V(2,2,t)$								
$V(2,3,t)$								
$V(2,4,t)$								
$V(2,5,t)$								
$V(2,6,t)$								
$V(2,7,t)$								

Pour quelles valeurs $[n, k, d]$ ci-dessus est-ce qu'il peut exister un codage linéaire en bloc parfait ?

6. Soit $\mathcal{C}_i : X \to \{0,1\}^n$ les codages linéaires avec matrices génératrices

$$G_1 = \begin{bmatrix} 1\,1\,0\,0\,0\,1 \\ 0\,0\,1\,1\,1\,0 \end{bmatrix},\ G_2 = \begin{bmatrix} 1\,1\,0\,0\,1\,1 \\ 0\,0\,1\,1\,1\,1 \end{bmatrix},\ \text{et } G_3 = \begin{bmatrix} 1\,0\,0\,0\,1\,1\,0 \\ 0\,1\,0\,0\,0\,1\,1 \\ 0\,0\,1\,0\,1\,1\,1 \\ 0\,0\,0\,1\,1\,0\,1 \end{bmatrix}.$$

Rappelons que pour chaque x dans $\mathcal{A}^n$ où $|\mathcal{A}| = q$, et entier t, le cardinal de la boule $B(x,t)$ est

$$|B(x,t)| = V(q,n,t) = \sum_{i=0}^{t} \binom{n}{i} (q-1)^i.$$

a. Trouver les paramètres $[n, k, d]$ pour chaque code, vérifier les bornes de Singleton et Hamming.

b. Pour chaque code $\mathcal{C}_i$ et t entre 0 et 2, compter $|B(\mathcal{C}_i, t)|$, le nombre de mots à distance t d'un mot de code.