le cnam

Codes et Automates finis

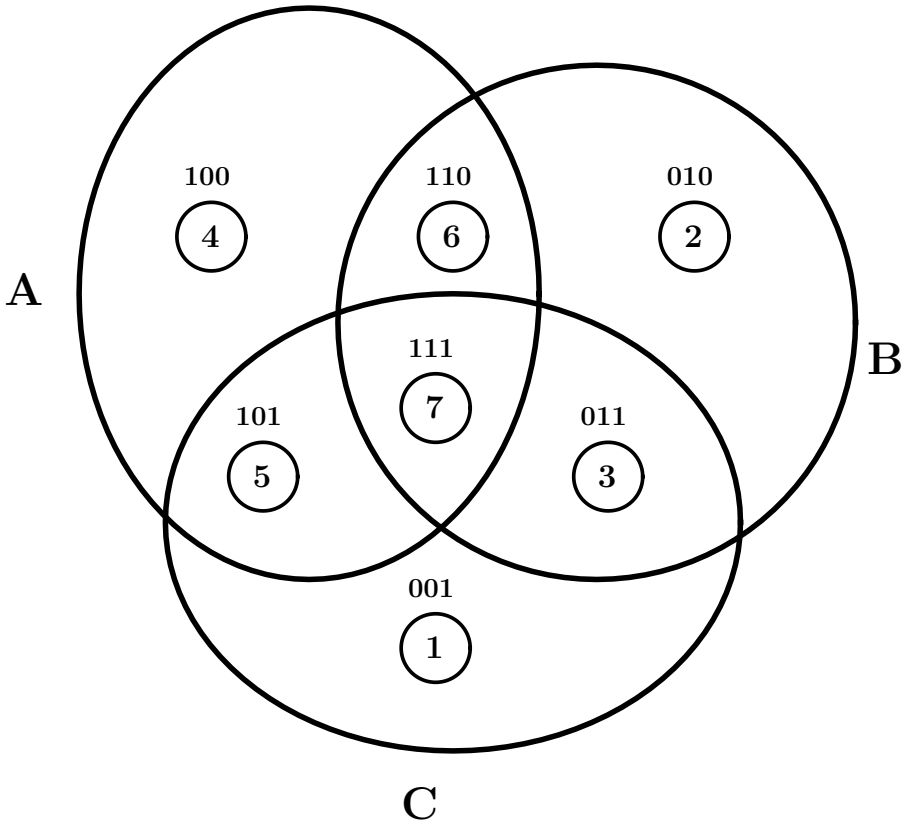
Cours 4 Code de Hamming H7

- Introduction

En 1948, Richard Hamming a proposé de coder un message de quatre bits ($a \in (\mathbb{F}_2)^4$) en utilisant un code de sept bits ($u \in (\mathbb{F}_2)^7$) généré de la façon suivante. On se donne d'abord les nombres de 1 à 7 écrits en base deux :

1	2	3	4	5	6	7
001	010	011	100	101	110	111

On se donne ensuite un diagramme de Venn formé de trois courbes fermées A, B et C. La courbe A est associée au premier bit (le plus à gauche par convention ici), la courbe B au second bit (celui du milieu) et la courbe C au troisième bit, le plus à droite. Si un nombre a est codé en base deux a l'aide de deux bits égaux a "1" en position gauche et un "0" en position droite, il est représenté par un symbole qui appartient à $A \cap B$ mais qui n'appartient pas à C.



Vision historique du code de Hamming H7

- Définition

Ensuite, Hamming numérote les bits du mot envoyé $u \in (\mathbb{F}_2)^7$ à l'aide du diagramme ci-dessus. Pour les indices qui ne sont pas des puissances de deux, on recopie le bit correspondant, ce qui permet de définir $u_3 = a_1$, $u_5 = a_2$, $u_6 = a_3$ et $u_7 = a_4$. Pour les indices numéros 1, 2 et 4 qui sont des puissances de deux, on construit un bit de contrôle de façon à annuler la somme des bits qui se situent dans les diagrammes C, B et A respectivement. De façon précise, les bits de contrôle u_1, u_2 et u_4 sont définis par les relations
$u_1 + u_3 + u_5 + u_7 = 0$, $u_2 + u_3 + u_6 + u_7 = 0$ et $u_4 + u_5 + u_6 + u_7 = 0$. On peut réécrire ces relations à l'aide des bits a_1, a_2, a_3 et a_4 du message initial : $u_1 + a_1 + a_2 + a_4 = 0$,
$u_2 + a_1 + a_3 + a_4 = 0$ et $u_4 + a_2 + a_3 + a_4 = 0$. Puis prendre en compte le fait que les nombres sont dans $\mathbb{F}_2$: $u_1 = a_1 + a_2 + a_4$, $u_2 = a_1 + a_3 + a_4$, $u_4 = a_2 + a_3 + a_4$.
Si on souhaite par exemple envoyer le message $a = 1101$, on place d'abord les quatre bits évidents : $u = **1*101$. Puis on calcule, dans $\mathbb{F}_2$, les trois bits manquants :
$u_1 = 1+1+1 = 1$, $u_2 = 1+0+1 = 0$ et $u_3 = 1+0+1 = 0$. Ainsi la fonction de codage φ est telle que $\varphi(1101) = 1010101$, qui est bien une chaîne de sept bits.

- Linéarité

La codage ainsi défini est linéaire. En effet, on peut écrire
$\varphi(a_1, a_2, a_3, a_4) = (a_1 + a_2 + a_4, a_1 + a_3 + a_4, a_1, a_2 + a_3 + a_4, a_2, a_3, a_4)$. La linéarité de cette application s'établit avec un calcul à la fois simple mais un peu long à écrire qui a été introduit au chapitre précédent.
Compte tenu de cette linéarité, toute l'information pour le codage est contenue dans les vecteurs $c_j = \varphi(e_j)$ pour les quatre vecteurs de base e_j de l'espace $(\mathbb{F}_2)^4$. On a ainsi
$c_1 = \varphi(1000) = 1110000$, $c_2 = \varphi(0100) = 1001100$, $c_3 = \varphi(0010) = 0101010$ et
$c_4 = \varphi(0001) = 1101001$.

- Matrice génératrice

La matrice génératrice G_0 s'obtient en écrivant les composantes des vecteurs lignes c_j les unes au dessous des autres. Il vient $G_0 = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$.
La calcul de $u = \varphi(1101)$ par exemple se réduit à la multiplication du vecteur ligne (1101) par la matrice G_0 : $u = (1101)\, G_0 = (1010101)$. On retrouve bien un résultat établi quelques lignes plus haut.

- Syndrome d'un mot reçu

Après transmission du message envoyé u, on reçoit un mot $v \in (\mathbb{F}_2)^7$. Il est égal au mot envoyé u, plus une erreur γ : $v = u + \gamma$. *A priori*, le mot reçu v appartient à $(\mathbb{F}_2)^7$, ensemble qui contient $2^7 = 128$ éléments. Notons bien que les mots envoyés vivent eux dans sous ensemble $\mathscr{C} = \{a\, G_0, a \in (\mathbb{F}_2)^4\}$ de $(\mathbb{F}_2)^7$ beaucoup plus petit puisqu'il compte seulement $2^4 = 16$ éléments. On note $|\mathscr{C}| = 16$ le nombre d'éléments du code $\mathscr{C}$.
Comment repérer qu'un mot $v \in (\mathbb{F}_2)^7$ appartient au code $\mathscr{C}$ ou pas ? On calcule la cohérence

du mot reçu avec les trois bits de parité. On pose $s_1 = v_4 + v_5 + v_6 + v_7$, $s_2 = v_2 + v_3 + v_6 + v_7$ et $s_3 = v_1 + v_3 + v_5 + v_7$. On construit ainsi un mot $s(v) = (s_1, s_2, s_3) \in (\mathbb{F}_2)^3$ qui est appelé le "syndrome" du mot reçu $v \in (\mathbb{F}_2)^7$. On remarque que l'application $(\mathbb{F}_2)^7 \ni v \longmapsto s(v) \in (\mathbb{F}_2)^3$ est linéaire.
Si $v \in \mathscr{C}$ est un mot du code, alors les trois bits de parité définis par s_1, s_2 et s_3 sont nuls et $s(v) = 000$. Le syndrome est nul si $v \in (\mathbb{F}_2)^7$ est un mot du code.
Si il y a eu au moins une erreur de transmission, alors en général le syndrome $s(v)$ est non nul. Ainsi, ayant calculé le syndrome $s(v) \in (\mathbb{F}_2)^3$ du mot reçu $v \in (\mathbb{F}_2)^7$, si on observe que $s(v) \neq 0$, on est certain qu'il y a eu au moins une erreur de transmission.

- Repérage d'un bit défaillant.

Proposition. Si le mot envoyé $u \in \mathscr{C}$ a été modifié une seule fois, alors le syndrome $s(v)$ du mot reçu indique, en base deux, le numéro du bit qui a été modifié.
Cette proposition justifie *a posteriori* la curieuse numérotation qui a été adoptée initialement pour le codage de Hamming.
Si on a reçu par exemple $v = 1111011$, on calcule le vecteur syndrome $s(v)$ sans difficulté : $s_1 = v_4 + v_5 + v_6 + v_7 = 1 + 0 + 1 + 1 = 1$, $s_2 = v_2 + v_3 + v_6 + v_7 = 1 + 1 + 1 + 1 = 0$ et $s_3 = v_1 + v_3 + v_5 + v_7 = 1 + 1 + 0 + 1 = 1$. Donc $s(v) = 101$, soit le nombre "5" écrit en base deux. La correction suggérée par la proposition précédente est que le mot envoyé u est identique au mot reçu v, au cinquième bit près, soit $u = 1111111$.

mot du code	u_1	u_2	u_3	u_4	u_5	u_6	u_7	numéro du mot
c_0	0	0	0	0	0	0	0	0
c_1	1	1	1	0	0	0	0	1
c_2	1	0	0	1	1	0	0	2
c_3	0	1	0	1	0	1	0	3
c_4	1	1	0	1	0	0	1	4
$c_1 + c_2$	0	1	1	1	1	0	0	5
$c_1 + c_3$	1	0	1	1	0	1	0	6
$c_1 + c_4$	0	0	1	1	0	0	1	7
$c_2 + c_3$	1	1	0	0	1	1	0	8
$c_2 + c_4$	0	1	0	0	1	0	1	9
$c_3 + c_4$	1	0	0	0	0	1	1	10
$c_1 + c_2 + c_3$	0	0	1	0	1	1	0	11
$c_1 + c_2 + c_4$	1	0	1	0	1	0	1	12
$c_1 + c_3 + c_4$	0	0	0	1	1	1	1	13
$c_2 + c_3 + c_4$	0	1	1	0	0	1	1	14
$c_1 + c_2 + c_3 + c_4$	1	1	1	1	1	1	1	15

Table des mots du code de Hamming H7

- Géométrie discrète

La propriété précédente est une conséquence du fait que les mots du code $\mathscr{C}$ sont "assez loin les uns des autres". Explicitons cette propriété pour les seize mots du code H7 en utilisant à nouveau la linérité. En effet, si u_1 et u_2 sont deux mots du code, c'est à dire deux mots de l'ensemble $\mathscr{C}$, alors la somme u_1+u_2 est encore un mot du code. Nous retenons que si φ est linéaire, la somme de deux mots du code est encore un mot du code :
$u_1, u_2 \in \mathscr{C} \Longrightarrow u_1+u_2 \in \mathscr{C}$.
Dans le tableau de la page précédente, on a fait la liste des seize mots reçus, numérotés de 0 à 15. Afin d'exprimer que deux chaînes de bits sont plus ou moins proches, on introduit une définition très générale.

- Distance de Hamming

Pour m entier supérieur ou égal à 1, soit $u=(u_1, u_2, \ldots, u_j, \ldots u_m) \in (\mathbb{F}_2)^m$ et
$v=(v_1, v_2, \ldots, v_j, \ldots v_m) \in (\mathbb{F}_2)^m$. Pour $1 \leq j \leq m$, on pose d'abord $\delta_j(u,v)=0$ si $u_j=v_j$ et $\delta_j(u,v)=1$ si $u_j \neq v_j$. On définit ensuite la distance de Hamming $d(u,v)$ par la somme $d(u,v)=\sum_{j=1}^m \delta_j(u,v)$ des nombres entiers $\delta_j(u,v)$. Le nombre $d(u,v)$ est un entier compris entre 0 et m. On observe que la fonction $(\mathbb{F}_2)^m \times (\mathbb{F}_2)^m \ni (u,v) \longmapsto d(u,v) \in \mathbb{N}$ a toutes les caractéristiques d'une distance dans un espace métrique.

- La distance de Hamming est une distance

On a en effet les propriétés suivantes : positivité : $d(u,v) \geq 0$, symétrie : $d(v,u)=d(u,v)$,
la distance de deux points identiques est toujours nulle : $d(u,u)=0$,
séparation : $(d(u,v)=0) \Longrightarrow (u=v)$ et inégalité triangulaire : $d(u,v) \leq d(u,w)+d(w,v)$.

- Distance minimale

On se donne un code $\varphi : (\mathbb{F}_2)^k \longrightarrow (\mathbb{F}_2)^n$ qui peut être linéaire ou non. Son image
$\mathscr{C}=\varphi\big((\mathbb{F}_2)^k\big)=\{\varphi(a), a \in (\mathbb{F}_2)^k\}$ est incluse dans $(\mathbb{F}_2)^n$. La distance minimale du code $\mathscr{C}$ est par définition le nombre $d=\min\{d(u,v), u \in \mathscr{C}, v \in \mathscr{C}, u \neq v\}$.
Pour calculer la distance minimale du code de Hamming H7, il faut *a priori* énumérer les couples de mots différents de l'ensemble $\mathscr{C}$ qui a été explicité à la page précédente. Compte tenu de la symétrie de la distance, on a *a priori* $\frac{16 \times 15}{2}=8 \times 15=120$ cas distincts à considérer. Dans le cas d'un code linéaire, et en particulier dans le cas du code de H7, on a une propriété qui permet un calcul plus rapide.

- Distance minimale d'un code linéaire

Si le code $\varphi : (\mathbb{F}_2)^k \longrightarrow (\mathbb{F}_2)^n$ est linéaire, alors la distance minimale d est donnée par la relation $d=\min\{d(u,0), u \in \mathscr{C}, u \neq 0\}$.
La distance minimale du code de Hamming H7 est égale à 3 : $d(\text{H7})=3$. Il suffit de compter le nombre minimal de "1" dans le tableau des mots du code présenté plus haut, en excluant bien sûr le mot nul.
Dans la suite de ce cours, on appelle poids d'un mot $v \in (\mathbb{F}_2)^m$ et on note $w(v)$ ou parfois $|v|$ le nombre entier $d(v,0)$: $|v|=d(v,0)$.

- Le code de Hamming H7 est un code parfait

En effet, tout mot reçu $v \in (\mathbb{F}_2)^7$ est toujours situé à une distance au plus égale à 1 d'au moins un mot du code $\mathscr{C}$. Pour tout mot $v \in (\mathbb{F}_2)^7$, il existe toujours au moins un mot du code $u \in \mathscr{C}$ de sorte que $d(u,v) \leq 1$. De plus, ce mot est unique.
Cette proposition est généralisée plus loin dans le cours avec la notion de "code parfait". Cette propriété est aussi une conséquence du fait que les mots du code H7 sont assez loin les uns des autres. Cette condition exprime le fait que la distance minimale du code de Hamming est égale à 3.

Exercices

- Ordonner autrement le code H7

a) Construire une variante $(\mathbb{F}_2)^4 \ni a \longmapsto \widetilde{u} = \widetilde{\varphi}(a) \in (\mathbb{F}_2)^7$ du code de Hamming H7 de sorte que $\widetilde{u}_1 = a_1$, $\widetilde{u}_2 = a_2$, $\widetilde{u}_3 = a_3$ et $\widetilde{u}_4 = a_4$.

b) Montrer qu'alors la matrice générarice G s'écrit $G = (\mathrm{I}_4 \ P)$, où l'on précisera la matrice P qui contient quatre lignes et trois colonnes.

c) Comment peut-on alors calculer le vecteur $s(v)$ des syndromes ?

d) Comment relier le syndrome $s(v)$ avec le numéro du bit à corriger ?

- Erreur pour le code H7

On suppose que le canal de transmission $u \longrightarrow v$ est symétrique et sans mémoire. Cette propriété entraîne que la défaillance ou non de deux bits différents constituent des événements indépendants. La probabilité d'erreur pour la transmission d'un bit est égale à p avec $0 < p < 1$.

a) On se donne un entier k avec $0 \leq k \leq 7$. Quelle est la probabilité d'avoir une distance $d(u,v) = k$ entre le mot envoyé u et le mot reçu v ?

b) Quel est l'ordre de grandeur des huit probabilités précédentes si $p = 10^{-3}$?

c) Quelle est la probabilité de fausse correction si on modifie un éventuel bit défaillant ?

- Comptage de mots

On se donne un entier $m \geq 1$ et un mot $u \in (\mathbb{F}_2)^m$.

a) Combien de mots $v \in (\mathbb{F}_2)^m$ vérifient la relation $d(u,v) = k$ si k est un etier entre 0 et m ?

b) Vérifier que la somme pour k variant de 0 à m des nombres précédents est égale à 2^m.