Théorie de l'information

D. Kohel

13/01/2022

Distance de Hamming

$d : A^n \times A^n \longrightarrow \mathbb{N}$, mesure le nombre de caractères différentes.

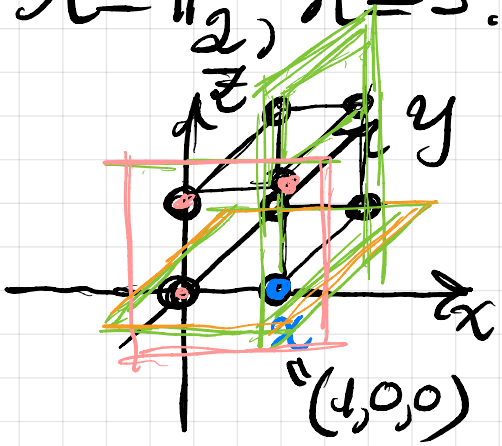
Remarque Pour un mot c transmis et mot r reçu, $d(c,r) = \#$ d'erreurs.

La boule de rayon t : Pour $x \in A^n$, $B(x,t) = \{y \in A^n : d(x,y) \leqslant t\}$.

Remarque. Tandis que $B(x,t)$ est nommé une boule, elle est discrète :

- On peut définir des sphères :
  $S(x,t) = \{y \in A^n : d(x,y) = t\}$
  et alors $B(x,t) = \bigcup_{i=0}^{t} S(x,i)$, union disjointe
- $B(x,t)$ est l'union de droites ($t=1$), plans ($t=2$), ou hyperplans de dimension t :

Ex. $A = \mathbb{F}_2$, $n=3$: $A^3 = \mathbb{F}_2^3$ est un espace vect.



$B((1,0,0),2)$ est l'union de trois plans de $\mathbb{F}_2^3$.

Attention. Le seul point qui n'est pas dans $B((1,0,0),2)$ est $(0,1,1)$.

Ex. Soit $C = \{\underbrace{00000}_{c_0}, \underbrace{10110}_{c_1}, \underbrace{01011}_{c_2}, \underbrace{11101}_{c_3}\} \subseteq \mathbb{F}_2^5$.

Les distances entre mots de code sont :

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

On voit que $c_3 = c_1 + c_2$, alors C est linéaire, de dimension 2.

Alors $d(c_1, c_2) = d(0, c_3)$, et $d(c_1, c_3) = d(0, c_2)$, $d(c_2, c_3) = d(0, c_1)$.

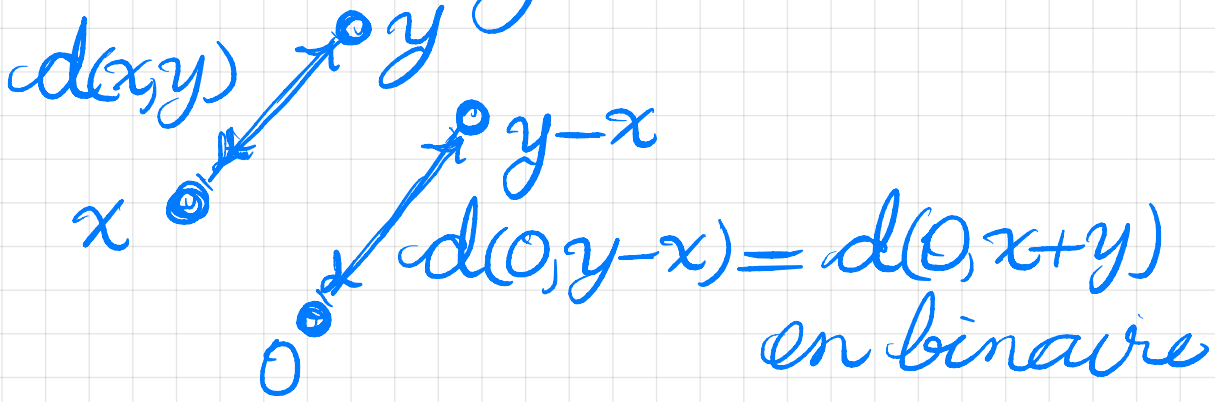
c_0 $c_1 + c_2 \neq (c_2 - c_1)$

Le tableau est déterminé par
(i) 0 sur le diagonale
(ii) les trois valeurs $d(0, c)$, $c \neq 0$.

Pour A^n en général, on a besoin de $|A^n| - 1$ distances non triviales, $q^n - 1$ en nombre, où $q = |A|$.

- On note que $c_1 + c_2 + c_3 = 0$, alors $c_1 + c_2 = c_3$, $c_2 + c_3 = c_1$, etc, car $-c = c$ en binaire ($A = \mathbb{F}_2$)
- $d(c_i, c_j) = d(0, c_i + c_j)$, par

linéarité de la distance de Hamming :



Invariance par translation.

La distance minimum,

$$d(\mathcal{C}) = \min\{d(x,y) : x,y \in \mathcal{C}, x \neq y\}$$
$$= \min\{d(0,x) : x \in \mathcal{C} \setminus \{0\}\}$$

(↑ linéarité)

est égale à 3.

Décodage : Soit c un mot transmis ($c \in \mathcal{C}$) et r le mot reçu ($r \in A^n$). On veut décoder r, en corrigeant des erreurs, et/ou détecter des erreurs. On dit que $\mathcal{C}$ est t-correcteur (d'erreurs) si les boules de rayon t autour des mots $c \in \mathcal{C}$ sont (deux à deux) disjointes.

Exercice. Supposer $|A| = q$.
Montrer que

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

Preuve.
$\binom{n}{i}$ = nombre de choix de i positions changées parmi n.
$(q-1)^i$ = cardinal des choix de coordonnées changées.

Ex. $A^n = \mathbb{F}_2^3$, $x = (1,0,0) = 100$

Indépendent de $x \in A^n$.

On note que $q=2$, alors $q-1 = 1$.

$|S(x,0)| = 1 = |\{x\}| = |\{(1,0,0)\}|$

$|S(x,1)| = \binom{3}{1} = 3 = |\{000, 110, 101\}|$

$|S(x,2)| = \binom{3}{2} = 3 = |\{111, 001, 010\}|$

$|S(x,3)| = 1 = |\{011\}|$

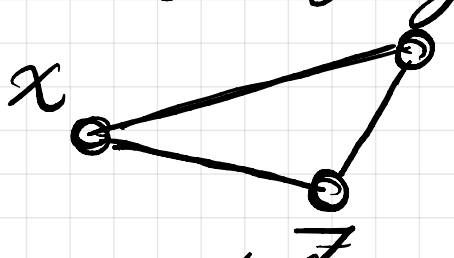
Total : 8 éléments de $\mathbb{F}_2^3$.

On en déduit de l'exercice que C est t-correcteur ssi

- $|B(C,t)| = |C| \sum_{i=0}^{t} \binom{n}{i} (q-1)^i$, ssi
- $d(C) \geq 2t+1$.

Indication $B(x,t) \cap B(y,t) = \emptyset$ ssi $d(x,y) \geq 2t+1$.

Attention. La distance de Hamming satisfait la loi du triangle:



$$d(x,y) \leq d(x,z) + d(z,y).$$

On a également : si $d(x,y) = s+t$, il existe $z (\in A^n)$ tel que

$$d(x,z) = s \text{ et } d(z,y) = t.$$

Strategies de décodage.

Avec une stategie de décoder r reçu à un mot de code dans $B(r,t)$ si $B(r,t) \cap C \neq \emptyset$, sinon d'annoncer l'existance des erreurs ("détecter" des erreurs : on peut :

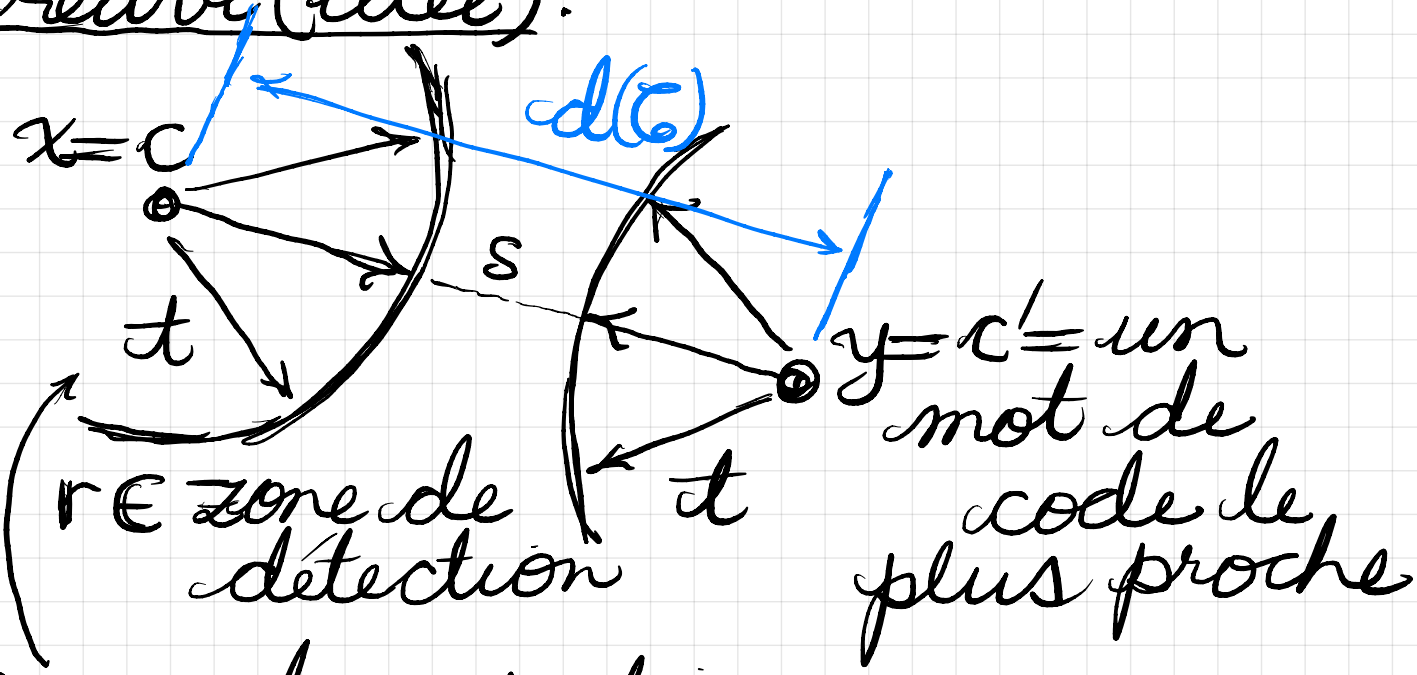
① corriger (correctement) t erreurs si $2t+1 \leq d(C)$,

② détecter s erreurs ($t=0$) si $s+1 \leq d(C)$, ou

③ corriger t erreurs et détecter entre $t+1$ et $s+t$ err.

si $2t+s+1 \leq d(C)$.

Preuve (idée):



Codes linéaires

Supposons que $A = \mathbb{Z}/q\mathbb{Z}$:

$\mathbb{Z}/q\mathbb{Z} := \{0, 1, \ldots, q-1\}$

avec addition et multiplication mod q (en prenant le reste après division par q).

Si q est premier (e.g. $q=2$), on écrit $\mathbb{F}_q$ pour $\mathbb{Z}/q\mathbb{Z}$, et $\mathbb{F}_q$ est un corps — tout élément non nul est inversible (comme $\mathbb{Q}$ (rationnels) $\mathbb{R}$ (réels) ou $\mathbb{C}$ (complexes)).

On dit que $\mathcal{C} \subseteq A^n = (\mathbb{Z}/q\mathbb{Z})^n = \mathbb{F}_q^n$ est <u>linéaire</u> si pour tout $x, y \in \mathcal{C}$ et $a, b \in \mathbb{F}_q$, alors $ax + by \in \mathcal{C}$.
Alors, on a $0 = (0, \dots, 0) \in \mathcal{C}$, en mettant $a = b = 0$, et $\mathcal{C}$ est un sous-espace vectoriel de $\mathbb{F}_q^n$.

<u>Defn</u> Si $A = \mathbb{Z}/q\mathbb{Z}$, on a $0 \in A$, élément distingué, ce qui permet de définir <u>le poids</u> de $x \in A^n$:

$$\|x\| = d(x, \underbrace{0}_{(0,\dots,0)}) = \#\{i : 1 \le i \le n, x_i \neq 0\}.$$

<u>Défn</u>. La <u>dimension</u> d'un codage linéaire sur un corps $\mathbb{F}_q$ est la taille minimum d'un sous-ensemble $\mathcal{B} \subseteq \mathcal{C}$ tel que

$$\mathcal{C} = \left\{ \sum_{x \in \mathcal{B}} a_x x : a_x \in \mathbb{F}_q \right\}.$$

Si $\mathcal{C}$ est un code linéaire sur $\mathbb{F}_q$ de dimension k, alors $|\mathcal{C}| = q^k$.

On rappelle...

Proposition Soit $C \subseteq \mathbb{F}_q^n$ un code linéaire. Alors $0=(0,\dots,0) \in C$ et

$$d(C) = \min\{d(x,y) : x \in C \setminus \{y\}\}$$

pour tout $y \in C$, et en particulier

$$d(C) = \min\{\|x\| : x \in C \setminus \{0\}\}.$$

Ex $C = \{000000, 010101, 101010, 111111\}$ est linéaire de dimension 2 $\leq \mathbb{F}_2^6$ sur $\mathbb{F}_2$, et $d(C) = 3$ (= poids non nul minimum).

Les boules de rayon $t=1$ autour des mots de code sont disjointes, donc

$$|B(C,1)| = |C||B(0,1)| = 4(1+6) = 28 < 2^6 = |\mathbb{F}_2^6| = 64.$$

Exercice. M.q. $B(C,2) = \mathbb{F}_2^6$.

Supposons qu'on utilise ce code pour transmission sur un canal binaire symétrique avec probabilité d'erreur ε.

Sans mémoire

Si c est le mot (de code) transmis et r le mot reçu, on a

$d(c,r)$	prob	
0	$(1-\varepsilon)^6$	$r \in B(c,1)$
1	$6(1-\varepsilon)^5\varepsilon$	
2	$15(1-\varepsilon)^4\varepsilon^2$	
3	$20(1-\varepsilon)^3\varepsilon^3$	dont 2 sont $\in \mathcal{C}$
4	$15(1-\varepsilon)^2\varepsilon^4$	dont 6 sont $\in$
5	$6(1-\varepsilon)\varepsilon^5$	tous dans $B(\mathcal{C},1)$
6	ε^6	dans $\mathcal{C}$

($r = c + 111111$)

mauvaise correction

Ce tableau nous permet de déterminer

(i) la probabilité d'une bonne correction (avec $t=1$)

(ii) la probabilité de détection d'erreur (non corrigé)

(iii) la probabilité d'une mauvaise correction (vers un mot de code $\neq c$).

Bornes sur les paramètres des codes

Les paramètres qui nous intéressent sont n = longueur ($\mathcal{C} \subseteq \mathbb{F}_q^n$), $k = \dim(\mathcal{C}) = \frac{\log_q |\mathcal{C}|}{\log q}$ et $d = d(\mathcal{C})$.

si linéaire, égalité

On dit que $\mathcal{C}$ est un code $[n,k,d]$. défn plus générale

Soit A un alphabet de q élément.

Théorème (Borne de Singleton). Soit $\mathcal{C}$ un code en bloc de longueur n ($\mathcal{C} \subseteq A^n$) et distance minimum d. Alors $|\mathcal{C}| \leq q^{n-d+1}$.

Remarque. Donc

$$k = \frac{\log |\mathcal{C}|}{\log q} \leq n - d + 1,$$

où k est un entier si $\mathcal{C}$ est linéaire.

Défn Si $|\mathcal{C}| = q^{n-d+1}$, on dit que $\mathcal{C}$ est MDS (separation de distance maximal)

Preuve. Soit π la projection

$$\begin{array}{ccc} \mathcal{C} & \xrightarrow{\pi} & A^{n-d+1} \\ \cap & & \cap \\ A^n & x_1 \dots x_n \longmapsto & x_1 \dots x_{n-d+1} \end{array}$$

La restriction à $\mathcal{C}$ est injective, car $\pi(x) = \pi(y)$ seulement si $d(x,y) \leq d-1$, alors $x = y \in \mathcal{C}$.

Donc

$$\mathcal{C} \xrightarrow{\pi} \pi(\mathcal{C}) \subseteq A^{n-d+1}$$

est une bijection, alors

$$|\mathcal{C}| = |\pi(\mathcal{C})| \leq |A^{n-d+1}| = q^{n-d+1} \quad \square$$

Ex Soit $\mathcal{C} = \langle 1001, 0101, 0011 \rangle \subseteq \mathbb{F}_2^4$.

Alors

$d(\mathcal{C}) = 2$ (car $d(\mathcal{C}) \leq 2$ et $d(\mathcal{C}) \neq 1$)

et $k = \dim_{\mathbb{F}_2} \mathcal{C} = 3$,

et donc

$$n - d + 1 = 4 - 2 + 1 = 3 = d(\mathcal{C}).$$

Alors $\mathcal{C}$ est MDS.

* Pourquoi?

* On observe que $\mathcal{C}$ est le sous espace de vecteurs de poids pair, donc $d(\mathcal{C})$ est pair.

Théorème (Borne de Hamming) Soit $\mathcal{C}$ un code en bloc de longueur n ($\mathcal{C} \subseteq A^n$) et distance minimum $d \geq 2t+1$. Alors

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

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

Preuve. Par l'hypothèse $d = d(\mathcal{C}) \geq 2t+1$, les boules de rayon t autour des mots de code sont disjointes, or

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

$\|$
q^n

rendement

Corollaire. Le taux de transmission de $\mathcal{C}$ est au maximum

$$1 - \frac{\log V(q,n,t)}{n \log q}.$$

Preuve. $R := \dfrac{\log |\mathcal{C}|}{n \log q} \leq \dfrac{\log V(q,n,t)}{n \log q}$, $1 -$

en prenant log de $(*)$, divisant par $n \log q$. $\square$

Défn Un code qui satisfait $|C| = q^n / V(q,n,t)$ s'appelle parfait.

Remarque. Un code est parfait ssi

- C est t-correcteur, avec $d(C) = 2t+1$, et
- les boules de rayon t autour de C couvrent tout A^n ($= B(C,t)$).

C'est à dire, qu'il n'y a pas d'espace entre les boules.