Codes correcteurs d'erreurs (au sens de Hamming)

Plan du cours

1. décodage au maximum de vraisemblance ; autres notions de décodage ;
2. codes linéaires ; codes de Reed-Solomon ;
3. bornes ;
4. codes concaténés.

1. Décodeur à maximum de vraisemblance

Soit un canal de communication $(\mathcal{A}, \mathcal{B}, \Pi)$.

Définition Soit C un code de longueur n sur A. Un *algorithme de décodage de* C est une procédure qui à tout élément de $\mathcal{B}^n$ associe un mot de C ou qui échoue (symbole ∞).

$$\begin{array}{rccc} \varphi : & \mathcal{B}^n & \to & C \cup \{\infty\} \\ & y & \mapsto & \varphi(y) \end{array}$$

Définition Un algorithme de décodage φ de C est dit à *maximum de vraisemblance* si pour tout $y \in B^n$, le mot $x = \varphi(y)$ est dans C et réalise le maximum de la probabilité $\mathbf{P}(\ll x \text{ émis} \gg \mid \ll y \text{ reçu} \gg)$.

Décodeur à maximum de vraisemblance

- Nécessite la connaissance du canal, du code et de la loi d'émission des x.

Mais

$$\mathbf{P}(\ll x \text{ émis} \gg \mid \ll y \text{ reçu} \gg) = \mathbf{P}(\ll y \text{ reçu} \gg \mid \ll x \text{ émis} \gg) \cdot \frac{\mathbf{P}(\ll x \text{ émis} \gg)}{\mathbf{P}(\ll y \text{ reçu} \gg)}$$

Les maximum coïncident quand $\mathbf{P}(\ll x \text{ émis} \gg)$ =constante= $1/|C| \Rightarrow$ hypothèse d'uniformité communément faite.

Canal q-aire symétrique

Canal symétrique $(\mathcal{A}, \mathcal{B}, \Pi)$ avec $\mathcal{A} = \mathcal{B}$, $|\mathcal{A}| = q$ et

$$\mathbf{P}_{\mathcal{B}|\mathcal{A}}(b|a) = \begin{cases} 1-p & \text{si } a = b \\ \frac{p}{q-1} & \text{si } a \neq b \end{cases} \qquad \begin{cases} p & \text{probabilité d'erreur} \\ \frac{p}{q-1} & \text{probabilité de transition} \end{cases}$$

Proposition 1. *Dans un canal q-aire symétrique sans mémoire de probabilité de transition $< 1/q$, si la loi d'émission des mots de code est uniforme, le mot $x \in C$ le plus probablement émis connaissant le mot reçu $y \in A^n$ est un mot réalisant le minimum de $d_H(x, y)$, où $d_H(x, y)$ est la distance de Hamming entre x et y : $d_H(x, y) \overset{\text{def}}{=} \#\{i | x_i \neq y_i\}$.*

Preuve

$$\mathbf{P}(y \mid x) = \left(\frac{p}{q-1}\right)^{d_H(x,y)} (1-p)^{n-d_H(x,y)}$$

Or $(1)\ \frac{p}{q-1} < \frac{1}{q} \implies p < 1 - \frac{1}{q} \implies 1 - p > \frac{1}{q} \implies \frac{1}{1-p} < q \quad (2)$

$$(1) \text{ et } (2) \implies \frac{p}{(q-1)(1-p)} < 1$$

$\Rightarrow$ recherche du mot de code le plus proche du mot reçu pour la distance de Hamming.

Diffentes problématiques

NCP (Nearest Codeword Problem, Maximum Likelihood Decoding)

LD (List Decoding) Une borne e est donnée. Le problème est de trouver *tous* (éventuellement aucun) les mots de code à distance e du mot reçu.

BDD (Bounded Distance Decoding) Une borne e est donnée. Le problème est de trouver *un* mot parmi les mots de code à distance e du mot reçu (s'il en existe).

UD (Unambiguous Decoding) Ici on se donne $e = (d-1)/2$, où d est la distance minimale du code, et on cherche le mot de code à distance e du mot reçu (s'il existe)

Distance minimale – Décodage

Soit $\mathcal{C}$ un code de distance minimale d.

— Deux boules de rayon $(d-1)/2$ centrées en deux mots de code distincts sont disjointes.
$\Rightarrow$ un code de distance minimale d peut corriger $\lfloor (d-1)/2 \rfloor$ erreurs

— Toute boule de rayon $d-1$ centrée en un mot de code ne contient aucun autre mot de code.
$\Rightarrow$ un code de distance minimale d peut détecter $d-1$ erreurs.

Performances

Définition Un algorithme de décodage φ d'un code C est dit *borné par t* si pour tout $x \in \mathcal{C}$, $d_H(x, y) \leq t \Rightarrow \varphi(y) = x$

Si la réciproque est vraie, l'algorithme est dit *strictement borné*. Tout code de distance minimale d possède un algorithme de décodage borné par $\lfloor (d-1)/2 \rfloor$.

Proposition 2. *La probabilité d'erreur après décodage d'un mot à travers un canal de probabilité d'erreur p décodé à l'aide d'un algorithme strictement borné par t est*

$$\sum_{i=t+1}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

2. Rappel : corps finis

Un corps fini $\mathbb{F}_q$ est un ensemble de cardinal q, avec les lois $(+, -, \times, /)$.

- On a nécessairement $q = p^m$, p premier.
- Structure : $\mathbb{F}_{p^m} = \mathbb{F}_p[X]/P_m(X)$ où P est un polynôme irréductible sur $\mathbb{F}_p[X]$ de degré m.

Exemple :

$$\begin{aligned} \mathbb{F}_4 &= \mathbb{F}_2[X]/(1+X+X^2) = \{0, 1, X, 1+X\} \\ X(1+X) &= X^2 + X \equiv 1 \bmod (1+X+X^2) \end{aligned}$$

Codes linéaires

Lorsque l'alphabet est un corps fini (par exemple $\mathcal{A} = \mathbf{F}_2 = \{0, 1\}$) l'espace de Hamming $\mathcal{A}^n$ est un espace vectoriel.

Définition Un code en bloc linéaire de longueur n sur $\mathbb{F}_q$ (le corps fini à q élément) est un sous-espace vectoriel de $\mathbb{F}_q^n$.

Nous parlerons de code $[n, k]_q$ si le code est de dimension k et de code $[n, k, d]_q$ si sa distance minimale est d.

La cardinal d'un code linéaire est q^k, son taux de transmission est donc

$$\frac{\log_q q^k}{n} = \frac{k}{n}$$

Les deux matrices

Un code $\mathcal{C}[n,k]_q$ peut se caractériser par

— une matrice génératrice G (de taille $k \times n$ sur $\mathbb{F}_q$) :

$$\mathcal{C} = \{(u_1, \ldots, u_k)G \mid (u_1, \ldots, u_k) \in \mathbb{F}_q^k\}$$

Les lignes de G forment une base de $\mathcal{C}$.

— ou une matrice de parité H (de taille $(n-k) \times n$ sur $\mathbb{F}_q$) :

$$\mathcal{C} = \{(x_1, \ldots, x_n) \in \mathbb{F}_q^n \mid H(x_1, \ldots, x_n)^T = 0\}$$

Les lignes de H forment une base du dual $\mathcal{C}^\perp$ de $\mathcal{C}$:

$$\mathcal{C}^\perp \stackrel{\text{def}}{=} \{(v_1, \ldots, v_n) | \forall (c_1, \ldots, c_n) \in \mathcal{C}, \sum_{i=1}^{n} v_i c_i = 0\}.$$

Codes linéaires – Propriétés

Proposition 3. *Pour tout code linéaire $\mathcal{C}$*

$$\min_{x \neq y | x,y \in \mathcal{C}} d_H(x,y) = \min_{x \neq 0 | x \in \mathcal{C}} w_H(x)$$

(la distance minimale est égale au poids minimal)

Proposition 4. *Soit $\mathcal{C}$ un code de matrice de parité H*

$$\left(\begin{array}{c} \mathcal{C} \text{ de distance} \\ \text{minimale} \geq d \end{array} \right) \Leftrightarrow \left(\begin{array}{c} d-1 \text{ colonnes quelconques} \\ \text{de } H \text{ sont libres} \end{array} \right)$$

Codes linéaires – Décodage par syndrome

À toute matrice de parité H de $\mathcal{C}$ on associe le syndrome

$$\begin{array}{rccc} \sigma : & \mathbb{F}_q^n & \to & \mathbb{F}_q^{n-k} \\ & y & \mapsto & Hy^T \end{array}$$

Nous noterons $\sigma^{-1}(s) = \{y \in \mathbb{F}_q^n \mid \sigma(y) = s\}$. Nous avons

$$\sigma^{-1}(Hy^T) = y + \mathcal{C} = \{y + c \mid c \in \mathcal{C}\}$$

Pour tout $s \in \mathbb{F}_q^{n-k}$, notons $L_H(s)$ un mot de poids minimal de $\sigma^{-1}(s)$.

Proposition 5. *Le décodeur $y \mapsto y - L_H(Hy^T)$ est à maximum de vraisemblance.*

Décodage par syndrome

Décodage par tableau standard : mettre $L_H(s)$ en table (précalcul).

Décodage algébrique : calculer de manière algébrique $L_H(s)$ pour certaines valeurs de s.

Exemple

Code de Hamming $[7, 4, 3]$.

- Peut être obtenu en répondant à la question : on veut le plus long code linéaire et binaire $[n, k, 3]$ qui soit tel que $n - k = 3$.

Le code de Hamming

Code de Hamming de $[7,4]_2$. Matrice de parité :

$$H = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

Pour un mot reçu $y \in \mathbb{F}_2^7$, il y a 8 syndromes possibles

$$Hy^T \in \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\},$$

$\Rightarrow$ tout mot de l'espace de Hamming $\{0,1\}^n$ s'écrit $x+e$ avec x dans le code et e de poids a au plus 1 (code parfait $=$ code de distance d dont les boules de rayon $\lfloor \frac{d-1}{2} \rfloor$ centrées sur les mots de code forment une partition de l'espace ambiant).

Code de Hamming binaire de longueur $2^m - 1$

C'est le code de longueur $2^m - 1$, de dimension $2^m - m - 1$, dont les colonnes de la matrice de parité sont tous les vecteurs de $\mathbb{F}_2^m \setminus \{0\}$.

C'est un code $[n = 2^m - 1, k = 2^m - m - 1, d = 3]$ parfait

$$\begin{aligned} 2^{2^m-m-1}\left(\binom{n}{1}+1\right) &= 2^{2^m-m-1}2^m \\ &= 2^n \end{aligned}$$

Théorème 1. *Les paramètres des codes parfaits sont connus : ceux du code à répétition de longueur impaire, ceux du Hamming, ceux du Golay binaire G_{23} $[23, 12, 7]_2$ et ceux du Golay ternaire G_{11}, $[11, 6, 5]_3$*

3. Bornes : borne de Singleton

Proposition 6. (Borne de Singleton)
Pour tout code $[n, k, d]$ *nous avons* $d \leq n - k + 1$.

Preuve

La matrice de parité est de rang $n-k$. Dans le meilleur des cas, toutes les familles de $n-k$ colonnes sont libres :

$$\implies d-1 \leq n-k.$$

Un code tel que $k+d=n+1$ est MDS (Maximum Distance Separable).

Codes de Reed-Solomon

Codes définis sur des gros alphabets $\mathbb{F}_q$. On prend $x_1, \ldots, x_n \in \mathbb{F}_q$ distincts.

Soit ev la fonction d'évaluation :

$$\begin{array}{rccl} \mathrm{ev}: & \mathbb{F}_q[X] & \to & \mathbb{F}_q^n \\ & f & \mapsto & \mathrm{ev}(f) = (f(x_1), \ldots, f(x_n)) \end{array}$$

et

$$L = \{f \in \mathbb{F}_q[X] \mid \deg f < k\}.$$

Alors le code de Reed-Solomon de dimension k est

$$C \stackrel{\text{def}}{=} \mathrm{ev}(L).$$

Paramètres

Proposition 7. *Si $k < n$, c'est un code de dimension k et de distance minimale $d = n - k + 1$, corrigeant $t = \lfloor \frac{n-k}{2} \rfloor$ erreurs.*

Preuve :

— Si $n > k$, alors ev est injectif.

— un polynôme de degré $< k$ a au plus $k - 1$ zéros. Il a donc au moins $n - k + 1$ composantes non nulles dans un mot de code non nul.

De plus, le polynôme $\prod_{i=1}^{k-1}(X - x_i)$ a exactement $k - 1$ racines. Les codes de Reed-Solomon sont MDS.

Décodage par interpolation des codes de Reed-Solomon

Soit $y = (y_1, \ldots, y_n)$ le mot reçu. Soit c le mot de code le plus proche, avec $c = \mathrm{ev}(f(X))$ où $\deg f(X) < k$.

Soit I l'ensemble des positions où il y a une erreur :

$$I = \{i \in \{1, \ldots, n\}, \quad f(x_i) \neq y_i\},$$

et construisons le polynôme $E(X) = \prod_{i \in I}(X - x_i)$. Alors nous avons

$$E(x_i)y_i = E(x_i)f(x_i), \quad i \in \{1, \ldots, n\}. \tag{1}$$

Décodage (II)

On pose

$$X^t + \sum_{i=0}^{t-1} e_i X^i \quad \stackrel{\text{def}}{=} \quad E(X)$$

$$\sum_{i=0}^{t+k-1} a_i X^i \quad \stackrel{\text{def}}{=} \quad E(X)f(X)$$

- $2t + k$ inconnues et n équations affines :

$$E(x_i)y_i = E(x_i)f(x_i), \quad i \in \{1, \ldots, n\}. \tag{2}$$

- On peut espérer corriger ainsi $\frac{n-k}{2} = \frac{d-1}{2}$ erreurs.

Borne de Hamming

Soit C un code de cardinal M, et de capacité de correction $t = \lfloor \frac{d-1}{2} \rfloor$, de longueur n sur l'alphabet $\mathbb{F}_q$. Alors

$$M \left(\sum_{i=0}^{t} (q-1)^i \binom{n}{i} \right) \leq q^n$$

Forme asymptotique

$$\begin{aligned} h_q(\delta/2) &\leq 1 - R \text{ avec} \\ \delta &\stackrel{\text{def}}{=} d/n \\ R &\stackrel{\text{def}}{=} \log_q M/n \\ h_q(x) &\stackrel{\text{def}}{=} -x \log_q \frac{x}{q-1} - (1-x) \log_q (1-x) \end{aligned}$$

Existence de bons codes – Borne de Varshamov-Gilbert

Théorème 2. ***(Borne de Varshamov-Gilbert)***

$$\sum_{i=0}^{d-2}(q-1)^i\binom{n-1}{i} < q^{n-k} \Rightarrow \begin{pmatrix} \textit{il existe un} \\ \textit{code } [n,k,d]_q \end{pmatrix}$$

Théorème 3. ***(Borne de Varshamov-Gilbert asymptotique)*** *Soit* $0 \leq \delta \leq (q-1)/q$. *Pour tout* $0 \leq R < 1 - h_q(\delta)$ *il existe une infinité de codes* $[n,k,d]_q$ *tels que* $d \geq \delta n$ *et* $k \geq Rn$ *où* $h_q(x) \stackrel{\text{def}}{=} -x\log_q\frac{x}{q-1} - (1-x)\log_q(1-x)$ *est la fonction d'entropie* q*-aire.*

Preuve

On construit progressivement, en rajoutant les colonnes une à une, une matrice de parité $r \times n$ avec la propriété que tout sous-ensemble de $d-1$ colonnes sont linéairement indépendantes.

Supposons maintenant que les i premières colonnes sont telles que tout sous ensemble de taille $d-1$ est linéairement indépendant.

Nombre N de combinaisons de $d-2$ colonnes ou moins, parmi i colonnes :

$$1 + \binom{i}{1}(q-1) + \cdots + \binom{i}{d-2}(q-1)^{d-2}$$

Si $N \leq q^r - 1$, on peut ajouter une colonne qui ne soit pas combinaison linéaire de $d-2$ ou moins colonnes.

On peut donc le faire tant que

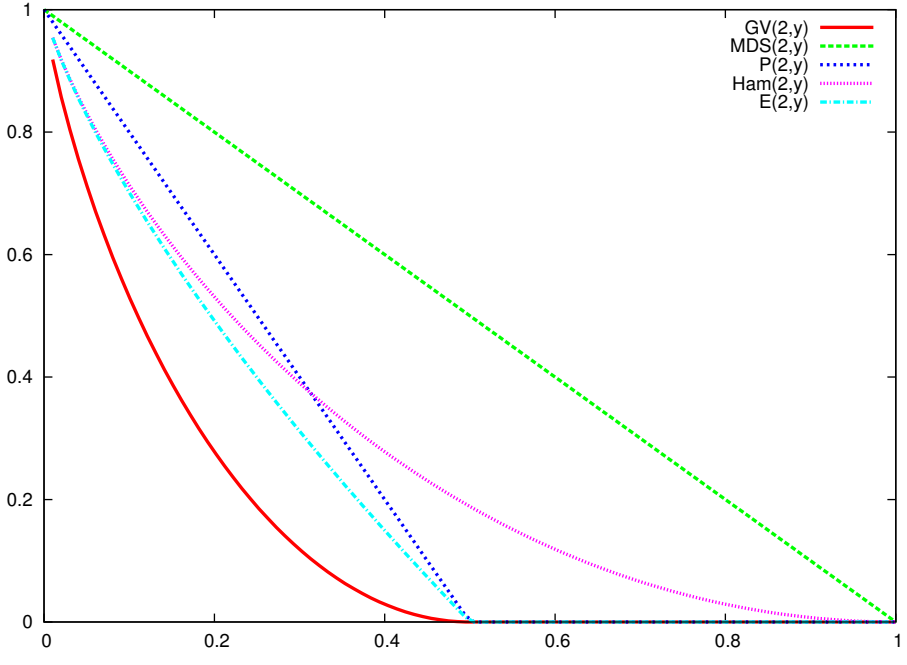
$$1 + \binom{i}{1}(q-1) + \cdots + \binom{i}{d-2}(q-1)^{d-2} < q^r$$

On finit la preuve avec les bornes pour $i \leq n-1$:

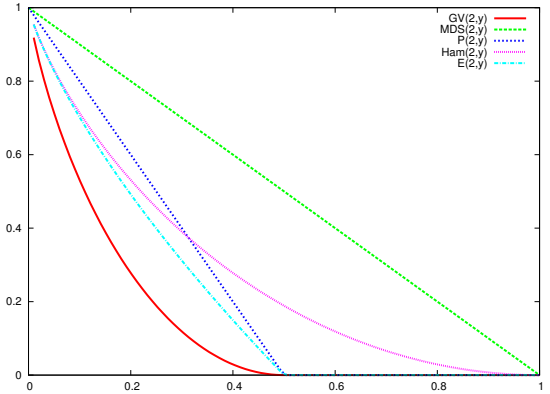
$$\begin{aligned}
& 1 + \binom{i}{1}(q-1) + \cdots + \binom{i}{d-2}(q-1)^{d-2} \\
\leq\ & 1 + \binom{n-1}{1}(q-1) + \cdots + \binom{n-1}{d-2}(q-1)^{d-2} \\
\leq\ & 2^{(n-1)h_q\left(\frac{d-2}{n-1}\right)} && (3) \\
\leq\ & 2^{nh_q\left(\frac{d}{n}\right)} && (4)
\end{aligned}$$

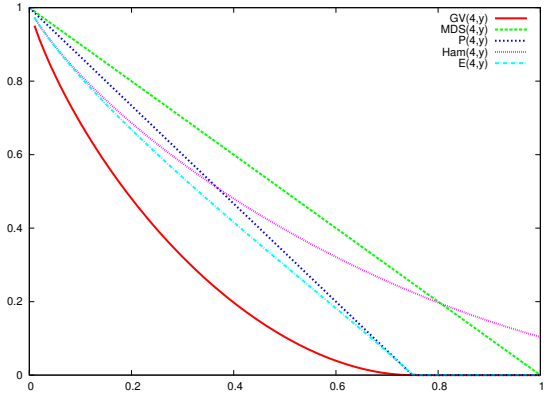
Exercice : Montrer (3) par une preuve utilisant les propriétés de l'entropie.

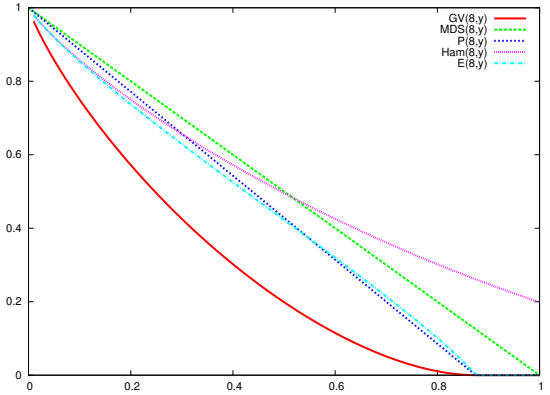
Varshamov-Gilbert – Cas binaire

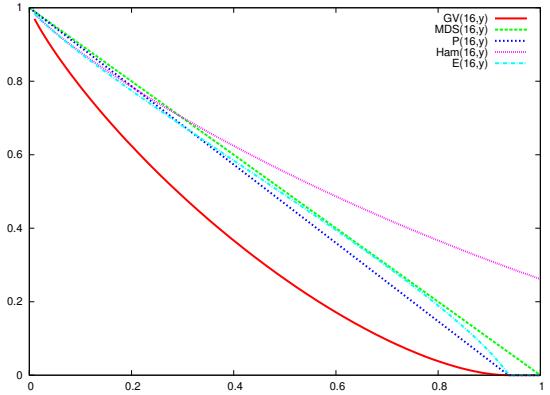


Courbes de 2 à 16









Codes atteignant les bornes

- la borne de Hamming : les codes parfaits : codes à répétition, codes de Hamming, codes de Golay binaires et ternaires.
- la borne de Singleton : les codes MDS : codes de Reed-Solomon ($n \leq q$).
- la borne de Varshamov-Gilbert : presque tous les codes...

Correction en moyenne/pire des cas

Sur un canal binaire symétrique de probabilité d'erreur p on a typiquement environ $\approx pn$ erreurs pour un code de longueur n et de rendement R.

On peut corriger presque toujours tant que $R < 1 - h(p)$, c.a.d.

$$p < h^{-1}(1 - R).$$

Or la distance d d'un code linéaire de longueur n et rendement R est presque toujours de la forme $d \approx nh^{-1}(1-R)$. On peut corriger $t = \frac{d-1}{2}$ erreurs dans tous les cas avec un tel code. Noter que dans ce cas

$$\frac{t}{n} \approx \frac{h^{-1}(1-R)}{2}$$

On corrige deux fois moins d'erreurs dans le pire cas qu'en moyenne.

4. Code concaténés

Idée : utiliser un deuxième niveau de codage pour réduire la probabilité d'erreur après décodage.

$$M \in \{0,1\}^k \xrightarrow{\text{codage}} \boxed{\underbrace{C \in \{0,1\}^n \xrightarrow{\text{canal}} C' \in A^n}_{\text{canal virtuel}}} \xrightarrow{\text{decodage}} \hat{M} \in \{0,1\}^k$$

Soit $B \stackrel{\text{def}}{=} \{0,1\}^k$, on choisit un code $[N, ?]$ sur B pour protéger les mots du code binaire (vus maintenant comme des symboles de B).

Codage

$$M = (x_1 \dots x_K) \in B^K \xrightarrow{\text{codage externe}} C = (y_1 \dots y_N) \in B^N$$

$$\xrightarrow{\text{codage interne des } y_i} C' = (c_1 \dots c_{nN}) \in \{0,1\}^{nN}$$

Code externe : code de longueur N et rendement $\frac{K}{N}$ sur $B = F_{2^k}$,

Code interne : code binaire de longueur n et rendement k/n.

$$\text{Rendement du code concatenaté} = \frac{kK}{nN}$$

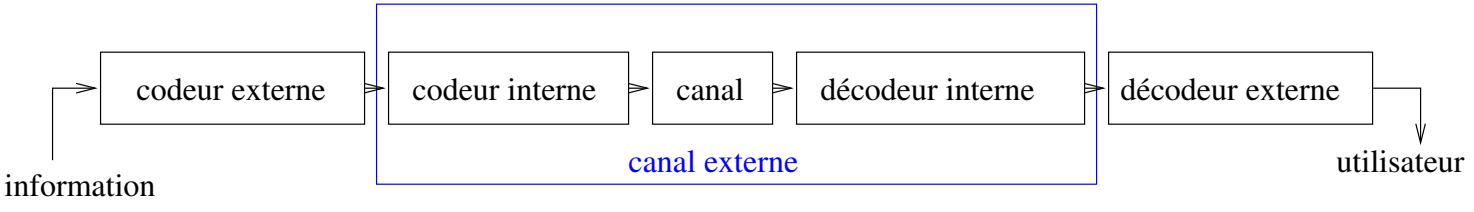
Décodage

$$C' = (c_1 \dots c_{nN}) \in \{0,1\}^{nN} \xrightarrow{\text{canal}} W = (a_1 \dots a_{nN}) \in A^{nN}$$

$$\xrightarrow{\text{décodage interne}} C'' = (y'_1 \dots y'_N) \in B^N$$

$$\xrightarrow{\text{décodage externe}} M' = (x'_1 \dots x'_k) \in B^K.$$

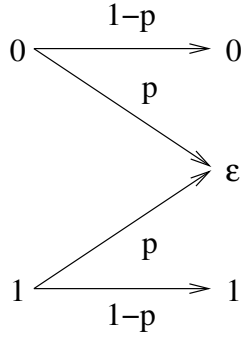
Dessin



1. On code un mot dont les symboles sont sur un gros alphabet
2. On code chaque symbole par un code interne
3. On transmet
4. On décode chaque symbole
5. On décode le mot dont les symboles sont sur un gros alphabet

Codes concaténés, outil : correction d'effacements

Un effacement est une « erreur localisée », c'est-à-dire que les symboles sont transmis à travers le canal suivant :



Pour tout code de distance minimale d, il existe un algorithme de décodage corrigeant $d-1$ effacements.

(Il n'existe qu'un mot de code coïncidant avec le mot reçu sur les positions non effacées)

Correction d'erreurs et d'effacements

Proposition 8. *Pour tout code de distance minimale d, il existe un algorithme de décodage corrigeant ν erreurs et ρ effacements ssi*

$$2\nu + \rho < d$$

Soit J l'ensemble des positions non effacées, et

$$C_J = \{c_J; \quad c \in C\}$$

La distance minimale de C_J est $\geq d - \rho$.

On peut donc décoder 2ν erreurs, si $2\nu < d - \rho$.

Ensuite, on corrige les effacements.

Codes concaténés – Décodage

Le mot reçu est de la forme

$$\underbrace{(y_{1,1},\ldots,y_{1,n})}_{y_1} \;\|\; \underbrace{(y_{2,1},\ldots,y_{2,n})}_{y_2} \;\|\; \cdots \;\|\; \underbrace{(y_{N,1},\ldots,y_{N,n})}_{y_N}$$

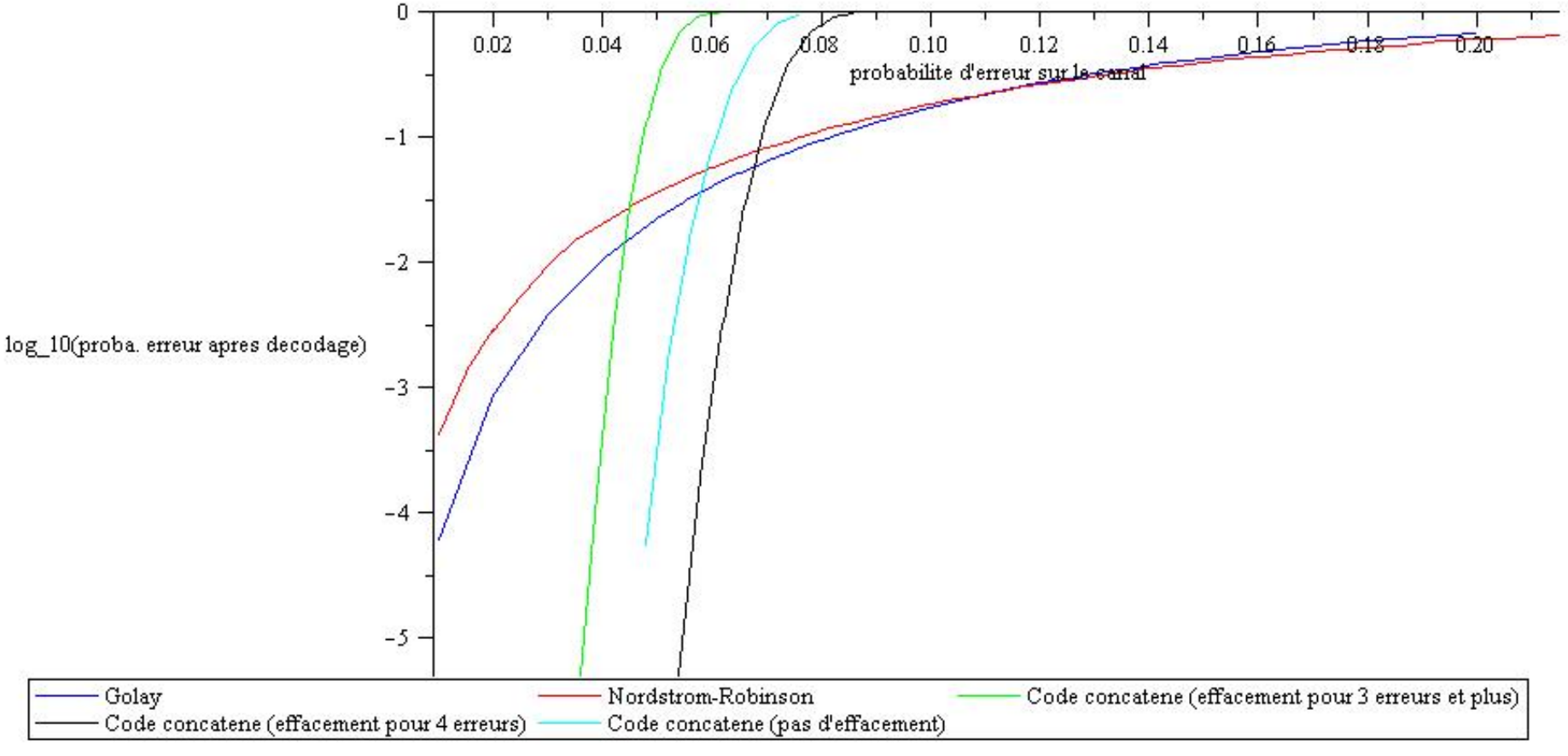
Chacun des N blocs est décodé avec le décodeur interne

$$\begin{array}{rccc} \varphi & : & \{0,1\} & \to & C_{\text{interne}} \cup \{\infty\} \\ & & y_i & \mapsto & z_i \end{array}$$

Chaque lettre du mot $(z_1,\ldots,z_N) \leftrightarrow$ un symbole de A ou bien avec un effacement (symbole ∞). L'ensemble est ensuite décodé à l'aide d'un décodeur d'erreurs et d'effacements du code externe C_{externe}.

$\Rightarrow$ pas forcément optimal d'utiliser le code interne au maximum de sa capacité de décodage, $\exists$ optimum nombre d'erreurs / nombre d'effacements.

TD


0
-1
-2
-3
-4
-5
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20
probabilite d'erreur sur le canal
log_10(proba. erreur apres decodage)
Golay
Nordstrom-Robinson
Code concatene (effacement pour 3 erreurs et plus)
Code concatene (effacement pour 4 erreurs)
Code concatene (pas d'effacement)