Réseaux, information et communications (INFO-F303)
Partie Théorie de l'Information
6. Codes correcteurs d'erreurs

Christophe Petit

Université libre de Bruxelles

Plan du cours

Contexte

- Messages transmis via un canal bruité
- Introduire de la redondance dans le message source pour compenser le bruit du canal
- Cas idéal : décodage déterministe

$$\mathbb{P}[x_i | y_j] \in \{0, 1\} \text{ pour tout } i, j$$

- En pratique : probabilité non nulle d'erreur au décodage, qu'on veut minimiser

Codes correcteurs d'erreurs

- But : introduire de la redondance dans le codage pour compenser les erreurs introduites par le canal bruité
- Code à répétition
- Fiabilité ; capacités de détection et correction d'erreurs
- Code par somme de contrôle
- Décodage par maximum de vraisemblance
- Débit d'un code (information minimale par symbole)
- Second théorème de Shannon : pour tout canal, il existe une famille de codes avec (asymptotiquement) un taux d'erreur nul et un débit égal à la capacité du canal
- Inverse du théorème de Shannon (borne de Fano) : information transmise limitée par la capacité du canal

Codes correcteurs d'erreurs (suite)

- Distance minimale d'un code ; liens avec les capacités de détection et correction d'erreurs
- Borne de Singleton
- Rayons d'empilement et de recouvrement
- Borne de Hamming et codes parfaits (i.e. codes atteignant cette borne)
- Borne de Gilbert-Varshamov

Exemple : code à répétition

- Alphabet C de r caractères
- Canal avec probabilité $p < \frac{1}{2}$ d'erreur
- Code à répétition

$$K : C \to C^n : c \mapsto \overbrace{cc \dots c}^{n}$$

- Décodage par majorité
- Quelle est la probabilité d'erreur de décodage ?

Code à répétition : probabilité d'erreur

- Erreur maximale dans le cas $r = 2$
(toutes les erreurs sur un caractère ont la même valeur)
- Nombre d'erreurs pour $r = 2 \sim$ distribution binomiale

$$\mathbb{P}[k \text{ erreurs}] = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mathbb{E}[\#\text{erreurs}] = np \quad \text{et} \quad \mathrm{Var}[\#\text{erreurs}] = np(1-p)$$

- Inégalité de Tchebychev $\mathbb{P}\left[|X - \mathbb{E}[X]| \geq \alpha\right] \leq \frac{\mathrm{Var}[X]}{\alpha^2}$ donne

$$\mathbb{P}_{\text{err}} \leq \frac{p\,(1-p)}{n\,(\frac{1}{2}-p)^2}$$

Fiabilité

- **Fiabilité d'un canal bruité** est $1 - \mathbb{P}_{\text{err}}$ probabilité qu'un message puisse être décodé correctement s'il a été transmis via ce canal

- Pour le code à répétition

$$\mathbb{P}_{\text{err}} \leq \frac{p\,(1-p)}{n\,(\frac{1}{2}-p)^2}$$

quand $p \neq 1/2$ on a $n \to \infty \Leftrightarrow \mathbb{P}_{\text{err}} \to 0$

- Peut-on faire mieux à moindre coût ?

Code correcteur

- Un **code correcteur** de t erreurs (error-correcting code) est un code en bloc de longueur n ($K \subset C^n$) pour lequel un mot reçu est toujours correctement décodé si au plus t caractères ont été altérés
- t est la **capacité de correction** du code

Maximum de vraisemblance

- Soit les probabilités $p_{ij} := \mathbb{P}[x_i|y_j]$
- Soit une fonction de décodage Cor : $y_j \to x_i = \mathrm{Cor}(y_j)$
- L'erreur de décodage pour cette fonction est

$$\begin{aligned}\mathbb{P}_{\text{err}} &= \sum_j \mathbb{P}[y_j] \sum_{x_i \neq \mathrm{Cor}(y_j)} \mathbb{P}[x_i|y_j] \\ &= \mathbb{P}_{\text{err}} = \sum_j \mathbb{P}[y_j]\left(1 - \mathbb{P}[\mathrm{Cor}(y_j)|y_j]\right)\end{aligned}$$

- Cette erreur est minimale pour la fonction de décodage par **maximum de vraisemblance**

$$\begin{aligned}\mathrm{Cor}: \quad & C^n \to K \cup \{\perp\} \\ & y \mapsto \begin{cases} \hat{x} : \mathbb{P}[\hat{x}|y] = \max_{x \in K} \mathbb{P}[x|y] & \text{s'il est unique} \\ \perp & \text{sinon} \end{cases}\end{aligned}$$

Maximum de vraisemblance pour source uniforme

- Si $\mathbb{P}[x_i] = \frac{1}{m}$ on a

$$\mathrm{Cor}(y_j) = \hat{x} \;\Rightarrow\; \mathbb{P}[y_j|\hat{x}] = \max_i \mathbb{P}[y_j|x_i]$$

- Preuve : on a $\mathbb{P}[x_i|y_j] = \frac{\mathbb{P}[x_i].\mathbb{P}[y_j|x_i]}{\mathbb{P}[y_j]} = \frac{\mathbb{P}[y_j|x_i]}{m\cdot\mathbb{P}[y_j]}$

Détection et correction

- Un **code détecteur** de τ erreurs (error-detecting code) est un code en bloc de longueur n ($K \subset C^n$) pour lequel un mot reçu dont au plus τ caractères ont été altérés n'appartient jamais au code (et peut donc être détecté comme erroné)
- Pour tout code en bloc,

$$t \leq \tau$$

- Exemple : pour code à répétition $t = \left\lfloor \frac{n-1}{2} \right\rfloor$ et $\tau = n - 1$

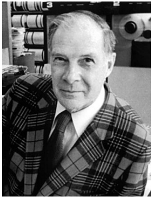
Code à somme de contrôle (checksum)

- Un code à somme de contrôle de longueur n sur un alphabet C de r caractères transmet des messages de $(n-1)$ caractères $c_1 c_2 \ldots c_{n-1}$ complétés d'un n-ième caractère c_n garantissant la relation

$$\sum_{i=1}^{n} c_i \equiv 0 \pmod{r}$$

- Détecte une erreur (mais ne peut pas la corriger)

Distance de Hamming



$$\begin{array}{rcl} \mathrm{d}_H : C^n \times C^n & \to & \mathbb{N} \\ (x, y) & \to & \mathrm{d}_H(x, y) = \#\{i \mid x_i \neq y_i\} \end{array}$$

- Distance au sens mathématique : satisfait les axiomes de séparation, de symétrie, et l'inégalité triangulaire
- Si $0 \in C$, le **poids de Hamming** d'un n-uplet est sa distance au n-uplet nul

$$\mathrm{w}_H(x) = \mathrm{d}_H(x, \vec{0}) = \#\{i \mid x_i \neq 0\}$$

Boules de Hamming

- La **boule de Hamming** de rayon ϱ centrée en c est l'ensemble des éléments à distance inférieure ou égale à ϱ

$$\mathcal{B}(c, \varrho) = \{x \in C^n \mid d_H(x, c) \leq \varrho\}$$

- Le nombre de mots B_ϱ dans la boule de Hamming $\mathcal{B}(c, \varrho)$ est

$$B_\varrho = \sum_{i=0}^{\varrho} \binom{n}{i} (r-1)^i$$

Boules de Hamming : exemple

- Soit le mot 10120 sur l'alphabet $\{0, 1, 2\}$
- Quels sont les mots à distance exactement 1 de 10120 ?
- Combien y a-t-il de mot dans la boule centrée en 10120 et de rayon 2 ?

Distance de Hamming et décodage

- Hypothèses supplémentaires courantes
  - Canal symétrique, probabilité d'erreur $p < \frac{1}{2}$
  - Code en bloc K, mots émis équiprobables
- Alors : méthode du maximum de vraisemblance donne

$$\mathrm{Cor}\colon y \mapsto \hat{x} : \mathrm{d}_H(\hat{x}, y) = \min_{x \in K} \mathrm{d}_H(x, y)$$

- Preuve :
  - Mots équiprobables donc $\mathrm{Cor}(y) = \arg\max_{x \in K} \mathbb{P}[y|x]$
  - $\mathbb{P}[y|x] = p^d (1-p)^{n-d}$ avec $d = \mathrm{d}_H(x, y)$

Débit d'un code (*code rate*)

- Le **débit** d'un code K sur un alphabet de r caractères et de longueur maximale $L_{\max}(K)$ est

$$R(K) = \frac{\log_r |K|}{L_{\max}(K)} \leq 1$$

- Pour un code en bloc de longueur n qui encode tous les mots de longueur $k \leq n$

$$R(K) = k/n$$

- On parle de **code** (n, k)
- Différence $n - k$ est la **redondance**

Rappel : capacité d'un canal

- **Capacité** d'un canal de communication est le maximum de l'information mutuelle, sur l'ensemble des distributions pour X
$$\mathcal{C} = \max_{\mathbb{P}[X]} \mathcal{I}(X, Y)$$

- Pour le canal symétrique avec probabilité d'erreur p on a
$$\mathcal{C}_r = 1 - \mathcal{H}_r(p)$$

Second théorème de Shannon

"il existe un code qui transmet sans erreur sur un canal bruité, tant que le débit du code est inférieur à la capacité du canal"

- Soit un canal bruité symétrique de probabilité d'erreur $p < \frac{1}{2}$, donc de capacité $\mathcal{C}_r = 1 - \mathcal{H}_r(p)$
- Shannon : il existe une famille $\{K_n\}$ de codes en bloc de longueur n, de probabilité d'erreur de décodage $\mathbb{P}_{\text{err}}(K_n)$ et de débit $R(K_n) < \mathcal{C}$ telle que, simultanément,

$$\lim_{n\to\infty} \mathbb{P}_{\text{err}}(K_n) = 0 \quad \text{et} \quad \lim_{n\to\infty} R(K_n) = \mathcal{C}$$

Second théorème de Shannon : remarques

- Pour le code à répétition on a

$$\lim_{n\to\infty} \mathbb{P}_{\text{err}}(K_n) = 0 \quad \text{mais} \quad \lim_{n\to\infty} R(K_n) = 0$$

- Wolfowitz : pour toute famille de codes, si $R(K_n) \geq \mathcal{C}_r + \epsilon$ (avec $\epsilon > 0$ fixé), alors $\lim_{n\to\infty} \mathbb{P}_{\text{err}}(K_n) = 1$

Second théorème de Shannon : intuitions

- Pour tout $\delta > 0$ et n assez grand, il existe k_n tel que

$$\mathcal{C} - \delta \leq \frac{k_n}{n} \leq \mathcal{C}$$

- Code (n, k_n) **aléatoire** satisfait le théorème (avec une grande probabilité)
  - Code aléatoire = choix aléatoire de r^{k_n} mots dans C^n
  - r^{k_n} mots choisis uniformément dans C^n sont en moyenne bien distribués, éloignés les uns des autres
  - Plus d'erreurs de décodage si beaucoup de mots proches
  - Moins d'erreurs si p petit ($\Leftrightarrow$ capacité $\mathcal{C}$ grande)

"Preuve" approximative

- Mots erronnés dans des boules de Hamming $\mathcal{B}(c, np)$
- On a

$$
\begin{aligned}
|\mathcal{B}(c,np)| &= \sum_{i=0}^{np} \binom{n}{i} (r-1)^i \approx \binom{n}{np} \approx \frac{e^{n \log n}}{e^{np \log np} e^{n(1-p) \log n(1-p)}} \\
&= e^{-np \log p - n(1-p) \log(1-p)} = r^{n\mathcal{H}_r(p)}
\end{aligned}
$$

- Boules disjointes si

$$
r^{n\mathcal{H}_r(p)} \cdot r^k \leq r^n
$$

ou

$$
\mathcal{C} = 1 - \mathcal{H}_r(p) \geq \frac{k}{n}
$$

Théorie des codes

- Shannon : *"il existe un code qui transmet sans erreur sur un canal bruité, tant que le débit du code est inférieur à la capacité du canal"*
- Théorème d'existence, preuve non constructive
- Objectif : codes avec bon débit et capacité de correction, et des fonctions d'encodage et de décodage efficaces
- Jusqu'au début des années '90 : codes algébriques
- Codes LDPC, turbocodes : débit approchant la capacité (pas couverts dans ce cours)

Distance minimale

- La **distance minimale** d'un code est la plus petite distance de Hamming séparant deux mots distincts du code

$$d = \mathrm{d}(K) = \min\{\mathrm{d}_H(x, y) \mid x, y \in K,\ x \neq y\}$$

- Le **poids d'un code** est le plus petit poids de Hamming des mots non nuls du code

Distance minimale et détection, correction

- Un code K de distance minimale d **détecte** τ erreurs si et seulement si
$$\tau < d$$

- Un code K de distance minimale d **corrige** t erreurs si et seulement si
$$t < \frac{d}{2}$$

Borne de Singleton

- Pour tout code K de longueur n et distance minimale d, on a
$$|K| \leq r^{n-d+1}$$
- Si K code tous les mots de k caractères alors $|K| = r^k$ et
$$d \leq n - k + 1$$

Borne de Singleton : démonstration

- Considérons les mots d'un code K' obtenus en supprimant les $(d-1)$ derniers symboles de chacun des mots de K
- On a $|K'| \leq r^{n-(d-1)}$
- Puisque la distance minimale de K est d, les mots de K' sont distincts et de distance minimale ≥ 1
- Donc $|K'| = |K|$
- On déduit $|K| \leq r^{n-(d-1)}$

Rayons d'empilement et de recouvrement

- On note $\Gamma(K, \varrho)$ l'ensemble des boules de Hamming centrées sur les mots de K et de rayon ϱ
- **Rayon d'empilement** (packing radius) d'un code K est le plus grand rayon s tel que les boules de $\Gamma(K, s)$ sont disjointes
- **Rayon de recouvrement** (covering radius) d'un code K est le plus petit rayon c tel que les boules de $\Gamma(K, c)$ recouvrent l'espace C^n
- On a

$$s = \left\lfloor \frac{d-1}{2} \right\rfloor \leq \left\lfloor \frac{d}{2} \right\rfloor \leq c$$

Rayons d'empilement et de recouvrement

- On a

$$s = \left\lfloor \frac{d-1}{2} \right\rfloor \le \left\lfloor \frac{d}{2} \right\rfloor \le c$$

- L'inégalité de droite peut être largement dépassée : si $K = \{00\ldots01, 00\ldots00\}$, on a $d = 1$ mais $c = n - 1$
- Un code est **parfait** si le rayon d'empilement est égal au rayon de recouvrement

$$t = c = s, \qquad d = 2t + 1$$

Borne de Hamming

- Borne de Hamming :

$$|K| \leq \frac{r^n}{B_s}$$

où $s = \lfloor (d-1)/2 \rfloor$ est le rayon d'empilement et B_s est le nombre de mots dans une boule de Hamming de rayon s

- Un code K est parfait si et seulement si

$$|K| = \frac{r^n}{B_s}$$

Code maximal

- Un code est dit **maximal** si tout ajout de mot dans le code réduit sa distance minimale
- Exemples :
  - $K = \{000, 001, 100\}$ n'est pas maximal
  - $K = \{000, 111\}$ est maximal
- Code non maximal $\Rightarrow$ moins d'information transmise, perte de débit pour les mêmes capacités de détection et correction

Borne de Gilbert-Varshamov

- Bornes de Singleton et Hamming : bornes supérieures du nombre de mots d'un code correcteur
- Borne de Gilbert-Varshamov est une borne inférieure du nombre de mots d'un code correcteur *maximal*
- Un code *maximal* K de distance minimale d a un rayon de recouvrement $c \leq d-1$, donc

$$r^n \leq |K| \cdot B_{d-1} \;\Leftrightarrow\; |K| \geq \frac{r^n}{B_{d-1}}$$

Exemple : code de Hamming binaire

- Famille de codes parfaits avec $n = 2^m - 1$, $k = 2^m - 1 - m$, $d = 3$
- Un mot $c = c_1 \ldots c_n \in \{0, 1\}^n$ du code de Hamming binaire est tel que les bits c_i dont l'indice i est une puissance de deux sont des bits de contrôle, les autres sont des bits de données
- Le bit de contrôle d'indice c_i pour $i = 2^\ell$ est la somme modulo 2 de tous les bits de données c_j dont l'indice j écrit en base 2 a le $(\ell + 1)^{\text{ème}}$ bit à 1

Exemple : code de Hamming binaire $(7, 4)$

- Pour $n = 7$, un mot $c = c_1 \ldots c_7$ du code de Hamming est tel que
  - les bits c_1, c_2, c_4 sont des bits de contrôle
  - les bits c_3, c_5, c_6, c_7 sont des bits de données
- On a

$$
\begin{aligned}
c_1 &= c_{110} + c_{101} + c_{111} = c_3 + c_5 + c_7 \mod 2\\
c_2 &= c_{110} + c_{011} + c_{111} = c_3 + c_6 + c_7 \mod 2\\
c_4 &= c_{101} + c_{011} + c_{111} = c_5 + c_6 + c_7 \mod 2
\end{aligned}
$$

Exemple : code de Hamming binaire

- On a $d = 3$: tout changement d'un bit de donnée $c_{\sum_{j=0}^{m-1} e_j 2^j}$ impacte tous les bits de contrôle c_{2^j} pour $e_j = 1$
- Le code est 1-correcteur. En effet, considérons la somme e des indices des bits de contrôle erronés. S'il n'y a qu'une seule erreur, elle ne peut provenir que du bit d'indice e

Questions ?

?

Crédits et remerciements

- Mes transparents suivent fortement les notes de cours développées par le Professeur Yves Roggeman pour le cours INFO-F303 à l'Université libre de Bruxelles
- Une partie des transparents et des exercices ont été repris ou adaptés des transparents développés par le Professeur Jean Cardinal pour ce même cours
- Je remercie chaleureusement Yves et Jean pour la mise à disposition de ce matériel pédagogique, et de manière plus large pour toute l'aide apportée pour la reprise de ce cours
- Les typos et erreurs sont exclusivement miennes (merci de les signaler !)