Réseaux, information et communications (INFO-F303) Partie Théorie de l'Information 2. Source aléatoire et codage efficace

Christophe Petit

Université libre de Bruxelles

Plan du cours

Source aléatoire et codes efficaces

- Hypothèses : alphabet fini, source “markovienne”
- But : longueur moyenne du code optimale

- Codes de Shannon et Shannon-Fano-Elias
- Code de Shannon-Fano
- Code de Huffman (optimal, utilisé pour ZIP, JPEG, MP3)

Source aléatoire, indépendante identiquement distribuée

- Chaque symbole s_i est associé à une probabilité

$$p_i = \mathbb{P}[s_{i_1}] > 0$$

  avec $\sum_{i=1}^{q} p_i = 1$

- Indépendance :

$$\mathbb{P}[s_{i_1} s_{i_2} \ldots s_{i_n}] = p_{i_1} \cdot p_{i_2} \cdot \cdots \cdot p_{i_n}$$

- Appelée *source markovienne* dans le syllabus

Longueur moyenne d'un code

- Longueur moyenne d'un code

$$L(K) = \mathbb{E}[\ell_i] = \sum_{i=1}^{q} \ell_i \cdot p_i$$

- Étant donnée une loi de probabilité sur S, quel est la plus petite longueur moyenne d'un code sans préfixe pour S ?

Codes efficaces

- Définition : un code est **efficace** ou **optimal** si sa longueur moyenne est minimale parmi celles de tous les codes possibles pour la source S de loi de probabilité $\mathbb{P}$
- Intuition : les symboles les plus probables doivent être associés aux mots les plus courts

Code de Shannon

- On suppose probabilités *décroissantes*

$$p_1 \geq p_2 \geq \cdots \geq p_q$$

- Soit $\ell_i = \lceil -\log_2 p_i \rceil \geq 1$. On a

$$\ell_1 \leq \ell_2 \leq \cdots \leq \ell_q$$

- **Code de Shannon** : $K(s_i)$ est formé des bits de

$$\left\lfloor 2^{\ell_i} \sum_{j=1}^{i-1} p_j \right\rfloor$$

Code de Shannon

- Exemple : $\mathbb{P} = \left\{ \frac{3}{8}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8} \right\}$
- On a $\ell_1 = \ell_2 = \ell_3 = 2$ et $\ell_4 = 3$
- Les probabilités cumulées sont (en base 2)
  $0, \frac{3}{8} = 0.011, \frac{5}{8} = 0.101, \frac{7}{8} = 0.111$
- On a $K(s_1) = 00$, $K(s_2) = 01$, $K(s_3) = 10$, $K(s_4) = 111$

- Remarque : clairement sous-optimal dans ce cas-ci

- Autre exemple : $\mathbb{P} = \left\{ \frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8} \right\}$

Propriétés du code de Shannon

- Simple, intuitif
- **Unique** (modulo permutation des codes de symboles de même probabilité)
- Code **sans préfixe** (voir transparent suivant)
- **Pas optimal** (donc peu utilisé)

Code de Shannon est sans préfixe

- Le préfixe strict de $K(s_i)$ de longueur ℓ_j, $j < i$, est

$$\left\lfloor 2^{\ell_j} \sum_{k=1}^{i-1} p_k \right\rfloor$$

- Montrons que sa différence avec $K(s_j)$ est non nulle

$$\begin{aligned} 2^{\ell_j} \left(\sum_{k=1}^{i-1} p_k - \sum_{k=1}^{j-1} p_k \right) &= 2^{\ell_j} \sum_{k=j}^{i-1} p_k \geq 2^{\ell_j} p_j \\ &\geq 2^{-\log_2 p_j} p_j = \frac{1}{p_j} p_j = 1 \end{aligned}$$

Variante : code de Shannon-Fano-Elias

- Shannon : $K(s_i) = \left\lfloor 2^{\ell_i} \sum_{j=1}^{i-1} p_j \right\rfloor$
- Shannon-Fano-Elias $K(s_i) = \left\lfloor 2^{\ell_i+1} (\sum_{j=1}^{i-1} p_j + \frac{1}{2} p_i) \right\rfloor$
- Egalement sans préfixe, unique
- Plus long que le code de Shannon
  (longueur moyenne $+1$)
- Base des codes arithmétiques
  (utilisés en compression des données)

Code de Shannon-Fano

- Résultat de l'algorithme suivant (pour $r = 2$), construisant l'arbre du code de façon "top-down"
  - Racine associée au mot vide et à toutes les probabilités décroissantes $p_1 \geq p_2 \geq \cdots \geq p_q$
  - Identifier i tel que $|\sum_{j=1}^{i} p_j - \sum_{j=i+1}^{q} p_j|$ minimal
  - $p_1, \ldots, p_i$ sont associées à un fils du sommet et une valeur de bit ajoutée en suffixe ; $p_{i+1}, \ldots, p_q$ sont associées à l'autre fils et l'autre valeur de bit
  - On procède par induction sur chaque branche
- Utilisation : algorithme *Implode* du format ZIP

Code de Shannon-Fano

- Exemple : $\mathbb{P} = \{3/8, 1/4, 1/4, 1/8\}$
  - $K = \{00, 01, 10, 11\}$, ou $K = \{0, 10, 110, 111\}$, ou ...
- Remarque : important d'ordonner les probabilités
  - Tentant de grouper $3/8$ et $1/8$ à la première étape pour équilibrer les deux branches de l'arbre
  - Mais dommage collatéral : mots plus longs pour les probabilités les plus élevées
    (Cas pathologique : $\mathbb{P} = \left\{3/8, 1/4, 1/4, 1/2^k, \ldots, 1/2^k\right\}$ pour k grand)

Propriétés du code de Shannon-Fano

- Code sans préfixe (par construction)
- Pas unique
  - Parfois, deux positions de scission possibles
  - Choix du bit de suffixe pour chaque branche
- Longueur moyenne unique
  (même si longueurs des mots peuvent varier)
- Longueur moyenne au plus la longueur moyenne du code de Shannon correspondant
  (même si certains mots peuvent être plus longs)
- Sous-optimal

Code de Huffman

- Code univoque (et sans préfixe) *optimal*
- Utilisé dans l'algorithme Deflate du format ZIP, formats JPEG et MP3
- Arbre du code peut être construit par un algorithme "bottom-up" de fusions d'arbres plus petits

Code de Huffman : algorithme ($r = 2$)

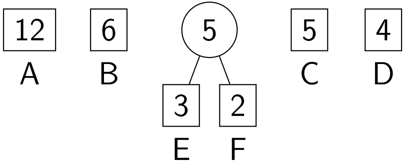
- Probabilités décroissantes $p_1 \geq p_2 \geq \cdots \geq p_q$
- Au départ, chaque symbole constitue son propre arbre au sein de la forêt, ayant pour poids la probabilité du symbole
- Fusionner deux arbres en un seuls
  1. Fusionner les 2 derniers arbres (poids les plus faibles)
  2. Poids de l'arbre fusion = somme de ceux des arbres fusionnés
  3. Valeurs 0 et 1 donnés en préfixe aux deux arbres fils (aux codes de leurs feuilles)
- Glisser cet arbre *à sa place* dans la forêt, afin d'y préserver l'ordre des arbres en poids non croissants
- Répéter jusqu'à obtenir un seul arbre

Code de Huffman : exemple

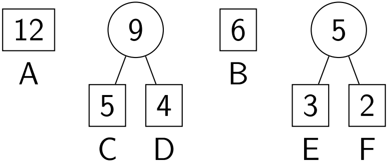
On considère $S = \{A, B, C, D, E, F\}$ avec les probabilités suivantes ($\times 32$) :

$\boxed{12}$	$\boxed{6}$	$\boxed{5}$	$\boxed{4}$	$\boxed{3}$	$\boxed{2}$
A	B	C	D	E	F

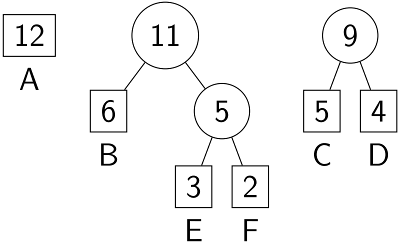
Code de Huffman : exemple



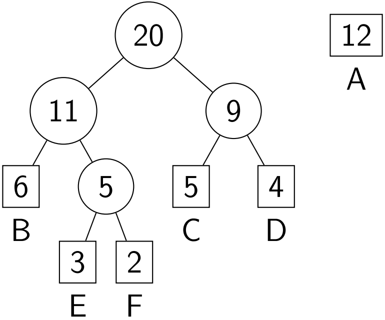
Code de Huffman : exemple



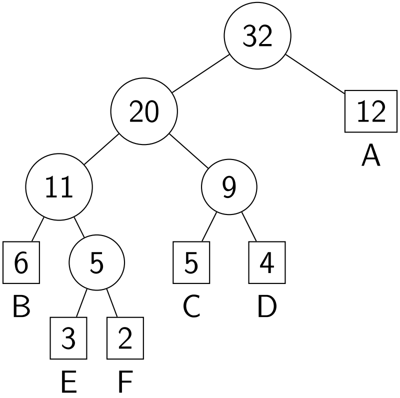
Code de Huffman : exemple



Code de Huffman : exemple



Code de Huffman : exemple



Les codes de Huffman sont efficaces

Pour une distribution de probabilité donnée sur les symboles, le code de Huffman a une longueur moyenne minimum.

Démonstration (1/2)

- Soit un code sans préfixe efficace
- $p_i > p_j \implies \ell_i \leq \ell_j$
- En échangeant éventuellement deux mots du code de longueur ℓ_{max}, on obtient un code sans préfixe efficace tel que dans l'arbre binaire associé, il existe un nœud de profondeur $\ell_{\text{max}} - 1$ ayant comme fils deux feuilles correspondant à deux symboles de probabilités les plus faibles

Démonstration (2/2)

- Notons p_{q-1} et p_q les deux probabilités les plus faibles
- La longueur moyenne du code fourni par le lemme précédent peut s'écrire

$$L = L' + 1 \times (p_{q-1} + p_q),$$

  où L' est la longueur moyenne du code sur l'ensemble des symboles $S' = \{s_1, \ldots, s_{q-1}\}$, avec la distribution $p'_k = p_k$ pour tout $k < q - 1$, et $p'_{q-1} = p_{q-1} + p_q$
- Si l'on veut minimiser L, il convient donc de minimiser L'
- On retrouve la procédure de construction du code de Huffman

Autres propriétés et comparaison

- Le code de Huffman n'est pas unique
  - Ordre ambigu si poids égaux
  - Nombre d'options maximal si les probabilités suivent une suite de Fibonacci
- Longueur moyenne optimale
  (mais peut avoir des mots plus longs ou plus courts que le code de Shannon correspondant)

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 !)