Codage des mots de poids constant

Vous joindrez le code source des programmes écrits pour répondre à une question.

Soient $n \geq t$ deux entiers positifs et $W_{n,t} \subset \{0,1\}^n$ l'ensemble des mots binaires de longueur n et de poids de Hamming[2] t. On considère la source $S_{n,t}$ constituée de l'alphabet $W_{n,t}$ muni de la loi uniforme. Nous noterons $\binom{n}{t}$ le coefficient binomial égal au nombre de parties de cardinal t d'un ensemble de cardinal n (égal à 0 si $n < t$).

Question 1. Quel est le cardinal de $W_{n,t}$? L'entropie de $S_{n,t}$?

Le cas typique qui nous intéresse est celui où t est petit par rapport à n (par exemple $n = 1024$ et $t = 50$). On représente alors un mot de $W_{n,t}$ par le t-uplet $0 \leq a_0 < a_1 < \ldots < a_{t-1} < n$ des indices de ses coordonnées non nulles.

Question 2. Montrez que l'application suivante

$$\begin{array}{ccc} W_{n,t} & \rightarrow & \left[0, \binom{n}{t}\right[\\ (a_0, \ldots, a_{t-1}) & \mapsto & \sum_{i=1}^{t} \binom{a_{i-1}}{i} \end{array}$$

où $0 \leq a_0 < a_1 < \ldots < a_{t-1} < n$, est une bijection.

Question 3. En déduire un code de longueur fixe de $S_{n,t}$ d'efficacité $E > 1 - 1/H(S_{n,t})$. On décrira les procédures de codage et de décodage correspondantes.

La méthode ci-dessus doit effectuer des calculs sur des entiers de grande taille qui ont une complexité algorithmique élevée. On souhaite réduire le coût du calcul, quitte à baisser l'efficacité. Plutôt que les indices des '1', nous allons considérer les écarts entre les indices :

$$(\underbrace{0,\ldots,0}_{\delta_0},\overset{a_0}{1},\underbrace{0,\ldots,0}_{\delta_1},\overset{a_1}{1},0\ldots\ldots 0,\overset{a_{t-2}}{1},\underbrace{0,\ldots,0}_{\delta_{t-1}},\overset{a_{t-1}}{1},0,\ldots,0).$$

Nous avons $\delta_0 = a_0$, et pour tout $i > 0$, $\delta_i = a_i - a_{i-1} - 1$.

Question 4. Pour tout i dans $[0, n-t]$, quelle est la probabilité p_i de l'évènement $\delta_0 = i$ (la loi de $W_{n,t}$ uniforme). La loi de δ_i, $i > 0$, est-elle différente ?

[1] ou par email avant le 5 novembre

[2] le poids de Hamming d'un mot est égal à son nombre de coordonnées non nulles

Question 5. Soit Δ la source d'alphabet $[0, n-t]$ munie de la loi de probabilité définie à la question précédente. Écrivez un programme prenant n et t en arguments et qui calcule l'entropie de cette source ainsi que la longueur moyenne d'un code de Huffman. Application numérique : $n = 1024$ et $t = 50$.

Question 6. On code un élément de $W_{n,t}$ comme un t-uplet d'éléments indépendants de Δ. Qualitativement, quelle erreur fait-on lorsque l'on code la source $S_{n,t}$ de cette manière ? Application numérique : $n = 1024$ et $t = 50$, comparez l'efficacité du codage de $S_{n,t}$ de la question 3 avec celui-ci.

Pour répondre à cette question, il faudra simuler le codage d'élément de $W_{n,t}$ choisis aléatoirement et uniformement. Tirer uniformément un mot de $W_{n,t}$ peut, par exemple, se faire de la façon suivante :

- *Pour $i = 0, \ldots, n-1$, poser $a_i = i$*
- *Pour $i = 0, \ldots, t-1$*
  - *tirer un entier j au hasard dans $[i, n-1[$*
  - *échanger a_i et a_j*
- *retourner $a_0, \ldots, a_{t-1}$ triés dans l'ordre croissant*

Le codage par la méthode de la question 6 est plus rapide mais nécessite le calcul d'un arbre de Huffman. Ce calcul n'est fait qu'une fois, mais nous voudrions éviter d'avoir à stoquer l'arbre. Nous allons donc dégrader encore le modèle et considérer pour tout $d > 0$ le codage suivant des entiers :

$$\begin{array}{rccl} \phi_d : & \mathbf{N} & \to & \{0,1\}^* \\ & i & \mapsto & \underbrace{1\cdots1}_{q} \parallel 0 \parallel \underbrace{\boxed{r}}_{d} \end{array}$$

où q et r sont le quotient et le reste de la division euclidienne de i par 2^d, $i = q2^d + r$, et $\underbrace{\boxed{r}}_{d}$ est l'écriture de r en base 2 (avec d bits exactement, on rajoute des '0' au début si besoin est).

Question 7. Montrez que ϕ_d est à décodage unique. On code l'élément $(\delta_0, \delta_1, \ldots, \delta_{t-1})$ de $W_{n,t}$ par $\phi_d(\delta_0) \parallel \phi_d(\delta_1) \parallel \cdots \parallel \phi_d(\delta_{t-1})$. Application numérique : $n = 1024$ et $t = 50$, déterminez et commentez l'efficacité du codage pour différentes valeurs de d.

Question 8. (facultative) Donnez un critère pour un choix optimal de d.