Compressão e Codificação de Dados

Secção de Telecomunicacções (DEEC), **Instituto Superior Técnico**

27 de Janeiro de 1999

Problema 1

Considere uma fonte que emite símbolos de um alfabeto $\mathcal{X} = \{A, B, C, D, E, F\}$ com probabilidades $p(A) = 0.3$, $p(B) = 0.2$, $p(C) = 0.15$, $p(D) = 0.15$, $p(E) = 0.1$, $p(F) = 0.1$. A entropia é igual a 2.471 bits/símbolo.

a) Desenhe um código de Huffman binário para esta fonte e calcule o seu comprimento médio. Compare com a entropia e comente, referindo como poderia melhorar a eficiência deste tipo de codificação.

b) Repita as duas alíneas anteriores, agora com $p(A) = 1/2$, $p(B) = 1/4$, $p(C) = 1/8$, $p(D) = 1/16$, $p(E) = 1/32$, $p(F) = 1/32$.

c) Considere finalmente um alfabeto infinito $\mathcal{X} = \{x_1, x_2, ..., x_n, ...\}$, cujas probabilidades são $p(x_i) = 2^{-i}$. Recorrendo, se necessário às conhecidas relações

$$\sum_{i=1}^{\infty} a^i = \frac{a}{1-a} \quad \text{e} \quad \sum_{i=1}^{\infty} i\, a^i = \frac{a}{(1-a)^2},$$

calcule a entropia desta fonte. Considere o código para esta fonte definido por

$$c(x_n) = \underbrace{11...1}_{(n-1)\ 1\text{'s}}\ 0.$$

Verifique que este código é instantâneo e óptimo e justifique intuitivamente este facto; sugestão: tente desenhar a árvore binária a que corresponde este código e relacione com o procedimento de Huffman.

Problema 2

Uma dada fonte de Markov é caracterizada pela distribuição de probabilidade do símbolo X_n, condicionada ao símbolo X_{n-1}, designada por $p(X_n|X_{n-1})$ e apresentada na seguinte tabela:

$p(X_n\|X_{n-1})$	$X_n = A$	$X_n = B$	$X_n = C$
$X_{n-1} = A$	1/4	1/2	1/4
$X_{n-1} = B$	1/2	1/2	0
$X_{n-1} = C$	0	1/2	1/2

a) A probabilidade, não condicionada, dos símbolos evolui de acordo com

$$[p(X_n = A),\ p(X_n = B),\ p(X_n = C)] = [p(X_{n-1} = A),\ p(X_{n-1} = B),\ p(X_{n-1} = C)]\ \mathbf{M},$$

em que $\mathbf{M}$ é a matriz descrita pela tabela acima. Verifique que $[1/3,\ 1/2,\ 1/6]$ é distribuição estacionária desta fonte de Markov.

b) Calcule a taxa de entropia condicional. Compare com a entropia da distribuição estacionária e comente.

c) Suponha agora que se pretente desenhar um sistema de codificação para esta fonte e que se opta por desenhar 3 códigos diferentes, um para cada possível símbolo anterior. Desenhe estes códigos por simples inspeção da tabela (isto é, não é necessário recorrer à técnica de Huffman) e calcule o comprimento médio de cada um deles, condicionado ao respectivo símbolo anterior. Com base no vector de distribuição estacionário acima indicado, calcule o comprimento médio geral deste procedimento de codificação; compare com a taxa de entropia e com a entropia da distribuição estaionária e comente.

d) Assumindo agora que a matriz de transição de probabilidades passa a ser a que apresenta na tabela seguinte, verifique que $[4/9,\ 2/9,\ 1/3]$ é distribuição estacionária desta nova fonte. Repita as alíneas **b)** e **c)** para estes novos valores.

$p(X_n \mid X_{n-1})$	$X_n = A$	$X_n = B$	$X_n = C$
$X_{n-1} = A$	1/4	1/2	1/4
$X_{n-1} = B$	0	0	1
$X_{n-1} = C$	1	0	0

Problema 3

Considere o canal binário simétrico (CBS) cujas probabilidades de transição são $p(Y = 1 \mid X = 0) = \alpha$, $p(Y = 0 \mid X = 1) = \alpha$. Recorde que $C_{\mathrm{CBS}} = 1 + \alpha \log \alpha + (1 - \alpha) \log (1 - \alpha)$.

a) Considere que na saída deste CBS se liga um outro CBS com probabilidade de erro β. Naturalmente, a concatenação dos dois CBS's constitui um novo canal. Sem recorrer à multiplicação das matrizes de canal, escreva a matriz do canal resultante e obtenha a sua capacidade.

b) Sem efectuar qualquer cálculo, diga, justificando, se capacidade do canal resultante da concatenação é superior, inferior, ou igual à do canal inicial.

Problema 4

Considere o seguinte procedimento de codificação do tipo Lempel-Ziv: cada subsequência de símbolos de comprimento superior ou igual a L que já tenha ocorrido anteriormente é substituida por um par de "ponteiros" com o formato (p_1, p_2), indicando que a ocorrência anterior se verificou na posição p_1, a contar do início, e que o comprimento da subsequência é p_2.

a) Com $L = 3$, ilustre detalhadamente o funcionamento desta técnica ao codificar e, em seguida, descodificar a sequência "*abccbabcabccbabca*".

b) Suponha que se utilizam 3 *bytes* para escrever o ponteiro p_1. Quanto a p_2, utiliza-se o seguinte esquema de codificação:

$$\underbrace{00...0}_{p_2 \text{ 0's}}\underbrace{11111}_{\text{5 1's}}$$

Sabendo que os símbolos das sequências a codificar correspondem a um *byte*, qual o valor mínimo de L para que este procedimento nunca resulte em expansão da sequência original.

Problema 5

Considere um quantizador escalar $\mathcal{Q}(\cdot)$ cuja gama de valores de entrada é $[-10, 10[$ e cujas células são os intervalos da forma $R_k = [-10 + k, -10 + k + 1[$, para $k = 0, 1, ..., 19$. Assuma que a entrada é uma variável aleatória X cuja função densidade de $f_X(x)$ probabilidade é a seguinte:

a) Determine o *code book* óptimo $\{y_0, y_1, ..., y_{19}\}$, isto é, aquele que minimiza o erro quadrático médio

$$E\left[(X - \mathcal{Q}(X))^2\right] = \int_{-10}^{10} f_X(x)\,(x - \mathcal{Q}(x))^2\, dx.$$

b) Tomando a resposta que obteve na alínea **(a)**, qual a distorção quadrática média resultante?

c) Supondo agora que $B = 0$, desenhe um quantizador óptimo com o mesmo número de células (20). Neste caso, tem liberdade total para escolher as células e o *code book*. Qual a distorção quadrática média resultante? Compare com o resultado da alínea **b)**.