# Exame de Compressão e Codificação de Dados

*Mário Figueiredo*
Secção de Telecomunicacções - DEEC
**Instituto Superior Técnico**

24 de Fevereiro de 1999

1. Considere uma fonte que emite símbolos de um alfabeto $\mathcal{X} = \{1, 2, 3, 4\}$ com probabilidades $p(1) = 0.3$, $p(2) = 0.4$, $p(3) = 0.3$, $p(4) = 0$. A entropia desta fonte é aproximadamente igual a 1.571 bits/símbolo.

   **a)** Desenhe um código de Huffman binário para esta fonte e calcule o seu comprimento médio. Compare com a entropia e comente.

   **b)** Considere a extensão de segunda ordem da fonte. Desenhe um código de Huffman binário para esta extensão da fonte, calcule o seu comprimento médio, compare com a entropia e comente.

   **c)** Qual a ordem de extensão necessária para garantir que o comprimento médio obtido não se afasta mais do que 0.01 bits/símbolo da entropia da fonte. Justifique.

2. Uma fonte de Markov é caracterizada pela distribuição de probabilidade do símbolo $X_n$, condicionada ao símbolo $X_{n-1}$, designada por $p(X_n|X_{n-1})$. Quando a fonte é estacionária, esta distribuição não depende de $n$, pelo que temos $p(X_n|X_{n-1}) = p(X_m|X_{m-1})$, para quaisquer dois instantes $n$ e $m$. Considere então uma fonte de Markov, gerando símbolos do alfabeto $\{A, B, C, D\}$, cuja distribuição de probabilidade condicionada é dada pela tabela seguinte:

| $p(X_n\|X_{n-1})$ | $X_n = A$ | $X_n = B$ | $X_n = C$ | $X_n = D$ |
|---|---|---|---|---|
| $X_{n-1} = A$ | 1/2 | 1/4 | 1/4 | 0 |
| $X_{n-1} = B$ | 1/4 | 1/4 | 1/4 | 1/4 |
| $X_{n-1} = C$ | 1/2 | 1/4 | 1/8 | 1/8 |
| $X_{n-1} = D$ | 1/8 | 1/8 | 1/4 | 1/2 |

   Sabe-se que [0.3791, 0.2288, 0.2222, 0.1699] é a distribuição estacionária desta fonte de Markov.

   **a)** Calcule a taxa de entropia condicional deste processo. Compare com a entropia da distribuição estacionária e comente.

   **b)** Suponha agora que se pretente desenhar um sistema de codificação para esta fonte e que se opta por desenhar 4 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.

   **c)** Considerando agora uma fonte cuja matriz de transição é que se apresenta de seguida, repita a alínea anterior após determinar a sua distribuição estacionária.

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

3. Considere o canal de apagamento simétrico (CAS) com erros, ilustrado abaixo, cujo alfabeto de entrada é $\mathcal{X} = \{0, 1\}$ e o de saída é $\mathcal{Y} = \{0, 1, \star\}$. As probabilidades de transição são $p(Y = \star|X = 0) = \varepsilon$, $p(Y = \star|X = 1) = \varepsilon$, $p(Y = 1|X = 0) = p$, $p(Y = 0|X = 0) = p$.

   **a)** Determine a matriz deste canal.

   **b)** Considere que na saída do canal acima descrito se liga um outro canal, cujo alfabeto de entrada é $\{0, 1, \star\}$ e o de saída é $\{0, 1\}$, tal como indicado na figura. Este segundo canal, quando recebe na sua entrada o símbolo $\star$, coloca na saída um 0 ou um 1, aleatoriamente com equiprobabilidade.

   Determine a matriz do canal resultante e calcule a sua capacidade.

   **c)** Determine a matriz e calcule a capacidade do chamado canal "Z", cujo diagrama é:

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 4 *bytes* para escrever o ponteiro $p_1$. Quanto a $p_2$, utiliza-se o seguinte esquema de codificação:

$$\underbrace{00...0}_{p_2+1 \text{ 0's}}\ 1001$$

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

5. Considere um quantizador escalar $\mathcal{Q}(\cdot)$ cuja gama de valores de entrada é $[0, 10[$ e cujas células são os intervalos da forma $R_k = [k/2, (k+1)/2[$, para $k = 0, 1, ..., 19$. Assuma que a entrada é uma variável aleatória $X$ cuja função densidade de probabilidade $f_X(x)$ é uniforme em $[0, 10][$.

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

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

   Qual a distorção quadrática média resultante?

   **b)** Considerando agora que a função densidade de probabilidade $f_X(x)$ é dada por $f_X(x) = x/50$, para $x \in [0, 10[$. O quantizador obtido na alínea **a)** é ainda óptimo? Se não for, calcule a expressão geral dos elementos do codebook óptimo $\{y_i, \quad i = 0, .., 19\}$ para as mesmas células. Calcule o valor de $y_0$ e interprete o resultado.