Compressão e Codificação de Dados. First Test

Mestrado em Engenharia Electrotécnica e de Computadores, IST November 8, 2014

Name: _____ Number: _____

NOTES: 1. Durations of the test is 90 minutes.

2. Part I: correct answer = 1 point; wrong answer = – 0.5 points.

3. Potentially useful facts: $\log_2 3 \simeq 1.585$; $\log_2 5 \simeq 2.322$; $\log_{10}(2) \simeq 0.30$; $\log_a b = (\log_c b)/(\log_c a)$.

Part I

1. Let $V \in \{2, 3, ..., 12\}$ be the random variable representing the total outcome when tossing two independent fair dice. Then,

   a) $H(V) < \log_2 11$ bits/symbol; ■

   b) $H(V) = \log_2 11$ bits/symbol; □

   c) $H(V) > \log_2 11$ bits/symbol. □

   Explanation: The random variable takes 11 possible values, and the distribution in not uniform.

2. Let $X, Y \in \{1, 2, ..., 6\}$ be two random variables representing the outcome of the toss of two independent fair dice. Consider $Z = |X - Y|$; then,

   a) $H(Z) < 1 + \log_2 3$ bits/symbol; ■

   b) $H(Z) = 1 + \log_2 3$ bits/symbol; □

   c) $H(Z) > 1 + \log_2 3$ bits/symbol. □

   Explanation: The random variable Z takes values in the set {0, 1, 2, 3, 4, 5}, with non-uniform probability; then, $H(Z) < \log_2 6 = \log_2 2 + \log_2 3 = 1 + \log_2 3$ bits/symbol.

3. Let $A, B \in \{0, 1, 2, ..., 9\}$ be a pair of independent random variables, both with uniform distribution, and let $X = A + 10\,B$.

   a) $I(X; A) < 3$ bits/symbol; □

   b) $I(X; A) = 3$ bits/symbol; □

   c) $I(X; A) > 3$ bits/symbol. ■

   Explanation: Clearly, if you know X, you know A, which means that $H(A|X) = 0$; thus $I(X; A) = H(A) - H(A|X) = H(A) = \log_2 10 > 3$ bits/symbol (because A has uniform distribution).

4. Let $A, B \in \{0, 1, 2\}$ be two independent random variables, both with uniform distribution, and $X = A + B$. Then,

   a) $H(X|A) < H(A|X)$; □

   b) $H(X|A) = H(A|X)$; □

   c) $H(X|A) > H(A|X)$. ■

   Explanation: Using Bayes law for entropies, we have that $H(X|A) = H(X, A) - H(A)$ and also that $H(A|X) = H(X, A) - H(X)$. Thus $H(X|A) - H(A|X) = H(X) - H(A) > 0$, because the sum of two independent random variables has larger entropy than each of the variables individually.

5. Consider a code where all the words have the same length, for a source where all the symbols have the same probability.

   a) This code is necessarily optimal; ☐

   b) This code is necessarily non-optimal; ☐

   c) This code may, or not, be optimal. ■

   Explanation: If the number of symbols is a power of D and the distribution is uniform (thus all probabilities are negative powers of D), then the optimal $D-$ary code has all the words with the same length. If the number of symbols is a not a power of D, and the distribution is uniform, then the optimal code has words of different lengths. Examples: an optimal binary code for $X \in \{1,2,3,4\}$, with uniform distribution, is $\{C(1) = 00, C(2) = 01, C(3) = 10, C(4) = 11\}$; an optimal binary code for $Y \in \{1,2,3\}$, with uniform distribution, is $\{C(1) = 0, C(2) = 10, C(3) = 11\}$;

6. For a source generating symbols from the alphabet $\{a,b,c\}$ with probabilities $P(a) > P(b) > P(c)$, the number of Huffman codes is

   a) 4; ■

   b) 6; ☐

   c) 8. ☐

   Explanation: There is only one possible structure of the Huffman tree, thus the number of Huffman codes is $2^{\text{number of nodes}} = 2^2 = 4$. The four Huffman codes are $\{C(1) = 0, C(2) = 10, C(3) = 11\}$, $\{C(1) = 0, C(2) = 11, C(3) = 10\}$, $\{C(1) = 1, C(2) = 00, C(3) = 01\}$, and $\{C(1) = 1, C(2) = 01, C(3) = 00\}$.

7. Consider a memoryless source $X \in \{a,b,c,d\}$, with probabilities $\{P(a) = 15/16, P(b) = P(c) = P(d) = 1/48\}$, generating 1000 symbols/second. Knowing that the entropy of this source is $H(X) \simeq 0.4364$ bits/symbol, what is the lowest extension order necessary to allow communicating the information produced by this source through a binary channel with maximum rate equal to 1000 bits/second?

   a) no extension is needed; ☐

   b) second order; ■

   c) third order. ☐

   Explanation: An optimal code for the source (without extension) is $\{C(a) = 0, C(b) = 10, C(3) = 110, C(4) = 111\}$, which has expected length $L(C) = 15/16 + (2+3+3)/48 = 15/16 + 8/48 = 53/48$ bits/symbol. Thus, without extension, this code produces $(53/48)1000 > 1000$ bits/second. For the n-th order extension, even without building the codes, we know that the expected length (in bits/symbol) is less than $H(X) + 1/n$. Since, $H(X) \simeq 0.4364$, it is clear that the second order is enough, because $H(X) + 1/2 \simeq 0.9364 < 1$ bits/symbol, thus the code would produce less than 1000 bits/second.

8. Consider a source $X \in \{a,b,c,d,e\}$, with probabilities $\{P(a) = 1/3, P(b) = 1/3, P(c) = 1/4, P(d) = 1/24, P(e) = 1/24\}$. The expected length of the optimal **quaternary** code for this source is

   a) 13/12 quads/symbol; ■

   b) 14/12 quads/symbol; ☐

   c) 15/12 quads/symbol ☐

   Explanation: Simply by inspection (although it can also be obtained by the Huffman procedure, after including three zero-probability symbols), an optimal quaternary code for source X is $\{C(a) = 0, C(b) = 1, C(c) = 2, C(d) = 30, C(e) = 31\}$. The resulting expected length is $1/3+1/3+1/4+2/24+2/24 = 13/12$ quads/symbol.

9. Consider a quaternary first-order Markov source, with the following transition probabilities

$P(X_t \mid X_{t-1})$	$X_t = a$	$X_t = b$	$X_t = c$	$X_t = d$
$X_{t-1} = a$	0	1	0	0
$X_{t-1} = b$	0	0	1	0
$X_{t-1} = c$	0	0	0	1
$X_{t-1} = d$	1/4	1/4	1/4	1/4

Knowing that the stationary distribution of this source is (0.1, 0.2, 0.3, 0.4) The expected length of the optimal binary coding scheme for this source is

a) 0.8 bits/symbol; ■

b) 1.0 bits/symbol; □

c) 1.2 bits/symbol □

Explanation: Since the transitions from $X_{t-1} = a$, $X_{t-1} = b$, and $X_{t-1} = c$ are deterministic, the corresponding codes have length zero, and only the code corresponding to the fourth row of the transition matrix needs to be considered. The optimal code for the distribution in the fourth row of the matrix obviously has expected length equal to 2 bits/symbol. Since the state d has stationary probability of occurrence equal to 0.4, the global expected length is $2 \times 0.4 = 0.8$ bits/symbol.

10. Consider a quaternary first-order Markov source, with the following transition probabilities:

$P(X_t \mid X_{t-1})$	$X_t = a$	$X_t = b$	$X_t = c$	$X_t = d$
$X_{t-1} = a$	1/2	1/4	1/8	1/8
$X_{t-1} = b$	1/4	1/2	1/8	1/8
$X_{t-1} = c$	1/8	1/8	1/2	1/4
$X_{t-1} = d$	1/4	1/8	1/8	1/2

The expected length of the optimal binary coding scheme for this source is

a) 5/4 bits/symbol; □

b) 6/4 bits/symbol; □

c) 7/4 bits/symbol ■

Explanation: The optimal codes for the distributions in all the rows of the transition matrix have expected length equal to 7/4 bits/symbol. Thus, regardless of the stationary distribution, the global expected length is also 7/4 bits/symbol.

Part II

Problem 1

Let $X \in \{1, 2\}$ be a random variable associated with the output of a discrete memoryless source (DMS) with probabilities $p_1 = 0.2$ and $p_2 = 0.8$.

1. Determine the entropies $H(X)$ and $H(X_{(2)})$ in bits/symbol and in trits/symbol.
   Note 1: $X_{(2)}$ denotes the random variable associated with the second order extension of the original DMS.
   Note 2: $0.2 \log_2 0.2 \simeq -0.4644$; $0.8 \log_2 0.8 \simeq -0.2575$; $0.2 \log_3 0.2 \simeq -0.2930$; $0.8 \log_3 0.8 \simeq -0.1625$;

   Solution:

   Binary entropies:

   - $H_2(X) = -0.2 \log_2 0.2 - 0.8 \log_2 0.8 \simeq 0.4644 + 0.2575 = 0.7219$ bits/symbol.
   - $H_2(X_{(2)}) = 2\, H_2(X)$ (the source is memoryless), thus $H_2(X_{(2)}) \simeq 1.4438$ bits/(2 symbols).

   Ternary entropies:

   - $H_3(X) \simeq 0.2930 + 0.1625 = 0.4555$ trits/symbol.
   - $H_3(X_{(2)}) = 2\, H_3(X)$ (the source is memoryless), thus $H_3(X_{(2)}) \simeq 0.9110$ trits/(2 symbols).

2. Determine the probabilities for $X_{(2)}$, *i.e.*, $p(x) \equiv P\{X_{(2)} = x\}$, for $x \in \{1,2\}^2 = \{(1,1),(1,2),(2,1),(2,2)\}$. Obtain a binary Huffman code C_a and ternary Huffman code C_b. Compute the average codeword lengths $L(C_a)$ and $L(C_b)$ and the respective efficiencies $\eta_a \equiv H(X_{(2)})/L(C_a)$ and $\eta_b \equiv H(X_{(2)})/L(C_b)$.

   Solution:

   - Probabilities for $X_{(2)}$: $\{p(1,1) = 0.04,\, p(1,2) = 0.16,\, p(2,1) = 0.16,\, p(2,2) = 0.64\}$ (since the source is memoryless).
   - Binary Huffman code (simply by inspection): $\{C_a(1,1) = 111,\, C_a(1,2) = 110,\, C_a(2,1) = 10,\, C_a(2,2) = 0\}$. Expected code length $L(C_a) = 3 \times 0.04 + 3 \times 0.16 + 2 \times 0.16 + 1 \times 0.64 = 1.56$ bits/(2 symbols). Efficiency: $\eta_a \simeq 1.4438/1.56 \simeq 0.93$ (or 93%).
   - Ternary Huffman code (simply by inspection): $\{C_b(1,1) = 21,\, C_a(1,2) = 20,\, C_b(2,1) = 1,\, C_b(2,2) = 0\}$. Expected code length $L(C_a) = 2 \times 0.04 + 2 \times 0.16 + 1 \times 0.16 + 1 \times 0.64 = 1.2$ trits/(2 symbols). Efficiency: $\eta_b \simeq 0.9110/1.20 \simeq 0.759$ (or 75.9%).

3. Modify the probabilities p_1 and p_2 such that the efficiency of code C_a obtained in point 2 is 100%. Under these conditions, the efficiency of code C_b obtained in point 2, is less than 100%. Justify this statement.

   Solution: For a binary code to have 100% efficiency, the probabilities have to be powers of 2; in this case, since there are only two symbols, the only possibility is $p_1 = p_2 = 1/2$.

   A ternary code has 100% efficiency if and only if the probabilities are powers of 3, and 1/2 is not a power of 3.

4. Let $Y \in \{1,2,3,4,5\}$ be a random variable associated with the output of a DMS with entropy $H(Y) = \log_2 5$ bits/symbol. Consider the binary codes listed in the table below. Compute the average codeword lengths $L(C_1)$ and $L(C_2)$ and the respective efficiencies. Is any of these two codes optimal?

Y	Code C_1	Code C_2
1	11	0
2	00	10
3	10	110
4	100	1110
5	101	1111

   Solution: Since $H(Y) = \log_2 5$ bits/symbol, the distribution is uniform: $\{P(Y = 1) = P(Y = 2) = P(Y = 3) = P(Y = 4) = P(Y = 5) = 1/5\}$. Thus, $L(C_1) = (2 + 2 + 2 + 3 + 3)/5 = 12/5$ bits/symbol, while $L(C_2) = (1 + 2 + 3 + 4 + 4)/5 = 14/5$ bits/symbol. The efficiencies are $\eta_1 = (5 \log_2 5)/12$ and $\eta_2 = (5 \log_2 5)/14$. Code C_1 is not optimal because it is not even instantaneous; $C(3)$ is a prefix of $C(4)$ and $C(5)$. To check if C_2 is optimal, we need to compare its expected length with that of the Huffman code; a possible Huffman code is $\{C(1) = 00\, C(2) = 01\, C(3) = 10\, C(4) = 110\, C(5) = 111\}$, which has expected length equal to 12/5 bits/symbol, thus C_2 is not optimal.

5. With respect to code C_2 shown in the above table, is it optimal for some probability distribution p_1, p_2, p_3, p_4, p_5 such that $p_2 = p_3 = p_4 = p_5$?

   Solution: Since $p_2 = p_3 = p_4 = p_5$, we may suspect that a code where the words for symbols 2, 3, 4, and 5 have the same length could be better than C_2. To confirm that, consider the following code:

Y	Code C_3
1	0
2	100
3	101
4	110
5	111

   The expected length of C_3 is $L(C_3) = p_1 + (3+3+3+3)\, p_2 = p_1 + 12\, p_2$ bits/symbol (recall that $p_2 = p_3 = p_4 = p_5$). The expected length of C_2 is $L(C_2) = p_1 + (2 + 3 + 4 + 4)\, p_2 = p_1 + 13\, p_2$ bits/symbol, which is larger than $L(C_3)$, thus C_2 cannot be optimal.

Problem 2

Consider a stationary first order Markovian source $\{X_t\}$ generating symbols from the alphabet $\mathcal{X} = \{a, b, c\}$. The Table below shows the joint distribution $P(X_t = x_t, X_{t-1} = x_{t-1})$ for $x_t, x_{t-1} \in \mathcal{X}$.

		X_t		
		a	b	c
X_{t-1}	a	2/18	1/18	1/18
	b	2/18	4/18	2/18
	c	0	3/18	3/18

Note: The above table represents a joint probability matrix and not a transition matrix

1. Determine the marginal probabilities $P(X_{t-1} = x_{t-1})$ and $P(X_t = x_t)$ for $x_t, x_{t-1} \in \mathcal{X}$.

   Solution: The marginals are easily obtained by summing the columns and the rows of the table: $P(X_t = a) = 2/9$, $P(X_t = b) = 4/9$, $P(X_t = c) = 3/9$ and $P(X_{t-1} = a) = 2/9$, $P(X_{t-1} = b) = 4/9$, $P(X_{t-1} = c) = 3/9$. as expected, since it is said that the source is stationary, the two distributions are equal.

2. Which is the stationary distribution $\mathbf{p}(\infty) = [p_1(\infty), p_2(\infty), p_3(\infty)]^T$ of this source?

   Solution: Since, as seen above, the distributions at two consecutive times instants is the same, this is precisely the stationary distribution (by definition), thus $p_1(\infty) = 2/9$, $p_2(\infty) = 4/9$, and $p_3(\infty) = 3/9$.

3. Determine the the transition matrix $P \equiv [P(X_t = x_t | X_{t-1} = x_{t-1})]$ and the conditional entropy rate.

   Solution: The transition matrix is obtained by Bayes law, $P(X_t = x_t | X_{t-1} = x_{t-1}) = P(X_t = x_t, X_{t-1} = x_{t-1})/P(X_{t-1} = x_{t-1})$, which corresponds to normalizing each row of the matrix above:

		X_t		
		a	b	c
X_{t-1}	a	1/2	1/4	1/4
	b	1/4	1/2	1/4
	c	0	1/2	1/2

4. Determine an optimal binary code for this source and compute the average codeword length. Compare with the conditional entropy rate and comment.

   Solution: The codes for this source are

		X_t		
		a	b	c
X_{t-1}	a	0	10	11
	b	10	0	11
	c	-	0	1

   The expected code length is equal to (3/2)(2/9) + (3/2)(4/9) + (3/9) = 12/9 bits/symbol. Of course, this coincides with the conditional entropy rate, since all the conditional probabilities are powers of 2.

5. Obtain an optimal binary code for the stationary distribution of this source and compare its length with the one obtained in the previous point.

   Solution: An optimal code for the stationary distribution is $\{C_(a) = 00, C(b) = 1, C(c) = 01\}$, which has expected length equal to 14/9 bits/symbols. Naturally, this code has larger expected length than the one in the previous question, since it ignores the dependency of each symbol on the previous one.