Data Coding and Compression.

First Exam and Second Test

DEEC, IST January 9, 2015

Name: _____ Number: _____

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

Parte I

(Exam: correct answer = 0.5, wrong answer = – 0.25. Test: correct answer = 1, wrong answer = – 0.5.)

1. The random variables $X, Y, Z \in \{1, 2\}$ represent the outcome of three throws of independent fair coins and $T = X + Y + Z$ denotes the sum of the three outcomes (note that $T \in \{3, \ldots, 6\}$); therefore,

   a) $H(T) < 2$ bits/symbol; ■

   b) $H(T) = 2$ bits/symbol; □

   c) $H(T) > 2$ bits/symbol. □

2. Let T be the random variable defined in the previous point and $U = |T - 3|$ another random variable; therefore,

   a) $H(U) < H(T)$; □

   b) $H(U) = H(T)$; ■

   c) $H(U) > H(T)$. □

3. Consider the random variables $X \in \{0, 1, ..., 49\}$, $A(X)$, and $B(X)$, where $A(X)$ and $B(X)$ represent, respectively, the digit of tens of X and the digit of units of X (for example for $X = 23$, we have $A(X) = 2$ and $B(X) = 3$). therefore,

   a) $H(X) < H(A, B)$. □

   b) $H(X) = H(A, B)$. ■

   c) $H(X) > H(A, B)$. □

4. Consider the random variables A e B of point 3 and assume that they have a uniform distribution; therefore,

   a) $H(X|A) < \log_2 10$ bits/symbol; □

   b) $H(X|A) = \log_2 10$ bits/symbol; ■

   c) $H(X|A) > \log_2 10$ bits/symbol. □

5. Consider the random variables A e B of the previous point, which have a uniform distribution; therefore,

   a) $I(A; B) < \log_2 5$ bits/symbol; ■

   b) $I(A; B) = \log_2 5$ bits/symbol; □

   c) $I(A; B) > \log_2 5$ bits/symbol. □

6. The random variable X takes values on an alphabet with N symbols uniformly distributed ($H(X) = \log_2 N$ bits/symbol). In an optimal code for this source it is always true that

   a) all codewords have the same length; □

   b) there are codewords with different lengths; □

   c) the average codeword length is never smaller than $H(X)$. ■

7. Consider the source $X \in \{a, b, c, d\}$ with probabilities $P(a) = 1/2$, $P(b) = 1/4$, $P(c) = 1/8$ e $P(d) = 1/8$. The expected length of an optimal code for this source is

   a) 1 trit/symbol; ☐

   b) 5/4 trit/symbol; ■

   c) 6/4 trit/symbol ☐

8. Consider a source generating symbols of the alphabet $\{a, b, c\}$ with probabilities $\{1/2 - \varepsilon,\ 1/4 + \varepsilon,\ 1/4\}$, where $\varepsilon \in [0,\ 1/4]$. The code $\{C(a) = 0,\ C(b) = 11,\ C(c) = 10\}$,

   a) is optimal for any value of $\varepsilon \in [0,\ 1/4]$. ☐

   b) is not optimal for any value of $\varepsilon \in [0,\ 1/4]$. ☐

   c) is only optimal for some values of $\varepsilon \in [0,\ 1/4]$. ■

9. Consider a first order Markovian source with the following transition matrix:

$P(X_n \mid X_{n-1})$	$X_t = a$	$X_t = b$	$X_t = c$	$X_t = d$
$X_{t-1} = a$	1/2	0	1/4	1/4
$X_{t-1} = b$	1/4	1/2	0	1/4
$X_{t-1} = c$	1/4	1/4	1/2	0
$X_{t-1} = d$	0	1/4	1/4	1/2

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

   a) $> 4/3$ bit/symbol; ■

   b) $= 4/3$ bit/symbol; ☐

   c) $< 4/3$ bit/symbol. ☐

10. Consider that the sequence *dcbaaa* was generated by the source described in question 9. Assuming that the initial distribution is (1/2, 1/4, 1/8, 1/8), what is the length of the resulting arithmetic codeword?

    a) 10 bits; ☐

    b) 11 bits; ☐

    c) 12 bits. ■

11. The Elias Delta code word for the natural number 19 is

    a) $C_\delta(19) = 000010011$; ☐

    b) $C_\delta(19) = 0010110011$; ☐

    c) $C_\delta(19) = 001010011$. ■

12. Which of the following sequences results from the Lempel-Ziv-Welch (LZW) decoding of the sequence *1,2,3,4,5,6,7*, assuming that the alphabet is $\{a, b, c, d\}$ and the first index of the dictionary is 1.

    a) *abccabbccd*; ☐

    b) *abcdabbccd*; ■

    c) *abcdabbcbd*. ☐

13. Consider an LZW coder for an alphabet with 32 symbols (thus represented by 5-bit symbols) using a dictionary of size 256 (thus indexed by 8-bit words). The minimum length of a sequence of consecutive *a*'s ("aaa...a") such that its LZW compression has fewer bits than the sequence itself is

    a) 5; ■

    b) 6; ☐

    c) 7. ☐

14. Consider a random variable $X \in [0, 1]$, with the probability density function $f_X(x) = 1 - \delta + 2\delta x$, where δ is a parameter in $[-1, +1]$. Then, for any $\delta \in [-1, +1]$,

a) $h(X) < 0$; ☐

b) $h(X) > 0$; ☐

c) none of the previous answers. ■

15. Consider the random variable $X \in [0, 1]$ defined in the previous question. Then,

a) $h(X)$ is a monotically increasing function of $\delta \in [-1, +1]$; ☐

b) $h(X)$ is a monotically decreasing function of $\delta \in [-1, +1]$; ☐

c) none of the previous answers. ■

16. Consider the random variable $X \in [0, 1]$ defined in the previous question, with $\delta = 1$, connected to a uniform quantizer with 8 regions. The optimal representative of each region

a) is located to left of its center; ☐

b) is located in the center of the region; ☐

c) is located to the right of its center. ■

17. Consider the random variable $X \in [0, 1]$ defined in the question 14, now with $\delta = 0$, connected to a uniform quantizer with 8 regions. The entropy of the discrete random variable at the output of the quantizer is

a) less than 3 bits/symbol; ☐

b) equal to 3 bits/symbol; ■

c) larger than 3 bits/symbol. ☐

18. Consider two uniform random variables $X \in [-1, 1]$ and $Y \in [-2, 2]$, both connected to similar 2-bit uniform quantizers with the following regions $R_0 = [-2, -0.5[$, $R_1 = [-0.5, 0[$, $R_2 = [0, 0.5[$, and $R_3 = [0.5, 2[$.

a) the quantizers are optimal for both X and Y; ☐

b) the quantizers are not optimal for X or Y; ☐

c) the quantizers are optimal for X but not for Y. ■

19. Consider a uniform random variables $X \in [-A, A]$ connected to a uniform 8-bit quantizer. By doubling the number of bits, the mean squared error

a) decreases by a factor of 2^8; ☐

b) decreases by a factor of 2^{16}; ■

c) decreases by a factor that depends on A. ☐

20. Consider a pair of random variables $X_1, X_2 \in [0, 1]$ with a given joint probability density function $f_{X_1,X_2}(x_1, x_2)$. The mean squared error (MSE) per component achieved by a 6-bit vector quantizer, compared to that obtained by a pair of separate scalar quantizers of 3 bits each,

a) is always no larger; ■

b) is the same if the variables are independent; ☐

c) is the same if the variables have uniform densities. ☐

Parte II

Problema 1

Consider the random variable $X \in \{1, 2, 3, 4, 5\}$ associated with the symbols emitted by a memoryless source, uniformly distributed.

1. Determine the entropy of the source, $H(X)$, and compute a binary Huffman code. What is the codeword length and the efficiency of the obtained code?

   Solution: Since the distribution is uniform, $H(X) = \log_2 5 \simeq 2.322$ bits/symbol.

   A possible Huffman code is $\{C(1) = 00, C(2) = 01, C(1) = 10, C(1) = 110, C(1) = 111\}$, which has expected length $L(C) = (2 + 2 + 2 + 3 + 3)/5 = 12/5 = 2.4$ bits/symbol.

   The efficiency of this code is $\eta = \frac{H(X)}{L(C)} = \frac{\log_2 5}{2.4} \simeq \frac{2.322}{2.4} \simeq 0.9675$.

2. For the code computed in point 1), we obtained another code by swaping the two first codewords. Is this code optimal? Why?

   Solution: Of course it is still optimal; since all the symbols have the same probability, the expected length of the code is the same.

3. Modify the probability distribution of X defined in point 1) such that the optimal code has an efficiency of 100%.

   Solution: For perfect efficiency (an ideal code), we need the probabilities to be powers of 2.

   For example: $P(1) = \frac{1}{2}$, $P(2) = P(3) = P(4) = P(4) = \frac{1}{8}$.

   In this case, the entropy is $H(X) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{4}{8}\log_2\frac{1}{8} = 2$ bits/symbol, and an optimal code (for example $\{C(1) = 0, C(2) = 100, C(1) = 101, C(1) = 110, C(1) = 111\}$) has expected length $L(C) = 2$ bits/symbol.

4. Let $Y \in \{1, 2, 3, 4\}$ a random variable associated with the symbols emitted by a memoryless source with probabilities $p_1 > p_2 > p_3 > p_4$. The table below shows two ternary codes for this source. For both codes say if they are optimal or not and justify.

Y	Code A	Code B
1	0	0
2	1	20
3	21	1
4	22	21

   Solution: Code A is optimal. First of all, it is an instantaneous ternary code; any ternary code for an alphabet with four symbols necessarily has two words with 1 trit and two words with 2 trits. Of course, the optimal code is obtained by assigning the two short words to the two most probable symbols.

   Code B is not optimal. Consider the expected length of code B: $L(C_B) = p_1 + 2\,p_2 + p_3 + 2\,p_4$. Now consider the expected length of code A and compute the difference: $L(C_A) = p_1 + p_2 + 2\,p_3 + 2\,p_4$,

$$L(C_B) - L(C_A) = p_1 + 2\,p_2 + p_3 + 2\,p_4 - \big(p_1 + p_2 + 2\,p_3 + 2\,p_4\big) = p_2 - p_3 < 0,$$

   which shows that code B is not optimal, because the expected length of code A is strictly smaller.

Problema 2

Consider a first order Markovian source with alphabet $\{a, b, c, d\}$ and transition matrix

$P(X_t \mid X_{t-1})$	$X_t = a$	$X_t = b$	$X_t = c$	$X_t = d$
$X_{t-1} = a$	1/4	1/4	1/4	1/4
$X_{t-1} = b$	1/4	1/2	1/8	1/8
$X_{t-1} = c$	1/16	1/16	3/4	1/8
$X_{t-1} = d$	1/32	1/32	1/16	7/8

1. Show that the probability vector $\mathbf{p} = (3/37,\ 4/37,\ 10/37,\ 20/37)$ is the stationary distribution of the source.

   Solution: We simply have to show that $\mathbf{P}^T\mathbf{p} = \mathbf{p}$, where $\mathbf{P}$ is the transition matrix given above. In fact,

$$\mathbf{P}^T\mathbf{p} = \begin{bmatrix} 1/4 & 1/4 & 1/16 & 1/32 \\ 1/4 & 1/2 & 1/16 & 1/32 \\ 1/4 & 1/8 & 3/4 & 1/16 \\ 1/4 & 1/8 & 1/8 & 7/8 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \\ 10 \\ 20 \end{bmatrix} \frac{1}{37} = \frac{1}{32} \begin{bmatrix} 8 & 8 & 2 & 1 \\ 8 & 16 & 2 & 1 \\ 8 & 4 & 24 & 2 \\ 8 & 4 & 4 & 28 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \\ 10 \\ 20 \end{bmatrix} \frac{1}{37} = \frac{1}{37} \begin{bmatrix} 96/32 \\ 128/32 \\ 320/32 \\ 640/32 \end{bmatrix} = \mathbf{p}.$$

2. Assuming that the source is in stationary state, among all sequences of two symbols $(x_i, x_j) \in \{(a,a), (a,b), (a,c), (a,d), (b,a), (b,b), (b,c), (b,d), (c,a), (c,b), (c,c), (c,d), (d,a), (d,b), (d,c), (d,d)\}$, which is the most probable?

   Solution: The most probable pair is (d,d), because $P(X_t = d | X_{t-1} = d) = 7/8$, thus $P(X_t = d, X_{t-1} = d) = (7/8)(20/37) = 140/296 \simeq 0.473$, which is not far from $1/2$. The next most probable pair is (c,c), which has probability $(3/4)(10/37) = 30/148 \simeq 0.2027$.

3. Determine an optimal binary coding scheme for this source and the respective average codeword length.

   Solution: An optimal coding scheme is the following

$C(X_t \mid X_{t-1})$	$X_t = a$	$X_t = b$	$X_t = c$	$X_t = d$
$X_{t-1} = a$	00	01	10	11
$X_{t-1} = b$	00	1	010	011
$X_{t-1} = c$	010	011	1	00
$X_{t-1} = d$	000	001	01	1

   The expected length of each of these four codes is:

$$\begin{aligned} L(C_a) &= 2 \text{ bits/symbol} \\ L(C_b) &= \frac{2}{4} + \frac{1}{2} + \frac{3}{8} + \frac{3}{8} = \frac{7}{4} \text{ bits/symbol} \\ L(C_c) &= \frac{3}{16} + \frac{3}{16} + \frac{3}{4} + \frac{2}{8} = \frac{11}{8} \text{ bits/symbol} \\ L(C_d) &= \frac{3}{32} + \frac{3}{32} + \frac{2}{16} + \frac{7}{8} = \frac{19}{16} \text{ bits/symbol} \end{aligned}$$

   Thus, the global expected code length is

$$L(C) = 2\,\frac{3}{37} + \frac{7}{4}\,\frac{4}{37} + \frac{11}{8}\,\frac{10}{37} + \frac{19}{16}\,\frac{20}{37} = \frac{1}{37}\left[\frac{1}{4}\,(24 + 28 + 55 + 95)\right] = \frac{101}{74} \simeq 1.365 \text{ bits/symbol.}$$

4. Based on the previous result, provide an upper bound (strictly better than the trivial $H'(X) \leq \log_2 4$ bits/symbol) for the conditional entropy rate for this source.

   Solution: Since we know that $L(C) \geq H'(X)$, the result from the previous question guarantees that $H'(X) \leq \frac{101}{74}$ bits/symbol.

Problema 3

Consider a random variable with the following probability density function

$$f_X(x) = \begin{cases} \frac{1}{2} + \alpha\left(\left||x| - \frac{1}{2}\right| - \frac{1}{4}\right) & \Leftarrow \quad x \in [-1, 1[, \\ 0 & \Leftarrow \quad x \notin [-1, 1]. \end{cases}$$

where $\alpha \in [-2, 2]$ is a parameter. Notice that $\int_{-1}^{0}\left(\left||x| - \frac{1}{2}\right| - \frac{1}{4}\right) dx = \int_{0}^{1}\left(\left||x| - \frac{1}{2}\right| - \frac{1}{4}\right) dx = 0$. Notice also that in the interval $[\frac{1}{2}, 1]$ the function simplifies to $f_X(x) = \frac{1}{2} - \frac{\alpha\, 3}{4} + \alpha\, x$.

1. Consider $\alpha = 0$ and that X is connected to a non-uniform 2-bit scalar quantizer with the following four regions: $R_0 = [-1, 0]$, $R_1 =]0, 1/2]$, $R_2 =]1/2, 3/4]$ e $R_3 =]3/4, 1]$. Determine the optimal representative of each region and the exact value of the corresponding mean squared error.

   Solution: For $\alpha = 0$, the probability density function is uniform, $f_X(x) = 1/2$, for $x \in [-1, 1]$. Thus, the optimal representative of each region is located in the center of the region,

   $$y_0 = \frac{-1+0}{2} = -\frac{1}{2}, \qquad y_1 = \frac{0+1/2}{2} = \frac{1}{4},$$
   $$y_2 = \frac{1/2+3/4}{2} = \frac{5}{8}, \qquad y_3 = \frac{3/4+1}{2} = \frac{7}{8}.$$

   Since the density is uniform, the high resolution approximation provides the exact value of the MSE. Because the quantizer is not uniform, the MSE is given by

   $$MSE = \frac{1}{12}\sum_{i=0}^{3} \Delta_i^2\, p_i,$$

   where $\Delta_0 = 1$, $\Delta_1 = 1/2$, $\Delta_3 = \Delta_4 = 1/4$, whereas $p_0 = 1/2$, $p_1 = 1/4$, and $p_2 = p_3 = 1/8$. Finally,

   $$MSE = \frac{1}{12}\left(\frac{1}{2} + \left(\frac{1}{2}\right)^2 \frac{1}{4} + \left(\frac{1}{4}\right)^2 \frac{1}{8} + \left(\frac{1}{4}\right)^2 \frac{1}{8}\right) = \frac{1}{12}\,\frac{(64+8+1+1)}{128} = \frac{37}{768} \simeq 0.0482.$$

2. Consider still that $\alpha = 0$ and determine the exact value of the MSE achieved by a uniform 2-bit quantizer; compare with the result of the previous question and comment.

   Solution: A 2-bit quantizer has 4 regions of equal length, thus $\Delta_i = \Delta = 1/2$ and the MSE is simply

   $$MSE = \Delta^2/12 = 1/48 \simeq 0.0208.$$

   This is clearly better than the result obtained in the previous question, which is natural, because the density is uniform, thus the optical quantizer is uniform.

3. Consider now a generic $\alpha \in [-2, 2]$ and a 1-bit quantizer with regions $R_0 = [-1, 0[$ and $R_1 = [0, 1]$. Justify that the optimal representatives of these regions, y_0 and y_1, do not depend on α and give their values.

   Solution: For any value $\alpha \in [-2, 2]$, the density function f_X is symmetric around the center of each region, thus the corresponding center of mass is precisely the geometric center. The optimal representatives are thus $y_0 = -1/2$ and $y_1 = 1/2$.

4. Consider still a generic $\alpha \in [-2, 2]$ and a 1-bit optimal quantizer with regions $R_0 = [-1, 0[$ and $R_1 = [0, 1]$; knowing that

   $$\int_{\frac{1}{2}}^{1}\left(\frac{1}{2} - \frac{\alpha\, 3}{4} + \alpha\, x\right)\left(x - \frac{1}{2}\right)^2 dx = \frac{1}{48} + \frac{\alpha}{192}, \tag{1}$$

   determine the corresponding MSE. Compute the high resolution approximation, compare with the result obtained, and comment.

Solution: The exact value of the MSE is given by

$$MSE = \int_{-1}^{0} \left(x + \frac{1}{2}\right)^2 f_X(x)\, dx + \int_{0}^{1} \left(x - \frac{1}{2}\right)^2 f_X(x)\, dx = 2\int_{0}^{1} \left(x - \frac{1}{2}\right)^2 f_X(x)\, dx,$$

due to the symmetry of the function. Moreover, since both $f_X(x)$ and $(x-1/2)^2$ are also symmetric around $x = 1/2$, we have that

$$MSE = 4\int_{\frac{1}{2}}^{1} \left(x - \frac{1}{2}\right)^2 \left(\frac{1}{2} - \frac{\alpha\, 3}{4} + \alpha\, x\right) dx = \frac{1}{12} + \frac{\alpha}{48},$$

where the final equality results from (1).

The high resolution approximation (since $\Delta = 1$) is $\widehat{MSE} = \Delta^2/12 = 1/12$, showing that the exact value can be either larger (for $\alpha > 0$), smaller (for $\alpha < 0$), or equal (for $\alpha = 0$) to the high resolution approximation.

5. Still for a generic $\alpha \in [-2, 2]$, consider now a 2-bit quantizer, with regions $R_0 = [-1, -\frac{1}{2}[$, $R_1 = [-\frac{1}{2}, 0[$, $R_2 = [0, \frac{1}{2}[$, and $R_3 = [\frac{1}{2}, 1]$. Knowing that

$$\int_{\frac{1}{2}}^{1} x \left(\frac{1}{2} - \frac{\alpha\, 3}{4} + \alpha\, x\right) dx = \frac{3}{16} + \frac{\alpha}{96},$$

find the optimal representatives of the four regions.

Solution: We begin by computing the representative of region R_3, exploiting the given integral:

$$y_3 = \frac{\displaystyle\int_{\frac{1}{2}}^{1} x \left(\frac{1}{2} - \frac{\alpha\, 3}{4} + \alpha\, x\right) dx}{\displaystyle\int_{\frac{1}{2}}^{1} \left(\frac{1}{2} - \frac{\alpha\, 3}{4} + \alpha\, x\right) dx} = 4\int_{\frac{1}{2}}^{1} x \left(\frac{1}{2} - \frac{\alpha\, 3}{4} + \alpha\, x\right) dx = \frac{3}{4} + \frac{\alpha}{24},$$

where the second equality results from the fact that the denominator is simply $P(X \in R_3) = 1/4$.

Now, observing the symmetries of the density function (illustrated in the picture below), it is clear that

$$y_0 = -y_3, \qquad y_2 = 1 - y_3, \qquad y_1 = -y2.$$