Version RV/5

EGT2
ENGINEERING TRIPOS PART IIA

Thursday 30 April 2026   9.30 to 11.10

Module 3F7

INFORMATION THEORY AND CODING

*Answer not more than **three** questions.*

All questions carry the same number of marks.

*The **approximate** percentage of marks allocated to each part of a question is indicated in the right margin.*

*Write your candidate number **not** your name on the cover sheet.*

STATIONERY REQUIREMENTS
Single-sided script paper

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM
CUED approved calculator allowed
Engineering Data Book

10 minutes reading time is allowed for this paper at the start of the exam.

You may not start to read the questions printed on the subsequent pages of this question paper until instructed to do so.

You may not remove any stationery from the Examination Room.

1 (a) There are two bags, labelled 1 and 2. Bag 1 contains two red balls and one blue ball, while Bag 2 contains two blue balls and one red ball. One of the bags is chosen (randomly and uniformly), and a ball is drawn from it. Let $X \in \{1, 2\}$ denote the chosen bag and let $Y \in$ {Red, Blue} denote the colour of the ball drawn from it.

(i) Compute $H(X), H(Y)$ and $H(X, Y)$. [20%]

(ii) A second ball is drawn from the same bag (uniformly at random from the two remaining balls in the bag). Let Z denote the colour of this second ball. Compute $H(Z \mid Y)$. [25%]

(iii) Compute $I(X; Z \mid Y)$ and $I(X; Y, Z)$. [20%]

(b) Consider two probability mass functions (pmf) P and Q on the same alphabet $\mathcal{X}$, and let $D(P\|Q)$ denote the relative entropy from P to Q. Consider a function $Y = f(X)$ for $X \in \mathcal{X}$, and let P_Y denote the pmf of Y when X is distributed according to P. Similarly, let Q_Y denote the pmf of Y when X is distributed according to Q.

(i) Prove that $D(P_Y\|Q_Y) \leq D(P\|Q)$. [25%]

Hint: The following "log-sum" inequality may be useful. For any non-negative numbers $a_1, \ldots, a_n, b_1, \ldots, b_n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}.$$

(ii) Give an interpretation for the inequality in part (b)(i) and a condition for when equality holds. [10%]

2 (a) Which of the following codes could be a binary Huffman code for some probability distribution over a 4-symbol alphabet? For those that cannot, provide a reason.

{1, 01, 001, 000}

{00, 01, 10, 110}

{0, 10, 01, 11} [15%]

(b) Consider the set of all discrete memoryless channels with input alphabet of size 4 and an output alphabet of size 3. Determine whether the following statement is TRUE or FALSE, and justify your answer. "The maximum capacity of a channel in this set is 2 bits per channel use." [15%]

(c) A source produces a sequence of symbols $X_1, X_2, \ldots$ from an alphabet $\{a, b, c\}$ as follows. The first symbol of the sequence is equally likely to be a, b, or c. For $k \geq 2$, each successive symbol X_k is the same as X_{k-1} with probability 0.5, or is equal to one of the other two symbols, with probability 0.25 each. An arithmetic encoder is used to compress the sequence produced by the source.

(i) Let $(X_1, X_2, X_3) = (b, a, c)$. Determine the interval produced by the arithmetic encoder for this sequence, and the corresponding binary codeword. [25%]

(ii) Compute the best entropy-based lower bound on the minimum expected code length of *any* prefix-free code for (X_1, X_2, X_3). [15%]

(d) A source produces a sequence of symbols from an alphabet $\{a, b, c\}$. These symbols are independent and identically distributed according to the distribution $P(a) = 0.5$, $P(b) = p$, $P(c) = 0.5 - p$, where p can take one of two values: $p = 0.1$ or $p = 0.2$. Knowing that p equals one of these two values, but not which one, we wish to design a compression code based on a distribution Q on $\{a, b, c\}$. Determine the distribution Q that minimizes the worst-case redundancy of the code among the two choices for p. [30%]

3 (a) A discrete memoryless channel has input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and transition probabilities given by the conditional distribution $P_{Y|X}$. The capacity of the channel is C bits/transmission.

(i) Consider a channel code of length n and rate R that achieves a small probability of decoding error. Briefly describe how the codebook is constructed, as well as the encoding and decoding operations. (You may assume that n is large and that you do not have to worry about the complexity of encoding and decoding.) [20%]

(ii) A ternary source with alphabet $\{a, b, c\}$ produces independent and identically distributed (i.i.d.) symbols according to the following probability distribution: $P(a) = 0.5, P(b) = P(c) = 0.25$. We wish to transmit m symbols of the source across the channel. The m symbols are first compressed into a sequence of bits. The output of the compressor is then mapped to a channel codeword of length n. Determine the minimum number of channel uses n required to reliably recover the m source symbols at the destination, for large m. Express your answer in terms of m and C. You may assume optimal compression and decompression, as well as optimal channel encoding and decoding. [10%]

(b) Consider a discrete memoryless channel with input alphabet $\{1, 2, \ldots, M\}$ and output alphabet $\{1, 2, \ldots, M, ?\}$, where ? denotes an erasure. Here M is an integer greater than 2. Given a channel input symbol $X \in \{1, \ldots, M\}$, the output Y is an erasure with probability ε or is equal to X with probability $(1-\varepsilon)$. Compute the capacity of the channel and determine a capacity-achieving input distribution. [25%]

(c) A phase-based communication channel has input X and output Y taking values in the interval $[0, 2\pi)$. Given an input $X \in [0, 2\pi)$, the output Y is generated as

$$Y = (X + Z) \text{ (mod) } 2\pi,$$

where $Z \in [0, 2\pi)$ is additive phase noise independent of X, with probability density function $p_Z(z)$. The (mod) 2π operation returns the remainder when divided by 2π. For example, if $X = 7\pi/4$ and $Z = 3\pi/2$, then $Y = 7\pi/4 + 3\pi/2 - 2\pi = 5\pi/4$.

(i) Show that for *any* noise density $p_Z(z)$, $z \in [0, 2\pi)$, the capacity is achieved by the uniform input distribution, i.e. the density $p_X(x) = \frac{1}{2\pi}$ for $x \in [0, 2\pi)$. [25%]
Hint: You may use the fact that among probability densities supported on an interval $[a, b)$, the uniform distribution on $[a, b)$ has the maximum differential entropy.

(ii) Compute the capacity when the noise distribution is $p_Z(z) = \frac{\alpha e^{-\alpha z}}{(1-e^{-2\pi\alpha})}$, for $z \in [0, 2\pi)$, where $\alpha > 0$ is a constant. [20%]

4 Consider a binary linear code which maps three information bits $\underline{x} = (x_1, x_2, x_3)$ into a six-bit codeword $\underline{c} = (c_1, c_2, c_3, c_4, c_5, c_6)$ as follows:

$$c_1 = x_2,\ c_2 = x_1 + x_3,\ c_3 = x_1 + x_2,\ c_4 = x_2 + x_3,\ c_5 = x_1 + x_2 + x_3,\ c_6 = x_3$$

where + denotes binary modulo 2 addition.

(a) What is the rate of the code? [5%]

(b) Write down the generator matrix $\mathbf{G}$ for the code. [15%]

(c) Write down the set of codewords in the code. [10%]

(d) Verify that the following matrix $\mathbf{H}$ is a valid parity check matrix for the code, and write down a systematic generator matrix $\mathbf{G}_{\text{sys}}$. [15%]

$$\mathbf{H} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

(e) A codeword from the code is transmitted over a binary symmetric channel with crossover probability 0.1, and the received sequence is $\underline{y} = [0,\ 1,\ 1,\ 0,\ 0,\ 1]$. Consider decoding the received sequence using a belief propagation decoder based on the parity check matrix in part (d), with messages in the log-likelihood ratio (LLR) format. Suppose that we stop the belief propagation decoder after one complete iteration of message passing, i.e. after one round of variable-to-check and check-to-variable messages. Compute the final LLR for the *third* code bit, and state whether it will be decoded to a 0 or 1. [25%]

(f) For the same channel and the received sequence $\underline{y}$ in part (e), compute the *exact* ratio of the posterior probabilities of the third code bit being a 0 versus a 1, given $\underline{y}$. Express this ratio in LLR format and compare with your answer in part (e). [30%]

END OF PAPER

THIS PAGE IS BLANK