Version RV/4

EGT2
ENGINEERING TRIPOS PART IIA

Thursday 27 April 2023   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 Consider two sources U_1 and U_2 taking values in the alphabet $\mathcal{U} = \{a, b, c, d, e\}$ with probability distributions $P_1 = \left\{\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right\}$ and $P_2 = \left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16}\right\}$, respectively. We consider a new source X which produces a symbol as follows. First, a selector independently chooses either source U_1 or U_2, with U_1 chosen with probability $\frac{1}{3}$ and U_2 with probability $\frac{2}{3}$. If the selector chooses U_1, then X is drawn according to P_1, otherwise it is drawn according to P_2.

(a) Find the probability mass function P_X of the source X. [15%]

(b) Specify an optimal symbol code for the source X and calculate its expected code length. [25%]

(c) Consider a sequence of symbols $(X_1, \ldots, X_n)$ produced independently and identically distributed according to P_X. The sequence is compressed using an arithmetic encoder. Compute an upper bound on the expected length of the codeword for $(X_1, \ldots, X_n)$. [10%]

(d) Consider again a sequence of symbols $(X_1, \ldots, X_n)$ produced independently and identically distributed according to P_X, but now suppose that both the encoder and decoder know which of U_1 or U_2 each X_i came from. Briefly describe how you would construct a practical compression scheme whose expected code length is close to the optimal value when n is large. Determine the expected number of bits per source symbol for the scheme as $n \to \infty$. [20%]

(e) Consider now a sequence of m symbols $(Z_1, \ldots, Z_m)$ from the source U_2, i.e., produced independently and identically distributed according to the probability distribution P_2. Assume that m is large. We wish to transmit the source sequence over a communication channel with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and capacity $\mathcal{C}$ bits/transmission. Suppose that we use a channel code with rate $R < \mathcal{C}$ bits/transmission.

(i) What is the minimum number of channel transmissions necessary so that $(Z_1, \ldots, Z_m)$ can be reconstructed at the decoder with high probability? Give your answer in terms of m and R. [15%]

(ii) Consider a channel code of rate $R < \mathcal{C}$ and code length n that is 5% larger than the minimum value determined above. With these parameters, explain why a unique codeword cannot be assigned to all possible length m source sequences generated from the source U_2. Despite this, why is it possible to reconstruct the source sequence $(Z_1, \ldots, Z_m)$ at the channel decoder with high probability? [15%]

2 (a) Let X be a binary random variable taking values in $\{0, 1\}$ with probability mass function $\{p, (1-p)\}$. Let Y be another random variable, independent from X, taking values in $\{1, \ldots, r\}$ with probability mass function $\{q_1, \ldots, q_r\}$. Let $Z = XY$ be the product of X and Y.

(i) Determine the probability distribution of Z. [10%]

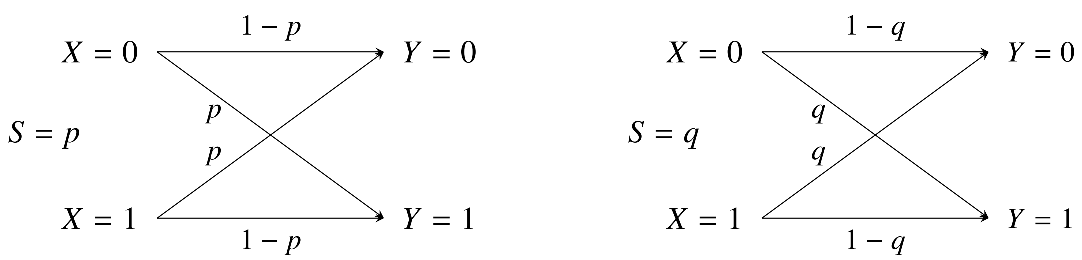
(ii) Determine $H(Z)$, the entropy of Z, in terms of p, $H(X)$ and $H(Y)$. [10%]

(b) A random channel code is constructed as follows. First, two random bits $a \in \{0, 1\}$ and $b \in \{0, 1\}$ are independently generated using the probability mass function $\left\{\frac{1}{3}, \frac{2}{3}\right\}$. Then, a linear code is constructed using the following generator matrix:

$$\mathbf{G} = [a \;\; a \;\; a \;\; a \;\; b].$$

What is the probability that the generated code is able to correct at least 2 channel errors (bit-flips)? *Hint*: Construct all the possible codes. [30%]

(c) Consider a binary channel with the input symbol X and output symbol Y both taking values in $\{0, 1\}$. At each time instant, the channel is a binary symmetric channel (BSC) with crossover probability either p or q with equal probability. Specifically, for each $i = 1, 2, \ldots$, there is a *state* variable $S_i \in \{p, q\}$, which is equal to either p or q with equal probability. As shown below, $P_{Y_i|X_i}$ corresponds to BSC(p) channel if $S_i = p$, and otherwise corresponds to a BSC(q) channel. Assume that $S_1, S_2, \ldots$ are independent from one another and also independent from the channel inputs $X_1, X_2, \ldots$.



(i) Determine the conditional distribution $P_{Y|X}$. [10%]

(ii) Compute the capacity of the channel when the state sequence is not known to either the encoder or the decoder. [10%]

(iii) Compute the capacity of the channel when the state sequence $S_1, S_2, \ldots$ is known to both the encoder and decoder. [20%]

(iv) For part (iii), describe how you could construct a capacity-achieving scheme using two capacity-achieving codes, one for BSC(p) and another for BSC(q). [10%]

3 (a) Consider a binary input, binary output memoryless channel. The channel input sequence is $X^n = (X_1, \ldots, X_n)$ and the output sequence is $Y^n = (Y_1, \ldots, Y_n)$, where each X_i and Y_i takes values in $\{0, 1\}$. The output sequence Y^n is processed via a function g to produce $Z = g(Y^n)$. Label each of the the following statements as TRUE or FALSE, with justifications.

(i) $H(Y^n \mid Z) = 0$. [10%]

(ii) $I(X^n ; Y^n) \leq n - H(X^n \mid Y^n)$. [15%]

(iii) $H(Z) \leq n$. [10%]

(b) Consider a channel for which the input and output symbols are each 8-bit binary vectors. Each time it is used, the channel flips *exactly one* of the bits in the input symbol, but the receiver does not know which one. The other seven bits are received without error. Each of the 8 bits in the input symbol is equally likely to be the one that is flipped.

(i) Determine the capacity of the channel. [30%]

(ii) Show by describing an explicit encoder and decoder that it is possible to communicate error-free over the channel at the rate of 5 bits/channel use. [35%]

Hint: The $(7, 4)$ Hamming code described by the following parity check matrix may be useful.

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

4 (a) Consider a binary linear code with code length n and a single parity check constraint on the code bits $(c_1, \ldots, c_n)$, given by: $c_1 + c_2 + \ldots + c_n = 0$. That is, the modulo-two sum of all the code bits equals zero.

(i) Specify a parity check matrix for the code, and determine the code rate. [15%]

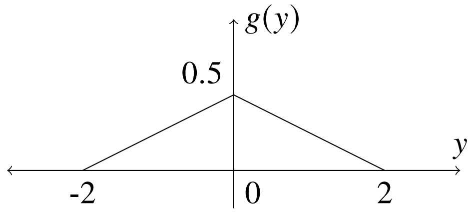
(ii) Find a systematic generator matrix for the code. [10%]

(b) Consider a binary repetition code where a single information bit is repeated to form a length n codeword. That is, 0 is mapped to the all-zeros codeword, and 1 to the all-ones codeword.

(i) Specify a generator matrix for the code. [10%]

(ii) Find a systematic parity matrix for the code. [10%]

(c) Let the function $g(y)$ be the probability density function shown in the figure below.



Consider a memoryless channel with binary inputs $x \in \{-1, 1\}$, and a continuous-valued output y generated according to a conditional density $f(y \mid x) = g(y - x)$, for $x \in \{-1, 1\}$. Compute the likelihood ratio $\frac{f(y|x=1)}{f(y|x=-1)}$ for y in the interval $(-3, 3)$. [20%]

(d) Consider a binary linear code with the following parity check matrix.

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

The code is used over the channel defined in part (c) with the following mapping. Each 0 in the codeword is mapped to a channel input $+1$ and each 1 in the codeword is mapped to a channel input -1. The received sequence is $\underline{y} = [0.1, 0.3, -0.4, 0.9, -1.1]$. A belief propagation decoder is used for decoding, with messages in log-likelihood ratio (LLR) format. Suppose that we stop the 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 *second* code-bit. [35%]

END OF PAPER

THIS PAGE IS BLANK