Version RV/4

EGT2
ENGINEERING TRIPOS PART IIA

Thursday 25 April 2024   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 Let X and Y be independent, identically distributed random variables with four equiprobable outcomes $\{1, 2, 3, 4\}$. Let $Z = X + Y$.

(a) Determine the probability mass function of Z. [25%]

(b) Suppose person A observes Z, but person B does not. B learns the value by asking A the following sequence of yes/no questions: "Is $Z = 2$? Is $Z = 3$? ... Is $Z = 7$?", stopping when Z is determined. What is the expected number of questions that person B asks? [20%]

(c) Design an optimal strategy to learn the value of Z via a sequence of yes/no questions. An optimal strategy is one that minimizes the expected number of questions. Compute the expected number of questions with this strategy. [25%]

(d) Compute the conditional joint entropy $H(X, Y \mid Z)$. [30%]

2 (a) Consider a random variable X which takes values in the set $\{1, \ldots, M\}$, with probability mass function P given by

$$P(1) = 2^{-k_1},\ P(2) = 2^{-k_2}, \ldots, P(M) = 2^{-k_M},$$

where $k_1 \leq k_2 \leq \ldots \leq k_M$ are suitable positive integers so that P is a valid probability distribution.

(i) Compute the expected codelength of the Shannon-Fano code for X and show that it is optimal. [25%]

(ii) Consider another Shannon-Fano code constructed for a pmf Q given by

$$Q(1) = 2^{-c_1},\ Q(2) = 2^{-c_2}, \ldots, Q(M) = 2^{-c_M},$$

where $c_1 \leq c_2 \leq \ldots \leq c_M$ are suitable positive integers that may differ from $k_1, \ldots, k_M$. If this code is used to represent a symbol generated according to the distribution P, compute the expected codelength. [15%]

(iii) Show that the expected codelength in part (a)(ii) is strictly larger than the one in part (a)(i) unless $P = Q$. That is, unless the two distributions are identical, we incur a penalty in the expected codelength by using an optimal code for Q when the true distribution is P. [20%]

(b) An urn contains r red, w white and b black balls.

(i) If we draw a ball from the urn uniformly at random and denote its colour by X_1, what is the entropy $H(X_1)$? Express your answer in terms of r, w, b. [10%]

(ii) Suppose we draw K balls from the urn independently *with replacement*, and denote their colours by $X_1, \ldots, X_K$. (With replacement means that before drawing the second ball, the first is put back into the urn, and so on.) Compute the joint entropy $H(X_1, \ldots, X_K)$. [10%]

(iii) Suppose that we draw K balls from the urn *without replacement*, and denote their colours by $Y_1, \ldots, Y_K$. (Without replacement means that the first ball is not put back into the urn before the second is drawn, and so on.) Compare $H(Y_1, \ldots, Y_K)$ with $H(X_1, \ldots, X_K)$ computed in part (b)(ii). You only need to specify which of the two joint entropies is larger and justify your answer; you are not expected to explicitly compute $H(Y_1, \ldots, Y_K)$ or give a rigorous proof. [20%]

3 (a) Let X be any continuous random variable whose probability density function is zero outside an interval $[A, B]$. Show that the differential entropy $h(X)$ satisfies

$$h(X) \leq \log_2(B - A),$$

with equality if and only if X is uniformly distributed on $[A, B]$. [20%]

(b) Consider a binary-input, continuous output channel $Y = V + Z$, where the binary input V takes values in the set $\{1, -1\}$, and the noise Z is independent of V and uniformly distributed in the interval $[-1, 1]$. Compute the capacity of the channel and specify a capacity-achieving input distribution. [20%]

(c) Let X be any continuous random variable with $\mathbb{E}[X^2] \leq \tau^2$. Then show that the differential entropy $h(X)$ satisfies

$$h(X) \leq \frac{1}{2}\log_2(2\pi e\tau^2),$$

with equality if and only if X is distributed as $\mathcal{N}(0, \tau^2)$, i.e., X is Gaussian with mean 0 and variance τ^2. [25%]

(d) Let X_1 and X_2 be any continuous random variables such that $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \mu$, and $\mathbb{E}[X_1^2] = \mathbb{E}[X_2^2] = P$ for some constants μ and P. Let $Y = X_1 + X_2 - 2\mu$.

(i) Show that for any valid probability distributions on X_1, X_2, we must have $\mu^2 \leq P$. [10%]

(ii) Find the maximum differential entropy $h(Y)$ if X_1 and X_2 are independent. Also specify probability distributions of X_1 and X_2 that maximize $h(Y)$. [10%]

(iii) Find the maximum differential entropy $h(Y)$ if X_1 and X_2 are allowed to be dependent. Also specify a joint probability distribution of (X_1, X_2) that maximizes $h(Y)$. [15%]

4 Consider a binary linear code with the following generator matrix:

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

(a) What are the dimension and the rate of the code? [10%]

(b) A codeword from the code was transmitted over a binary erasure channel, and the received sequence was $\underline{y} = [1, ?, ?, ?, 0, 1]$ where the ? outputs are erasures. Find the transmitted codeword. [15%]

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

(d) Find a systematic parity check matrix $\mathbf{H}$ for the code. [10%]

(e) Find the minimum distance of the code. [10%]

(f) Draw the factor graph corresponding to the parity check matrix $\mathbf{H}$ found in part (d). [10%]

(g) A codeword from the code is transmitted over a binary-input additive white Gaussian noise (AWGN) channel 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 $\underline{y} = [-0.6,\ 0.5,\ 0.2,\ 0.4,\ -1.2,\ 0.7]$. The channel noise variance is $\sigma^2 = 1$. We run belief propagation decoding to recover the transmitted codeword, with messages in the log-likelihood ratio (LLR) format.

(i) In the first iteration, what is the message sent by the first variable node to the check nodes connected to it? [5%]

(ii) 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 LLRs for the *third* and *sixth* code-bits and state whether each will be decoded to a 0 or a 1. [30%]

END OF PAPER

THIS PAGE IS BLANK