EGT2
ENGINEERING TRIPOS PART IIA

Thursday 1 May 2025 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) Company A has a return on investment that is uniformly distributed in the interval $[-1, 1]$, while company B has a return that is uniformly distributed in $\left[-\frac{1}{2}, \frac{1}{2}\right]$. An investor makes a random choice, investing in company A with probability $\frac{2}{3}$ and in company B with probability $\frac{1}{3}$. Let $X \in \{A, B\}$ be the random variable denoting the investor's choice, and let Y be the random variable denoting the return on the chosen investment.

(i) Compute the probability density function of Y. [15%]

(ii) Determine the mutual information $I(X;Y)$. [15%]

(b) Let X be a continuous random variable with differential entropy $h(X)$. Let $Y = aX$ for a constant $a > 0$. Then show that [20%]

$$h(Y) = h(X) + \log_2(a).$$

Hint: Note that the cumulative distribution function of Y is related to that of X via $P(Y \leq y) = P(X \leq y/a)$, for all y. Also recall that the probability density function is the derivative of the cumulative distribution function.

(c) Compare the following expressions and label each with $\geq$, $=$, or $\leq$. Justify your answers.

(i) $H(X,Y,Z) - H(X,Y)$ versus $H(X,Z) - H(X)$. [15%]

(ii) $I(X\,;\,Z \mid Y) - I(Z\,;\,Y \mid X)$ versus $I(X\,;\,Z) - I(Z\,;\,Y)$. [15%]

(d) Prove that for any jointly distributed discrete random variables X, Y, Z, the following inequality holds: [20%]

$$2H(X,Y,Z) \leq H(X,Y) + H(Y,Z) + H(Z,X).$$

2 (a) A discrete source produces i.i.d. symbols in the alphabet $\{a, b\}$ with the distribution $P(a) = 0.99$, $P(b) = 0.01$. We consider source sequences of length $n = 100$ symbols, and wish to construct a binary code assigning a unique codeword to each source sequence.

(i) Give the best possible lower bound on the minimum expected code length per source symbol of any binary code that assigns a unique codeword to each source sequence. [10%]

(ii) An equal length binary codeword is assigned to each length-100 sequence with 97 or more a symbols. Find the minimal length of each such codeword. [10%]

(iii) A length-100 binary codeword is assigned to each of the sequences not covered in part (ii). This, together with the codewords assigned in part (ii), defines a code that assigns a unique codeword to each length-100 source sequence. Calculate the expected code length per source symbol for this code. (You can assume that no additional flag bits are included to distinguish between the two types of codewords.) [20%]

(iv) Briefly describe how the binary code in parts (ii) and (iii) should be modified so that its expected length per source symbol approaches the optimal value. [15%]

(b) Consider Huffman coding for an alphabet $\{1, 2, \ldots, m\}$ with symbol probabilities $p_1 > p_2 \geq \ldots \geq p_m$. Assume $m > 2$.

(i) Prove that for any binary Huffman code, if the most probable message symbol has probability $p_1 > \frac{2}{5}$, that symbol must be assigned a codeword of length 1. [25%]

(ii) Prove that for any binary Huffman code, if the most probable message symbol has probability $p_1 < \frac{1}{3}$, that symbol must be assigned a codeword of length at least 2. [20%]

3 (a) Consider a channel with a ternary input alphabet $\{0, 1, 2\}$ and binary output alphabet $\{0, 1\}$. For inputs $X = 0$ and $X = 1$, the output $Y = X$, and for $X = 2$, the output Y is equally likely to be 0 or 1. Determine the capacity of the channel. [15%]

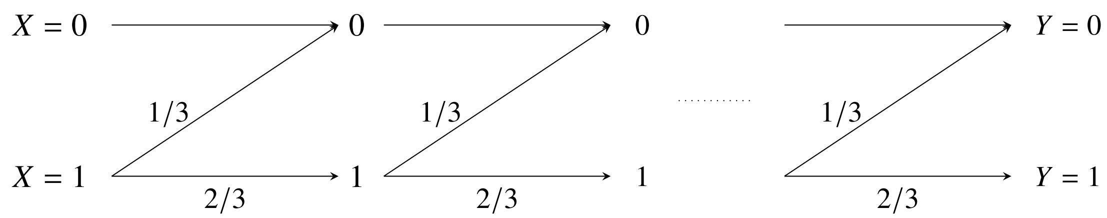
(b) Consider a Z-channel with binary input and output alphabets, and transition probabilities $P_{Y|X}$ given by the following matrix:

	$P_{Y\|X}$	$Y = 0$	$Y = 1$
X	0	1	0
	1	1/3	2/3

(i) Determine the capacity of the channel and a capacity-achieving input distribution. [25%]

Hint: The binary entropy function $H_2(x) = x \log_2 \frac{1}{x} + (1 - x) \log_2 \frac{1}{(1-x)}$ has derivative $\frac{dH_2(x)}{dx} = \log_2 \frac{(1-x)}{x}$, for $x \in (0, 1)$.

(ii) Consider a cascade of m independent Z-channels as shown in the figure below. Write down the transition probability matrix for this cascade channel. What is its capacity as $m \to \infty$? [15%]



(c) Consider a channel whose input and output alphabets are the set $\{0, 1, \ldots, 23\}$. The output Y is produced from the input X according to

$$Y = (X + Z) \mod 24,$$

where the noise variable Z takes one of three values $\{0, 1, 2\}$ with equal probability. The mod 24 operation returns the remainder when divided by 24. For example, $(23 + 1) \mod 24 = 0$, $(23 + 2) \mod 24 = 1$, and $(23 + 0) \mod 24 = 23$.

(i) Compute the capacity C of the channel (in bits/channel use), and specify a capacity-achieving input distribution. [20%]

(ii) Specify a coding scheme to communicate information with *zero* error at the rate of C bits/channel use. You should describe the encoding and the decoding used for your code. [25%]

4 (a) Consider an (n, k) binary linear code defined via a generator matrix $\mathbf{G}$. Show that the linear code is a subspace of $\{0, 1\}^n$, the space of length-n binary sequences. [20%]

Hint: You need to show that any linear combination of codewords is also a codeword, where the coefficients of linear combination are in $\{0, 1\}$.

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

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

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

(ii) Consider a memoryless channel with binary inputs $x \in \{-1, 1\}$, and a continuous-valued output $y \in \mathbb{R}$ generated according to a conditional density $f(y \mid x)$, where

$$f(y \mid x = 1) = \begin{cases} \frac{1}{3}e^{-y/2}, & y \geq 0, \\ \frac{1}{3}e^{y}, & y < 0, \end{cases}$$

and $f(y \mid x = -1) = f(-y \mid x = 1)$. Compute and sketch the log-likelihood ratio $\ln \frac{f(y|x=1)}{f(y|x=-1)}$ for $y \in \mathbb{R}$. [15%]

(iii) A codeword from the code is transmitted over the channel defined in part (ii) 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 $\underline{y} = [2.1, -0.1, 0.6, 0.4, -1.5]$.

We run belief propagation decoding to recover the transmitted codeword, 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 LLRs for the *fourth* code bit and state whether it will be decoded to a 0 or a 1. [25%]

(iv) Determine the set of codewords in the code. [10%]

(v) For the same channel and the received sequence $\underline{y}$ in part (iii), compute the *exact* ratio of the posterior probabilities of the fourth code bit being a 0 versus a 1. [20%]

END OF PAPER

THIS PAGE IS BLANK