Version RV/4

EGT2
ENGINEERING TRIPOS PART IIA

Thursday 29 April 2021 9.00 to 10.40

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 and at the top of each answer sheet.*

STATIONERY REQUIREMENTS
Write on single-sided paper.

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM
CUED approved calculator allowed.
You are allowed access to the electronic version of the Engineering Data Books.

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

The time taken for scanning/uploading answers is 15 minutes.

Your script is to be uploaded as a single consolidated pdf containing all answers.

1 (a) Consider random variables X, Y, Z which form a Markov chain. That is,

$$P_{X,Y,Z} = P_X\, P_{Y|X}\, P_{Z|Y}.$$

Suppose that X takes values in the set $\{1, \ldots, n\}$, Y takes values in $\{1, \ldots, k\}$, and Z takes values in $\{1, \ldots, m\}$, where $n > k$ and $m > k$.
Prove that $I(X; Z) \leq \log_2 k$. [20%]

(b) A source emits N symbols from a 9-character alphabet $\{$`a, b, c, ..., i`$\}$. The first symbol X_1 is equally likely to be any of the 9 symbols. The successive symbols $(X_2, \ldots, X_N)$ are generated as follows. For all $k \geq 2$, the symbol X_k is equal to X_{k-1} with probability 0.99, and different from X_{k-1} with probability 0.01; and if it is different, X_k is equally likely to be any of the 8 symbols that are not X_{k-1}.

(i) Specify an optimal binary code for encoding the *first* symbol X_1, and calculate its expected codelength. [25%]

(ii) Calculate $H(X_k \mid X_{k-1})$, for $k \geq 2$. [25%]
Hint: To simplify computation, you may use $H_2(0.01) \approx 0.08$, where $H_2(p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{(1-p)}$, for $p \in (0, 1)$.

(iii) Calculate the joint entropy $H(X_1, \ldots, X_N)$, expressing your answer in terms of N. [20%]

(iv) Name a practical compression scheme whose expected code length would be close to the minimum value when N is large. Give a bound on its expected code length. [10%]

2 (a) A source produces n independent and identically distributed (i.i.d.) symbols from the alphabet $\{a, b, c\}$, with probability distribution $P(a) = 0.98, P(b) = P(c) = 0.01$. Its set of typical length-n sequences is defined as

$$A_{\epsilon,n} = \left\{x^n \,:\, H - \epsilon \,\leq\, \frac{-1}{n}\log_2 P(x^n) \,\leq\, H + \epsilon\right\}.$$

Here $P(x^n) = \prod_{i=1}^{n} P(x_i)$, and H is the entropy of the distribution P.

(i) For $\epsilon = 0.01$ and $n = 100$, precisely describe what the sequences in $A_{0.01,100}$ look like (in terms of the number of a, b, and c symbols in the sequences). In your calculations, you may round all numbers to two decimal places. [30%]

(ii) Suppose that we assign a unique codeword to each sequence in $A_{0.01,100}$. Assuming all the codewords have equal length, what is the minimum code length in bits? [15%]

(b) Consider any random variable X over the set $\{1, 2, \ldots, N\}$, such that $\mathbb{E}[X] = \alpha$, for some constant α in the interval $[1, N]$.

(i) Among all such random variables, show that the one with maximum entropy has the following probability mass function (pmf):

$$P_X^*(i) = 2^{-(\lambda_0+\lambda_1 i)}, \qquad i = 1, \ldots, N.$$

Here λ_0, λ_1 are constants chosen to ensure that it is a valid pmf, and that the expected value equals α. [30%]

Hint: Consider $D(Q_X \| P_X^*)$, for any other pmf Q_X over $\{1, 2, \ldots, N\}$ with expected value equal to α.

(ii) Among random variables taking values in the set $\{1, 2, 3\}$ with expected value equal to 1.5, determine the pmf of the one with maximum entropy. [25%]

3 (a) Consider a channel with input alphabet $\{-1, 0, 1\}$. Given input symbol X, the channel generates the output symbol $Y \in \{-1, 0, 1\}$ in two steps: it first generates $Z = X + N$, where N is a (continuous) noise random variable uniformly distributed on the interval $[-1, 1]$. It then maps Z to Y according to the following rule:

$$Y = \begin{cases} -1, & \text{if } Z \leq -1 \\ 0, & \text{if } -1 < Z < 1 \\ 1, & \text{if } Z \geq 1. \end{cases}$$

(i) Write down the transition probability matrix containing the values of $P_{Y|X}$ for the channel. [20%]

(ii) Compute the capacity of the channel, and the capacity achieving input distribution. [30%]

Hint: The derivative of the binary entropy function is $\frac{\mathrm{d}}{\mathrm{d}\alpha} H_2(\alpha) = \log_2 \frac{(1-\alpha)}{\alpha}$ for $\alpha \in (0, 1)$.

(b) Consider the channel $Z = X + N$, where as in part (a), the input symbol $X \in \{-1, 0, 1\}$, and N is a continuous random variable uniformly distributed on the interval $[-1, 1]$.

(i) For the input distribution $P_X(1) = P_X(-1) = p, P_X(0) = 1 - 2p$, compute and sketch the probability density function of Z. Here p is a constant that lies in the interval $(0, \frac{1}{2})$. [20%]

(ii) Compute the capacity of the channel with input X and output Z, and specify a capacity achieving input distribution. [30%]

Hint: Among probability density functions that are non-zero only on an interval $[A, B]$, the uniform distribution on $[A, B]$ has maximum differential entropy.

4 Consider a binary linear code with the following parity check matrix.

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

(a) Determine the dimension and rate of the code. [10%]

(b) Find a systematic parity check matrix for the code. [20%]

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

(d) A codeword from this code is transmitted over a binary symmetric channel (BSC), and the received sequence is [1, 1, 1, 1, 1, 1]. If we use a maximum-likelihood (minimum-distance) decoder, what is the decoded codeword? [10%]

(e) Draw the factor graph corresponding to the parity check matrix $\mathbf{H}$ in (1). [15%]

(f) A codeword 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 $\underline{y} = [-0.5, 0.5, 1.2, 0.6, 0.2, 0.8]$. The channel noise variance is $\sigma^2 = 1$.

(i) Consider decoding the received sequence using belief propagation, with messages in the log-likelihood ratio (LLR) format. 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 *fourth* code-bit and the *fifth* code-bit, and state whether each of these will be decoded to a 0 or a 1. [30%]

END OF PAPER

THIS PAGE IS BLANK