National University of Singapore
Department of Electrical & Computer Engineering

Midterm for

EE5139 Information Theory and its Applications
(Semester I, 2021/22)
September 2021

Time Allowed: 1.0 hours

INSTRUCTIONS FOR CANDIDATES:

- This paper contains **THREE (3)** questions, printed on **THREE (3)** pages.
- The total number of marks is 25.
- You can consult the lecture notes, your own notes and all the homework problems and other materials distributed in class.
- You can also use the internet but you CANNOT communicate with anybody during the exam.
- Try to answer all questions. Not all questions are equally difficult. Invest your time wisely.
- The notation used is the same as in the lecture notes. All logarithms (log) are to the base 2.

Question 1: Warmup (total: 5 points)

The following are TRUE/FALSE questions. Explanations are not needed. Please just write TRUE or FALSE. Please indicate the part of the question (1, 2, etc.) in your answer script clearly.

1. **(1 point)** We always have $I(X;Y|Z) \leq I(X;Y)$.

2. **(1 point)** We have $D(P\|Q) \neq 0$ unless $P = Q$ for any two pmfs P and Q.

3. **(1 point)** There exists a binary prefix-free variable-length source code with lengths $(1, 3, 3, 3, 4, 4)$.

Please answer the following question in at most two sentences.

4. **(2 points)** Explain the difference between a strong and a weak converse for the example of source coding for a DMS.

Question 2: Variable-length source coding (total: 10 points)

The Rotokas language is spoken by about 4000 people on an island in Papua New Guinea. It is one the languages with the fewest numbers of different sounds (phonemes) and thus we can write it with an alphabet of only 12 letters. A sample of the language is given on `http://en.wikipedia.org/wiki/Rotokas_language`:

> Vo tuariri rovoaia Pauto vuvuiua ora rasito pura-rovoreva. Vo osia rasito raga toureva, uva viapau oavu avuvai. Oire Pauto urauraaro tuepaepa aue ivaraia uukovi. Vara rutuia rupa toupaiva. Oa iava Pauto oisio puraroepa, Aviavia rorove. Oire aviavia rorova.

Consider now a DMS that picks letters from the Rotokas' alphabet (ignoring cases, spaces and special characters) with probabilities given by the frequencies observed in the above sample text. For your convenience these frequencies are given in the following table.

letter	A	E	I	O	U	G	K	P	R	S	T	V
frequency	0.23	0.04	0.11	0.13	0.12	0.005	0.005	0.05	0.12	0.02	0.05	0.12

Your job is to come up with a compression scheme for this source.

1. **(2 points)** Compute the entropy of this source.

2. **(4 points)** Design a Huffman code for this source.

3. **(4 points)** Design a Shannon code for this source.

Question 3: From variable to fixed-length source coding (total: 10 points)

We have quickly discussed in the lecture that it is possible to use a variable-length code to achieve rates down to the entropy in fixed-length block coding. In this question you will fill in the gaps we left open.

Consider a DMS $\mathbf{X}$ which produces sequences of letters from an alphabet $\mathcal{X}$ such that the the entropy of each letter is $H(X)$. Assume also that for any $k \in \mathbb{N}$, we have a Huffman code for k-tuples produced by this source. That is, a Huffman code that maps sequences $x^k \in \mathcal{X}^{\times k}$ of length k to codewords $C_k(x^k) \in \{0,1\}^*$ with lengths $\ell_k(x^k)$.

1. (**3 points**) Give an upper bound on $\mathbb{E}[\ell_k(X^k)]$ for X^k produced from this source in terms of the entropy $H(X)$. You might want to consider the case $k = 1$ first and recall properties of the Shannon code and its relation to the Huffman code.

2. (**4 points**) Use the weak law of large numbers to show that, for any $k \in \mathbb{N}$,
$$\lim_{m\to\infty} \Pr\left[\frac{1}{mk}\sum_{i=1}^{m} \ell_k(X_i^k) \geq H(X) + \frac{2}{k}\right] = 0\,. \tag{1}$$

3. (**4 points**) Now consider a rate $R = H(X) + \delta$ for any $\delta > 0$. Choose $k = \lceil \frac{2}{\delta} \rceil + 1$. For each $m \in \mathbb{N}$, design a block encoder for blocks of length $n = km$, that is, an encoder $e_n : X^n \to \left[2^{\lfloor Rn \rfloor}\right]$ and a decoder d_n creating an estimate $\hat{X}^n$ such that
$$\lim_{n\to\infty} \Pr\left[X^n \neq \hat{X}^n\right] = 0\,. \tag{2}$$

You have now shown that variable-length codes are also optimal also for block-coding.

END OF PAPER