PART II CODING AND CRYPTOGRAPHY
EXAMPLE SHEET 1

The first 11 examples are minimal to cover the course; you are also encouraged to try questions 12–16.

1 Consider Morse code[1]:

$$\begin{array}{rrr} A \mapsto \bullet - * & B \mapsto - \bullet \bullet \bullet * & C \mapsto - \bullet - \bullet * \\ D \mapsto - \bullet \bullet * & E \mapsto \bullet * & F \mapsto \bullet \bullet - \bullet * \\ O \mapsto - - - * & S \mapsto \bullet \bullet \bullet * & 7 \mapsto - - \bullet \bullet \bullet * \end{array}$$

Decode $- \bullet - \bullet * - - - * - \bullet \bullet * \bullet *$.

(ii) Consider the ASCII (American Standard Code for Information Interchange):

$$\begin{array}{rrr} A \mapsto 1000001 & B \mapsto 1000010 & C \mapsto 1000011 \\ a \mapsto 1100001 & b \mapsto 1100010 & c \mapsto 1100011 \\ + \mapsto 0101011 & ! \mapsto 0100001 & 7 \mapsto 0110111 \end{array}$$

Encode b7!. Decode 110001111000011100010.

2 Consider two alphabets $\mathcal{A}$ and $\mathcal{B}$ and a coding function $c : \mathcal{A} \to \mathcal{B}^*$.

(i) Explain, without using the notion of prefix-free codes, why, if c is injective and fixed length, c is decodable. Explain why, if c is injective and fixed length, c is prefix-free.

(ii) Let $\mathcal{A} = \mathcal{B} = \{0, 1\}$. If $c(0) = 0$, $c(1) = 00$ show that c is injective but c^* is not.

(iii) Let $\mathcal{A} = \{1, 2, 3, 4, 5, 6\}$ and $\mathcal{B} = \{0, 1\}$. Show that there is a variable length coding c such that c is injective and all codewords have length 2 or less. Show that there is no decodable coding c such that all codewords have length 2 or less.

3 (i) Give an example of a decodable code which is not prefix-free.

(ii) Give an example of a non-decodable code which satisfies Kraft's inequality.

(iii) A *comma code* (like Morse code) is one where a special letter – comma – occurs at the end of each codeword and nowhere else. Show that a comma code is prefix-free and give a direct argument to show that it must satisfy Kraft's inequality.

4 The product of two codes $c_j : \mathcal{A}_j \to \mathcal{B}_j^*$ is the code

$$g : \mathcal{A}_1 \times \mathcal{A}_2 \to (\mathcal{B}_1 \cup \mathcal{B}_2)^*$$

given by $g(a_1, a_2) = c_1(a_1)c_2(a_2)$.

Show that the product of two prefix-free codes is prefix-free, but the product of a decodable code and a prefix-free code need not even be decodable.

[1] After RMS *Titanic* hit an iceberg at 11.40 pm on 14 April 1912, the wireless operators Jack Philips and Harold Bride initially transmitted 'CQD-MGY, sinking, need immediate assistance,' later interspersed with the newer 'SOS' at the suggestion of Bride (CQD was still a widely understood distress signal at the time, and MGY was *Titanic*'s call sign). Morse code was used as an international standard for maritime communication until 1999, when it was replaced by the Global Maritime Distress Safety System. When the French Navy ceased using Morse code on 31 January 1997, the final message transmitted was "Calling all. This is our last cry before our eternal silence".

5 (i) Apply Huffman's algorithm to the nine messages M_j where M_j has probability $j/45$ for $1 \leq j \leq 9$.

(ii) Consider four messages with the following properties: M_1 has probability .23, M_2 has probability .24, M_3 has probability .26 and M_4 has probability .27. Show that any assignment of the codewords 00, 01, 10 and 11 produces a best code in the sense of this course.

6 Consider 64 messages M_j with the following properties: M_1 has probability 1/2, M_2 has probability 1/4 and M_j has probability 1/248 for $3 \leqslant j \leqslant 64$. Explain why, if we use codewords of equal length, then the length of a codeword must be at least 6. By using the ideas of Huffman's algorithm (you should not need to go through all the steps) obtain a set of codewords such that the *expected* length of a codeword sent is no more than 3.

7 (i) Let $\mathcal{A} = \{1, 2, 3, 4\}$. Suppose that the probability that letter k is chosen is $k/10$. Use your calculator to find $\lceil -\log_2 p_k \rceil$ and write down a Shannon–Fano code c.

(ii) Apply Huffman's algorithm to the four messages M_j, where M_j has probability $j/10$ for $1 \leqslant j \leqslant 4$. Denoting by c_H the Huffman code for this system, show that the entropy is approximately 1.85, that $\mathbb{E}|c(A)| = 2.4$ and that $\mathbb{E}|c_H(A)| = 1.9$. Check that these results are consistent with the appropriate results in the course.

8 Use the methods of Shannon-Fano and Huffman to construct prefix-free binary codes for messages $M_1, \dots, M_5$ emitted either (a) with equal probabilities, or (b) with probabilities 0.3, 0.3, 0.2, 0.15, 0.05. Compare the expected word lengths in each case.

9 Messages $M_1, \dots, M_5$ are emitted with probabilities 0.4, 0.2, 0.2, 0.1, 0.1. Find an optimal binary code. Determine whether there are optimal binary codes with (a) all but one codeword of the same length, or (b) each codeword a different length.

10 You are playing bridge with a partner and two opponents. The pack (of 52 cards) is dealt to provide 4 hands of 13 cards each. A simple representation for the hand you get would assign a unique 6-bit binary number to represent each card; then a 78-bit message represents your hand, a 156-bit message your pair's hands and a 312-bit message the whole deal. Let's try to do better. Assume all possible deals are equally likely. Show that there are 52!/(13!39!) different hands you might obtain. Show also that there are $52!/(13!)^4$ different deals.

(i) If the contents of a hand are conveyed by one player to their partner by a series of nods and shakes of the head how many movements of the head are required? Show that at least 40 movements are required. Give a simple code requiring 52 movements.

[You may assume for simplicity that the player to whom the information is being communicated does not look at her own cards. (In fact this does not make a difference since the two players do not acquire any shared information by looking at their own cards.)]

(ii) If instead the player uses the initial letters of words (say using the 16 most common letters), how many words will she need to utter?

11 Show that if an optimal binary code has word lengths s_1, s_2, $\dots s_m$ then

$$m \log_2 m \leqslant s_1 + s_2 + \cdots + s_m \leqslant (m^2 + m - 2)/2.$$

12 (i) It is known that exactly one member of the starship U.S.S. Emphasise has contracted the Macguffin virus. A test is available that will detect the virus at any dilution. However, the power required is such that the ship's force shields must be switched off for a minute during each test. Blood samples are taken from all crew members. The ship's computer has worked out that the probability of crew member number i harbouring the virus is p_i. (Thus the probability that the captain, who is, of course, number 1, has the disease is p_1.) Explain how, by testing pooled samples, the expected number of tests can be minimised. Write down the exact form of the test when there are 2^n crew members and $p_i = 2^{-n}$.

(ii) Questions like (i) are rather artificial, since they require that exactly one person carries the virus. Suppose that the probability that any member of a population of 2^n has a certain disease is p (and that the probability is independent of the health of the others) and there exists an error free test which can be carried out on pooled blood samples which indicates the presence of the disease in at least one of the samples or its absence from all.

Explain why there cannot be a testing scheme which can be guaranteed to require less than 2^n tests to diagnose all members of the population. How does the scheme suggested in the last sentence of (i) need to be modified to take account of the fact that more than one person may be ill (or, indeed, no one may be ill)? Show that the expected number of tests required by the modified scheme is no greater than $pn2^{n+1} + 1$. Explain why the cost of testing a large population of size x is no more than about $2pcx \log_2 x$ with c the cost of a test.

(iii) In practice, pooling schemes will be less complicated. Usually a group of x people are tested jointly and, if the joint test shows the disease, each is tested individually. Explain why this is not sensible if p is large but is sensible (with a reasonable choice of x) if p is small. If p is small, explain why there is an optimum value for x. Write down (but do not attempt to solve) an equation which indicates (in a 'mathematical methods' sense) that optimum value in terms of p, the probability that an individual has the disease.

Schemes like these are only worthwhile if the disease is rare and the test both is expensive and will work on pooled samples. However, these circumstances do occur together from time to time and the idea then produces public health benefits much more cheaply than would otherwise be possible.

13 In the binary case the Huffman code is defined by combining the two letters with smallest probabilities.

(i) Give the appropriate generalisation of Huffman's algorithm to an alphabet with a symbols when you have m messages and $m \equiv 1 \pmod{a-1}$.

(ii) Prove that your algorithm gives an optimal solution.

(iii) Extend the algorithm to cover general m by introducing messages of probability zero.

(iv) Carry this out for a ternary ($a = 3$) coding of a source with probabilities 0.2, 0.2, 0.15, 0.15, 0.1, 0.1, 0.05, 0.05.

(v) Show that if a ternary decodable code of size m achieves the lower bound in the Noiseless Coding Theorem then m is odd.

14 You are given m apparently identical coins, one of which may be a forgery. Counterfeit coins are either too light or too heavy. You have a balance, on which you may place any of the coins you like and determine whether the coins in one pan are together lighter than, heavier than or the same weight as those in the other. Using the balance you wish to detect whether there is a forgery and, if so, which coin it is and whether it is lighter or heavier.

Prove that, in any system of weighings which solves this problem, the maximum number of weighings involved cannot be less than $\log_3(2m+1)$.

[Optional] Show that for $m = 12$ three weighings suffice. [This problem 'is said to have been planted during the war ... by enemy agents since Operational Research spent so many man-hours on its solution.'[2]]

15 Extend the definition of entropy to a random variable X taking values in the non-negative integers. (You must allow for the possibility of infinite entropy.)

Compute the expected value $\mathbb{E}Y$ and entropy $H(Y)$ in the case when Y has the geometric distribution, that is to say $\mathbb{P}(Y = k) = p^k(1-p)$ $[0 < p < 1]$. Show that, amongst all random variables X taking values in the non-negative integers with the same expected value μ $[0 < \mu < \infty]$, the geometric distribution maximises the entropy.

16 Suppose that a source emits letters from the finite alphabet $\mathcal{A} = \{1, 2, \ldots, n\}$, each letter i occurring with (known) probability $p_i > 0$. Let S be the random codeword-length when the message is encoded by a decodable code $c : \mathcal{A} \to \mathcal{B}^*$ where $\mathcal{B}$ is an alphabet of k letters. It is desired to find a decodable code that minimizes the expected value of k^S. Establish the lower bound

$$\left(\sum_{i=1}^{n} \sqrt{p_i}\right)^2 \le \mathbb{E}(k^S)$$

and characterise when equality occurs. [Hint: Cauchy–Schwarz, $p_i^{1/2} = p_i^{1/2} k^{s_i/2} k^{-s_i/2}$.]

Prove that an optimal code for the above criterion must satisfy

$$\mathbb{E}(k^S) < k \left(\sum_{i=1}^{n} \sqrt{p_i}\right)^2 .$$

[Hint: Look for a code with codeword lengths $s_i = \lceil -\log_k p_i^{1/2}/\lambda \rceil$ for an appropriate λ.]

SM, Lent Term 2012
Comments on and corrections to this sheet may be emailed to `sm@dpmms.cam.ac.uk`

[2] The quotation is lifted from Pedoe's *The Gentle Art of Mathematics* which also gives an attractive solution. Niobe, the protagonist of Piers Anthony's novel *With a Tangled Skein*, must solve the twelve-coin variation of this puzzle to find her son in Hell: Satan has disguised the son to look identical to eleven other demons, and he is heavier or lighter depending on whether he is cursed to lie or able to speak truthfully.