**MATHEMATICAL TRIPOS PART II (2023–2024)**
**CODING AND CRYPTOGRAPHY**
**EXAMPLE SHEET 1 OF 4**

**1** (i) Give an example of a decipherable code which is not prefix-free. (Hint: What happens if you reverse all the codewords in a prefix-free code?)

(ii) Give an example of a non-decipherable code which satisfies the Kraft inequality.

(iii) Recall that a *comma code* is one where a special letter—the comma—occurs at the end of each codeword and nowhere else. Show that a comma code is prefix-free and check directly that comma codes satisfy the Kraft inequality.

**2** For a code $f : \Sigma_1 \to \Sigma_2^*$ and a code $f' : \Sigma_1' \to \Sigma_2'^*$ the *product code* is $g : \Sigma_1 \times \Sigma_1' \to (\Sigma_2 \cup \Sigma_2')^*$ given by $g(x, y) = f(x)f'(y)$. Show that the product of two prefix-free codes is prefix-free, but that the product of a decipherable code and a prefix-free code need not even be decipherable.

**3** Jensen's inequality states that if $f : \mathbb{R} \to \mathbb{R}$ is a convex function and $p_1, \ldots, p_n$ is a probability distribution (i.e. $0 \leqslant p_i \leqslant 1$ and $\sum p_i = 1$) then $f(\sum p_i x_i) \leq \sum p_i f(x_i)$ for any $x_1, \ldots, x_n \in \mathbb{R}$. Deduce Gibbs' inequality from Jensen's inequality applied to the convex function $f(x) = -\log x$.

**4** Show that $H(p_1, p_2, p_3) \leq H(p_1, 1 - p_1) + (1 - p_1)$ and determine when equality occurs.

**5** Use the methods of Shannon-Fano and Huffman to construct prefix-free binary codes for messages $\mu_1, \ldots, \mu_5$ emitted (i) with equal probabilities, or (ii) with probabilities $0.3, 0.3, 0.2, 0.15, 0.05$. Compare the expected word lengths in each case.

**6** Messages $\mu_1, \ldots, \mu_5$ are emitted with probabilities $0.4, 0.2, 0.2, 0.1, 0.1$. Determine whether there are optimal binary codings with (i) all but one codeword of the same length, or (ii) each codeword a different length.

**7** A binary Huffman code is used for encoding symbols $1, \ldots, m$ occurring with probabilities $p_1 \geq p_2 \geq \cdots \geq p_m > 0$ where $\sum_{1 \leq j \leq m} p_j = 1$. Let $s_1$ be the length of the shortest codeword and $s_m$ the length of the longest codeword. Determine the maximal and minimal values of $s_1$ and $s_m$ and find binary trees for which they are attained.

**8** Show that if an optimal binary code has word lengths $s_1, \ldots, s_m$ then

$$m \log m \leq s_1 + \cdots + s_m \leq (m^2 + m - 2)/2.$$

**9** 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 (binary) codewords of equal length, then the length of the 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.

**10** Suppose that a covid infection is known to originate in exactly one of $m$ rooms in College, the probability it originates in the $j^{th}$ being $p_j$. A health inspector has samples from all of the $m$ rooms and by testing the pooled samples from a set $A$ of them can determine with certainty whether the infection originates in $A$ or its complement. Let $N(p_1, \ldots, p_m)$ denote the minimum expected number of such tests needed to locate the infection. Show that $H(p_1, \ldots, p_m) \leq N(p_1, \ldots, p_m) < H(p_1, \ldots, p_m) + 1$, and determine when the lower bound is attained.

**11** A source emits messages $\mu_1, \ldots, \mu_m$ with non-zero probabilities $p_1, \ldots, p_m$. Let $S$ be the codeword length random variable for a decipherable code $f : \Sigma_1 \to \Sigma_2^*$ where $\Sigma_1 = \{\mu_1, \ldots, \mu_m\}$ and $|\Sigma_2| = a$. Show that the minimum possible value of $E(a^S)$ satisfies

$$\left(\sum_{i=1}^{m} \sqrt{p_i}\right)^2 \leq E(a^S) < a\left(\sum_{i=1}^{m} \sqrt{p_i}\right)^2.$$

(Hint: for the left-hand inequality consider the Cauchy-Schwarz inequality. For the right-hand inequality look for a code with codeword lengths $\ell_i = \lceil -\log_a(p_i^{1/2}/\lambda) \rceil$ for an appropriate $\lambda$.)

**12** (i) In lectures we only described Huffman coding in the binary case, *i.e.* $a = 2$. In general we add extra messages of probability zero so that the number of messages $m$ satisfies $m \equiv 1 \pmod{a-1}$. Then at each stage we group together the $a$ smallest probabilities. Carry this out for a ternary coding of a source with probabilities $0.2, 0.2, 0.15, 0.15, 0.1, 0.1, 0.05, 0.05$.

(ii) Show that if a ternary decipherable code of size $m$ meets the lower bound in the noiseless coding theorem then $m$ is odd.

**Further Problems**

**13** A balance puzzle: you are given $m$ apparently identical coins, one of which may be a forgery. Forged coins are either too light or too heavy. You are also given a balance, on which you may place any of the coins you like. The coins placed in either pan may be together heavier or lighter than those in the other pan or the pans may balance.

You are allowed at most 3 uses of the balance. Show that if $m > 13$ then you cannot be sure of detecting the forgery and its nature. [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.'[1]]

**14** In an unreleased episode of *The Queen's Gambit*, there is a game on a standard chessboard in which one player (Beth) has to guess where her opponent has placed the queen. Beth is allowed six questions which must be answered truthfully by a yes/no reply. Prove that there is a strategy by which Beth can always win this game, but that she cannot ensure winning if she is allowed only five questions.

In the forthcoming Netflix sequel, the game is played on an $n \times n$ chessboard. How many questions does Beth need in order to be certain of winning?

[1]The quotation is lifted from Dan Pedoe's *The Gentle Art of Mathematics* (Dover reprint, 1982) 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. In the episode 'Captain Peralta' of *Brooklyn Nine-Nine*, Holt presents to his team a version of the twelve-coin problem involving twelve men and a seesaw. The original 12 coin version was solved in 1945 by H. Grossman.

**15** Consider the following method for generating a code for a random variable $X$ which takes $m$ values $\{1, 2, \ldots, m\}$ with probabilities $p_1, p_2, \ldots, p_m$. Assume that the probabilities are ordered so that $p_1 \geqslant p_2 \geqslant \cdots \geqslant p_m$. Define

$$F_i = \sum_{k=1}^{i-1} p_k,$$

i.e. for the sum of the probabilities of all symbols less than $i$. Then the codeword for $i$ is the number $F_i \in [0, 1]$ rounded off to $\ell_i$ bits, where $\ell_i = \lceil \log \frac{1}{p_i} \rceil$.

(i) Show that the code constructed by this process is prefix-free and the expected word length $L$ satisfies

$$H(X) \leqslant L < H(X) + 1.$$

(ii) Construct the code for the probability distribution $(0.5, 0.25, 0.125, 0.125)$.

(This is called a *Shannon code*. It is suboptimal in the sense that it does not in general achieve the lowest possible expected codeword length like Huffman coding does.)

SM, Lent Term 2024
Comments on and corrections to this sheet may be emailed to `sm137@cam.ac.uk`