UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFORMATION THEORY

Saturday 15th December 2012

09:30 to 11:30

MSc Courses

Convener: B. Franke
External Examiners: T. Attwood, R. Connor, R. Cooper, S. Denham

INSTRUCTIONS TO CANDIDATES

Answer QUESTION 1 and ONE other question.

Question 1 is COMPULSORY.

All questions carry equal weight.

CALCULATORS MAY NOT BE USED IN THIS EXAMINATION

1. **You MUST answer this question.**

   (a) An ensemble X, has outcomes drawn from a discrete alphabet, $x \in \mathcal{A}_X = \{a_1, a_2, \ldots, a_I\}$, with corresponding probabilities $\mathcal{P}_X = \{p_1, p_2, \ldots, p_I\}$. The variance of the information content of outcomes from the ensemble is σ^2.

      i. Write an expression for the information content of an outcome $h(x = a_i)$. [*1 mark*]

      ii. Write an expression for the entropy of the ensemble, $H(X)$. [*1 mark*]

      iii. Explain what is meant by the 'extended ensemble', X^N. Express the mean and variance of $h(\mathbf{x})$, the information content of an extended outcome, in terms of $H(X)$ and σ^2. Also give the mean and variance of $h(\mathbf{x})/N$, the information content per symbol in an extended outcome. [*5 marks*]

   (b) The ban is the unit of information that measures log-probability using base-10 logarithms. We wish to express an information content of x bans as y bits. Write an expression for y in terms of x, using only the standard arithmetic operations `+`, `-`, `*`, `/`, and a function to evaluate natural logarithms, $\log_e$. [*2 marks*]

   (c) Construct a Huffman code for an ensemble X, with symbols $\mathcal{A}_X = \{\texttt{a}, \texttt{b}, \texttt{c}, \texttt{d}, \texttt{e}, \texttt{f}\}$ and corresponding probabilities $\mathcal{P}_X = \{0.1, 0.1, 0.1, 0.1, 0.3, 0.3\}$. Compute the expected length of the encoding in bits per symbol. Hence write down upper and lower bounds for the Entropy of this ensemble. [*5 marks*]

   (d) The Huffman algorithm generates codewords with lengths $\{\ell_i\}$. Give an ensemble for which this symbol code is an optimal compressor. [*2 marks*]

   (e) State an ensemble that is not compressed well by a Huffman code. Explain why Huffman coding does not work well, and describe a better method, other than arithmetic coding, to compress many outcomes from your ensemble. [*4 marks*]

   (f) A set of positive values $\{x_1, x_2, \ldots, x_N\}$ has arithmetic mean A, and geometric mean G, where by definition:

$$A = \frac{1}{N}\sum_{n=1}^{N} x_n\,, \qquad \text{and} \qquad G = \exp\left(\frac{1}{N}\sum_{n=1}^{N} \log_e x_n\right).$$

   Prove that, for all sets of positive values, either $G \leq A$ or $G \geq A$. State for which sets the geometric mean is equal to the arithmetic mean.

   Hint: you may wish to rewrite the geometric mean as a function of an expectation. [*5 marks*]

2. (a) Three probabilistic models are used, in conjunction with an arithmetic coder, to compress a megabyte of English text. In context c, a model predicts character a_i, which belongs to alphabet $\mathcal{A}_X$, with probability

$$P(a_i \,|\, c) = \frac{n_c(a_i) + \alpha}{N_c + \alpha|\mathcal{A}_X|},$$

where $n_c(a_i)$ gives the number of times the character has previously been observed in this context, out of the $N_c = \sum_i n_c(a_i)$ times that this context has previously been used for predictions. The contexts used by the different models are: 1) empty, 2) the previous character, 3) the previous 10 characters.

i. Why must the frequencies be smoothed with $\alpha > 0$? [*2 marks*]

ii. Explain how the parameter α affects predictions, and how it relates to prior knowledge about the distributions for different contexts. [*4 marks*]

iii. Explain in detail why model 2 compresses the file better than model 1, but model 3 compresses the file less well than either of the other models. (The setting of α is not critical, but you may assume $\alpha = 0.1$.)

Briefly outline a way that long contexts could be used to obtain better predictions. [*8 marks*]

The remainder of this question considers a storage device that represents files using sequences of "trits", symbols in $\{\texttt{0}, \texttt{1}, \texttt{2}\}$. Files are represented on the device using a ternary symbol code: each symbol from the file is mapped to a codeword made from trits. These codewords are concatenated together into a single sequence on the device, without any explicit marks showing where one symbol ends and another begins.

(b) Each source symbol $x \in \mathcal{A}_X$ is encoded into a codeword of length $l(x)$ trits. Show that a uniquely decodable ternary symbol code must satisfy

$$\sum_{x \in \mathcal{A}_X} 3^{-l(x)} \leq 1.$$

Hint: one approach involves constructing a probability distribution implied by the codewords, normalizing it, and comparing the expected length of a codeword to the entropy (both measured in trits). Regardless of your approach, explain each step of your argument carefully. [*5 marks*]

(c) The storage device is used to store binary files that have been compressed as well as possible. Explain why a reasonable model for the source files is that each bit has an independent probability of $1/2$ of being a $\texttt{0}$ or $\texttt{1}$. [*2 marks*]

(d) Let X be the ensemble resulting from drawing a single bit $x \in \{\texttt{0}, \texttt{1}\}$ with even probabilities, $\mathcal{P}_X = \{1/2, 1/2\}$. Construct instantaneously decodable, *ternary* symbol codes for the extended ensembles X^2 and X^3. For each code, compute the expected number of trits used per source bit. [*4 marks*]

3. (a) Consider the following two joint distributions over two variables x and y:

$P_1(x,y)$	$x=1$	$x=2$
$y=1$	0.4	0
$y=2$	0	0.6

$P_2(x,y)$	$x=1$	$x=2$
$y=1$	0.04	0.06
$y=2$	0.36	0.54

i. For each distribution, P_1 and P_2, show whether x and y are independent. [*4 marks*]

ii. One of these joint distributions is chosen at random with probability $\frac{1}{2}$ each. A joint (x,y) pair is sampled from the chosen distribution, and the value $x=1$ is observed. What is the probability that the other value in the pair is $y=1$? [*4 marks*]

A channel takes 4 bits of input (half a byte, or a 'nibble') at a time. Exactly one of these bits is always flipped before being received. The corrupted bit is chosen uniformly at random.

(b) Carefully explain why the symmetry of the channel means that the input distribution to the channel can be set to uniform when maximizing the mutual information between the input and output. [*3 marks*]

(c) Show that the capacity of this channel is 2 bits. [*3 marks*]

Now assume that each input bit is flipped independently with probability $1/4$, so that each nibble still has one bit flipped on average. The optimal input distribution is still uniform.

(d) Show that the capacity of this channel is $4(1-H_2(1/4))$ bits, where $H_2(p) = x\log_2\frac{1}{x} + (1-x)\log_2\frac{1}{1-x}$ is the binary entropy function. [*4 marks*]

Consider a Binary Symmetric Channel (BSC), which transmits binary symbols in $\{\mathtt{0},\mathtt{1}\}$, and changes the transmitted bit with flip probability $f=\frac{1}{5}$. The repetition R_{101} code copies each bit of a binary source file 101 times before transmission, and decodes the bit by majority vote.

(e) When a source bit is a `1`, 101 `1`s are sent over the channel. Sketch the probability distribution over the number of `1`s received after the block has been corrupted by noise. Give an approximate indication of the location and width of the distribution, and label your axes. [*3 marks*]

(f) The R_{101} code is an example of an $[N,K]$ block code.

i. How many codewords, S, does it use? [*1 mark*]

ii. How many bits, K, does it communicate per block? [*1 mark*]

iii. A more sophisticated $N=101$ block code has a rate five times larger than that of the R_{101} code. How many codewords does the code have? [*1 mark*]

iv. Briefly explain whether another $N=101$ block code could have a better bit-error probability than the R_{101} code, when used over a Binary Symmetric Channel. [*1 mark*]