CSC 310S
Spring, 2003

March 10, 2003
DUE: April 2 / 4, 2003

# Information Theory
# Assignment #3
**[FINAL(!!) Release]**

[Note: There are four questions in this assignment, but your basic mark for the assignment (out of 60) will be based only on the three questions that you answer best. If you answer all four, then your fourth best answer will give you "extra credit" in the course (between 1% and 5% added to your final mark).]

## 1. Lempel-Ziv-Welch Coding [20 marks]

Consider using Lempel-Ziv-Welch for encoding *binary strings*, so that the source alphabet is $\{0, 1\}$. Use knowledge of the relationships among (i) $T$, the length of text to be encoded, (ii) $c(T)$, the number of codes transmitted, (iii) the number of dictionary entries at the end of the transmission, and (iv) the sum lengths of all the strings in the dictionary entries to answer the following questions. (In all parts, assume that source symbols are coded in 8 bits each, so that the compression ratio is: $\frac{B \times c(T)}{8 \times T}$)

A. If the LZW algorithm is applied to a source stream of length $T$ that consists only of *one* bits:

"11111111111111111111111 ... 1111111111111"

(i). Obtain an *approximate* expression for the value of $c(T)$ in terms of $T$.

(ii). If the dictionary codes are $B$ bits long, find an *approximate* expression for $R$, the ratio of the length of the encoded stream to the length of the original stream.

(iii). Evaluate the expressions obtained in parts (i) and (ii) for the three indicated pairs of values for $B$ and $T$:

- ($B = 8$ , $T = 1{,}000$)
- ($B = 12$ , $T = 1{,}000{,}000$)
- ($B = 16$ , $T = 1{,}000{,}000{,}000$)

B. Perform all the same calculations as in part A, but this time assuming that the source string of length $T$ consists entirely of *strictly* alternating ones and zeroes:

"10101010101010101010101 ... 101010101010"

C. Finally, consider the case where the source string is a truly random bit string with each bit being independent of previous bits, and being equally likely to be either a "0" or a "1". Produce expressions for rough (but justifiable) upper bounds on $c(T)$ and $R$ for this case, and use them to express bounds on the values of $c(T)$ and $R$ for the following two cases:

- ($B = 8$ , $T = 1{,}000$)
- ($B = 16$ , $T = 1{,}000{,}000$)

**2. Information Channels with Noise [20 marks]**

Consider an Information Channel (with noise) which is a generalization of the Binary Symmetric Channel (BSC) and the Binary Erasure Channel (BEC): The input alphabet is {0, 1} with input distribution $\{ p_0, p_1 \}$, and the output alphabet is {0, 1, ?}. When a bit (either zero or one) is transmitted, it is correctly received with probability $P$. When it is not received correctly, it is flipped (from 0 to 1 or 1 to 0) with probability $f(1-P)$, and it is erased (becoming "?") with probability $(1-f)(1-P)$. In order to obtain numerical answers to the questions below, assume that $p_0 = .6$, $P = .7$ and $f = 1/3$.

A. What is the entropy of the input distribution, $H(A)$?

B. Show the *channel matrix*, $P_{ij}$, for the information channel, and calculate the entropy of the symbol received given the symbol transmitted, $H(B \mid A)$.

C. Show the joint probability matrix, $R_{ij}$, and calculate the entropy of the joint distribution, $H(A, B)$.

D. What is the output distribution, $(q_0, q_1, q_?)$, and what is its entropy, $H(B)$?

E. Show the backward probability matrix, $Q_{ij}$, and calculate the entropy of the backward probabilities, $H(A \mid B)$.

F. What is the mutual information between the symbols sent and the symbols received, $I(A, B)$?

G. By considering other values for $p_0$ (and $p_1$) but keeping $P = .7$ and $f = 1/3$ fixed, give an estimate of the *capacity*, $C$, of the information channel, and indicate the approximate value of $p_0$ for which the channel capacity is attained. (It may be helpful to try to sketch a graph of $I(A, B)$ versus $p_0$ when $P = .7$ and $f = 1/3$.)

**3. Error Correcting Codes [20 marks]**

Investigate three *Error Correcting Codes [n,k]*, characterized by the parameters $k$ (where $2^k$ is the number of codewords), and $n$ (the length of each codeword), for pairs [4,1], [5,2], and [6,3].

A. For each of the three *generator matrices* below composed of $k$ codewords of length $k+3$, generate a full code of $2^k$ codewords.

| [4,1] code | [5,2] code | [6,3] code |
|---|---|---|
| 1111 | 10101<br>01110 | 010101<br>100110<br>111000 |

B. How do these three codes compare to the [7,4] Hamming code with respect to rate of transmission?

C. For each of the three codes, give the distribution of pairwise distances between all code word pairs. (This distribution consists of $2^{k-1}(2^k - 1)$ values.)

D. Based on the answer to part C, comment on how these three codes compare to the [7,4] Hamming code with respect to ability to correct and/or detect various numbers of erroneous bits.

E. In the [7,4] Hamming code, the *syndrome* can be used to identify the specific bit out of the seven that was most likely transmitted erroneously. Specify an appropriate parity check matrix, $H$, that will yeild the possibility of syndrome decoding for the [6,3] code given above.

## 4. Encoding of Real Numbers [20 marks]

A sequence of $n$ real numbers, $x_1, x_2, \cdots, x_n$, is composed of *independent* selections from the distribution $f(x)$ defined below. Consider encodings of the sequence of numbers in which each real number is encoded as a bit string of length $b$, and the $i^{th}$ bit string (i.e., $i-1$ in binary) corresponds to real value $r_i$. With *distortion* defined as:

$$distortion = \frac{1}{n} \sum_{j=1}^{n} \min_{i} \; | x_j - r_i |$$

devise coding schemes by selecting the $r_i$ values to minimize distortion for each of the cases below. The function $f(x)$ is a "trapezoidal" distribution of real numbers:

$$f(x) = \begin{cases} \dfrac{\beta+x}{\beta-\alpha} \left[ \dfrac{1}{\beta+\alpha} \right] & -\beta \le x \le -\alpha \\ \left[ \dfrac{1}{\beta+\alpha} \right] & -\alpha \le x \le +\alpha \\ \dfrac{\beta-x}{\beta-\alpha} \left[ \dfrac{1}{\beta+\alpha} \right] & +\alpha \le x \le +\beta \end{cases}$$

Let $\beta = 60$ always, and let $\alpha = 0$ or 30 or 60 to form "isoceles triangular" or "trapezoidal" or "uniform" distributions, respectively.

A. Give an expression for the *differential entropy* of this distribution, and evaluate the expression for the case of $\alpha = \beta = 60$ (i.e., the uniform distribution).

B. For each of the three values of $\alpha$, find boundaries $(b_0, b_1, b_2)$ and values $(r_1, r_2)$ so that distortion is minimized when a *one-bit* code is used for the real numbers. What is the distortion with your chosen values?

C. For each of the three values of $\alpha$, find boundaries $(b_0, b_1, b_2, b_3, b_4)$ and values $(r_1, r_2, r_3, r_4)$ so that distortion is minimized when a *two-bit* code is used for the real numbers. What is the distortion with your chosen values?

D. Assess the benefits of using *vector quantization* (with vectors of length two) rather than scalar quantization for input values from distribution f(x) with an encoding rate of one bit per input value as in part (B). Each *pair* of values from the input steam, $(x_k, x_{k+1})$, is encoded in two bits by associating the pair with one of four *representatives*, $((r_{x1}, r_{y1}), (r_{x2}, r_{y2}), (r_{x3}, r_{y3}), (r_{x4}, r_{y4}))$. Then the contribution to distortion of this pair is:

$$\frac{1}{2} \min_{i=1,2,3,4} \left( | x_k - r_{xi} | + | x_{k+1} - r_{yi} | \right)$$

For any of the three values of $\alpha$, can distortion be made lower than in part (B)? Discuss briefly to support your answer.

[Note: To produce answers for parts (B), (C) and (D), you may use any combination of analysis, heuristic search or algorithmic search. Just indicate the method(s) you use.]