Information Theory: Principles and Applications
Homework 3 - Due: April 23, 2010

1. The following problem concerns a technique known as run-length coding. Suppose $X_1, X_2, \ldots$ is a sequence of random source symbols with $p_X(a) = 0.9$ and $p_X(b) = 0.1$. We encode this source by a variable-length-to-variable-length coding technique known as run-length coding. The source output is first mapped into intermediate digits by counting the number of occurrences of a between each b. Thus, an intermediate digit occurs on each occurrence of the symbol b. However, since we do not want the intermediate digits to get too large, the intermediate digit 8 is used on the eighth consecutive a, and the counting restarts at this point. Thus, outputs appear on each b and on each eighth a. For example, the two lines below illustrate a string of source symbols and the corresponding intermediate digits

b	a	a	a	b	a	a	a	a	a	a	a	a	a	a	b	b	a	a	a	a	b
0				3								8			2	0					4
0000				0011								1			0010	0000					0010

The final stage of encoding assigns the codeword 1 to the intermediate integer 8, and assigns a 4 bit codeword consisting of 0 followed by the 3 bit binary representation for each integer 0 to 7. This is illustrated in the third line above.

   (a) Show why the overall code is uniquely decodable.

   (b) Find the average number n_1 of source symbols per intermediate digits.

   (c) Find the average number n_2 of output bits per intermediate digits.

   (d) Show, by appeal to the law of the large numbers, that for a very long sequence of source symbols, the ratio of the number of encoded bits to the number of source symbols will, with high probability, be close to n_2/n_1. Compare this ratio to the average number of code letters per source letter for a Huffman code encoding 4 source digits at a time.

2. Lempel-Ziv LZ78

   (a) Give the Lempel-Ziv parsing and encoding of 000000110101000000110101.

   (b) Decode the following sequence encoded by the LZ78 algorithm

$$00101011101100100100011010101000011.$$

3. Consider a binary, stationary, Markov source described by $P(X_{k+1} = 0|X_k = 0) = P(X_{k+1} = 1|X_k = 1) = \alpha$ where $0 < \alpha < 1$.

   (a) Find the entropy rate of this source.

   Given a sequence $X_1, X_2, \ldots$ we can think of it as an alternating series of repetitions. For example if $X_1, X_2, \ldots, = 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, \ldots$ it can be thought of as 0 repeated 3 times, 1 repeated 2 times, 0 repeated 1 time, 1 repeated 5 times, etc. Let $R_1, R_2, \ldots$ be the lengths of these repetitions. In the example above, these are $3, 2, 1, 5, \ldots$

   (b) Argue that $R_1, R_2, \ldots$ form an i.i.d. sequence.

   (c) Find the probability distribution of R_k.

   (d) Find the expectation $E[R_1]$, and the entropy $H(R_1)$.

   (e) The sequence $X_1, X_2, \ldots$ can be described by first describing X_1 using 1 bit and then describing $R_1, R_2, \ldots$. Suppose the sequence $R_1, R_2, \ldots$ is effciently encoded into $H(R)$ bits per symbol. How many bits per symbol does this method use in encoding the sequence $X_1, X_2, \ldots$? How does this compare to $H(\mathcal{X})$ found in (a)?

4. The binary sequence

$$s = 1111111110000001111111111111111100001 = 1^9 0^6 1^{16} 0^4 1$$

   was generated by a stationary two state Markov chain with transition probabilities $P(X_{i+1} = 1|X_i = 0) = 2P(X_{i+1} = 0|X_i = 1) = 0.2$.

   Encode s using:

   (a) a Huffman code for 3-bit symbols based on the source model.

   (b) a Huffman code for 3-bit symbols based on relative frequencies in s.

   (c) a Shannon-Fano-Elias code for 3-bit symbols based on the source model.

   (d) a Shannon-Fano-Elias code for 3-bit symbols based on relative frequencies in s.

(e) The LZ78 algorithm.

(f) Relate your answers to the entropy rate of the Markov source and the entropy of s based on relative frequencies.

5. The output of a discrete memoryless channel K_1 is connected to the input of another discrete memoryless channel K_2. Show that the capacity of the cascade combination can never exceed the capacity of K_i, $i = 1, 2$. ("Information cannot be increased by data processing").

6. Consider the discrete memoryless channel $Y = X + Z \mod 13$, where $P(Z = 1) = P(Z = 2) = P(Z = 3) = 1/3$, and $X \in \{0, 1, ..., 12\}$. Assume that Z is independent of X.

   (a) Find the capacity.

   (b) What is the maximizing input distribution $p_X^*(x)$?

7. The Z-channel has binary input and output alphabets and transition probabilities $P(Y|X)$ given by $P(0|0) = 1$ and $P(0|1) = \epsilon$. Find the capacity of the Z-channel and the maximizing input probability distribution in terms of ϵ.

8. (Optional) The binary errors-and-erasures channel is given by

$$P(Y|X) = \begin{bmatrix} 1-p-\alpha & \alpha & p \\ p & \alpha & 1-p-\alpha \end{bmatrix}$$

   (a) Find the capacity.

   (b) Specialize to erasures only ($p = 0$).

   (c) Specialize to the binary symmetric channel ($\alpha = 0$).

9. (Optional) Show that for a weakly symmetric channel

$$C = \log |\mathcal{Y}| - H(\text{row of transition matrix})$$

   and is achieved by a uniform distribution on the input alphabet.

Useful formula

$$\sum_{n=0}^{\infty} nr^n = \frac{r}{(1-r)^2}, |r| < 1$$