Information Theory (H.-M. Hang; 2009/10/5)

(Cover and Thomas, Chaps 3 and 5)

Homework #2

Due Date: October 20, 2009

(1)

7. **The AEP and source coding.** A discrete memoryless source emits a sequence of statistically independent binary digits with probabilities $p(1) = 0.005$ and $p(0) = 0.995$. The digits are taken 100 at a time and a binary codeword is provided for every sequence of 100 digits containing three or fewer ones.

   (a) Assuming that all codewords are the same length, find the minimum length required to provide codewords for all sequences with three or fewer ones.

   (b) Calculate the probability of observing a source sequence for which no codeword has been assigned.

   (c) Use Chebyshev's inequality to bound the probability of observing a source sequence for which no codeword has been assigned. Compare this bound with the actual probability computed in part (b).

(2)

13. **Calculation of typical set** To clarify the notion of a typical set $A_\epsilon^{(n)}$ and the smallest set of high probability $B_\delta^{(n)}$, we will calculate the set for a simple example. Consider a sequence of i.i.d. binary random variables, $X_1, X_2, \ldots, X_n$, where the probability that $X_i = 1$ is 0.6 (and therefore the probability that $X_i = 0$ is 0.4).

   (a) Calculate $H(X)$.

   (b) With $n = 25$ and $\epsilon = 0.1$, which sequences fall in the typical set $A_\epsilon^{(n)}$? What is the probability of the typical set? How many elements are there in the typical set? (This involves computation of a table of probabilities for sequences with k 1's, $0 \le k \le 25$, and finding those sequences that are in the typical set.)

k	(n k)	prob.	entropy	cumulative prob.
0	1	0	1.321928	0
1	25	0	1.29853	0
2	300	0	1.275131	0
3	2300	0.000001	1.251733	0.000001
4	12650	0.000007	1.228334	0.000008
5	53130	0.000045	1.204936	0.000054
6	177100	0.000227	1.181537	0.000281
7	480700	0.000925	1.158139	0.001205
8	1081575	0.003121	1.13474	0.004326
9	2042975	0.008843	1.111342	0.013169
10	3268760	0.021222	1.087943	0.034392
11	4457400	0.04341	1.064545	0.077801
12	5200300	0.075967	1.041146	0.153768
13	5200300	0.11395	1.017748	0.267718
14	4457400	0.146507	0.994349	0.414225
15	3268760	0.161158	0.970951	0.575383
16	2042975	0.151086	0.947552	0.726469
17	1081575	0.11998	0.924154	0.846448
18	480700	0.079986	0.900755	0.926435
19	177100	0.044203	0.877357	0.970638
20	53130	0.019891	0.853958	0.990529
21	12650	0.007104	0.83056	0.997633
22	2300	0.001937	0.807161	0.999571
23	300	0.000379	0.783763	0.99995
24	25	0.000047	0.760364	0.999997
25	1	0.000003	0.736966	1

(3)

4. **Huffman coding.** Consider the random variable

$$X = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ 0.49 & 0.26 & 0.12 & 0.04 & 0.04 & 0.03 & 0.02 \end{pmatrix}$$

(a) Find a binary Huffman code for X.

(b) Find the expected codelength for this encoding.

(c) Find a ternary Huffman code for X.

(4)

12. **Shannon codes and Huffman codes.** Consider a random variable X which takes on four values with probabilities $(\frac{1}{3}, \frac{1}{3}, \frac{1}{4}, \frac{1}{12})$.

(a) Construct a Huffman code for this random variable.

(b) Show that there exist two different sets of optimal lengths for the codewords, namely, show that codeword length assignments $(1,2,3,3)$ and $(2,2,2,2)$ are both optimal.

(c) Conclude that there are optimal codes with codeword lengths for some symbols that exceed the Shannon code length $\lceil \log \frac{1}{p(x)} \rceil$.

(5)

30. **Relative entropy is cost of miscoding:** Let the random variable X have five possible outcomes $\{1,2,3,4,5\}$. Consider two distributions $p(x)$ and $q(x)$ on this random variable

Symbol	$p(x)$	$q(x)$	$C_1(x)$	$C_2(x)$
1	1/2	1/2	0	0
2	1/4	1/8	10	100
3	1/8	1/8	110	101
4	1/16	1/8	1110	110
5	1/16	1/8	1111	111

(a) Calculate $H(p)$, $H(q)$, $D(p||q)$ and $D(q||p)$.

(b) The last two columns above represent codes for the random variable. Verify that the average length of C_1 under p is equal to the entropy $H(p)$. Thus C_1 is optimal for p. Verify that C_2 is optimal for q.

(c) Now assume that we use code C_2 when the distribution is p. What is the average length of the codewords. By how much does it exceed the entropy p?

(d) What is the loss if we use code C_1 when the distribution is q?

(6)

37. **Codes.** Which of the following codes are

(a) uniquely decodable?

(b) instantaneous?

$$\begin{aligned} C_1 &= \{00, 01, 0\} \\ C_2 &= \{00, 01, 100, 101, 11\} \\ C_3 &= \{0, 10, 110, 1110, \ldots\} \\ C_4 &= \{0, 00, 000, 0000\} \end{aligned}$$

--- The End ---