CSC 310S
Spring, 2003

February 3, 2003
DUE: March 5 / 7, 2003

Information Theory
Assignment #2

1. Instantaneous Code Construction [20 marks]

Assume that a source produces independent symbols with relative frequencies:

(.41, .16, .15, .14, .08, .03, .02, .01)

(a). Devise *binary* (base 2) codes for this source using each of the three algorithms:

- Shannon Construction algorithm
- Shannon-Fano algorithm
- Huffman's algorithm

Then calculate and compare the average code word lengths and the relative redundancies of the three codes.

(b). For the same set of source symbol probabilities, do the same as in part (a), but this time producing *ternary* (base 3) codes by each of the three algorithms.

[NOTE: Produce the codes prescribed by the algorithms. Don't go on to do "obvious optimizations" to the resulting codes.]

		base 2 codes			base 3 codes		
symbol	prob	Shan	Sh-Fa	Huff	Shan	Sh-Fa	Huff
a	.41	00	00	0	0	0	0
b	.16	011	01	100	10	10	1
c	.15	100	10	101	12	11	20
d	.14	101	110	110	20	20	21
e	.08	1101	1110	1110	212	21	220
f	.03	111100	11110	11110	2211	220	221
g	.02	111110	111110	111110	2220	221	2220
h	.01	1111110	111111	111111	22220	222	2221
$L_D(X) =$		2.86	2.51	2.41	1.80	1.65	1.60
$H_D(X) =$		2.38065	2.38065	2.38065	1.50202	1.50202	1.50202
Rel.Redund. =		16.761%	5.154%	1.218%	16.554%	8.968%	6.124%

Comments: For the base 2 codes, there are substantial differences in the relative redundancies, with the Shannon Construction leading to more than 12 times higher relative redundancy than with the Huffman code. For the base 3 codes, the same ratio is less than a factor of 3. In both cases, the Shannon Construction code has a relative redundancy over 15%, while the Huffman code relative redundancy is about 1% in base 2, but about 6% in base 3.

2. Arithmetic Coding [20 marks]

Either by hand or using a program, carry out the tasks below involving *Arithmetic Coding*. (Show your intermediate work by writing down, or printing out from your program, the intermediate values obtained at each stage of the computations.) You may assume that the strings to be encoded and decoded require at most nine decimal places of accuracy in their arithmetic encoded representation (so regular floating point representations can be used; there is no need to employ shifting of bit strings).

Assume a source alphabet of four symbols (A, B, C, and D) with respective *independent* probabilities of occurrence (.4, .3, .2, and .1), associating A with the range [0.0,0.4), D with the range [0.9,1.0), etc.

(a). i. If the string to be encoded starts out with "AAAAA", what is the range of real numbers in which the arithmetic encoding of that string must fall?

Since $(.4)^5 = .0102$, the range is: [0.0, .0102).

ii. And what if the string to be encoded starts out with "DDDDD"?

Since $(.1)^5 = .00001$, and $1.0 - .00001 = .99999$,
the range is: [.99999,1.0).

(b). i. If the *binary represenation* of the number to be decoded starts out with "00000" in the first five places, how many initial characters of the decoded string are known with certainty?

A binary fraction starting with "00000" must be less than $1/32 = .03125$. The range corresponding to "AAA..." is [0.0, .064), which covers all numbers up to $1/32$, but the range for "AAAA..." is [0.0, .0256), which does not. Therefore, "00000..." guarantees that the first three symbols are "A", but the fourth one may not be an "A".

ii. And what if the number to be decoded starts out with "11111"?

A binary fraction starting with "11111" must be at least $31/32 = .96875$. The range corresponding to "D..." is [.9, 1.0), which covers all numbers above $31/32$, but the range for "DD..." is [0.99, 1.0), which does not. Therefore, "11111..." guarantees that the first symbol is "D", but the second one may not be an "D".

(c). Encode the string "ABACADABA" producing a real number in the interval [0,1).

Strings that start with "ABACADABA" correspond to the range [0.197117440, 0.197135872).

(d). Decode the value .929576450 to produce a string of length nine composed of the four symbols.

The string encoded by the number 0.929576450 starts with the sequence "DACABACAD".

3. Calculating Entropy Rates [20 marks]

Consider sources that emit the three symbols, X, Y, and Z. Such sources can be modeled either as *independent*, giving just the relative frequencies of the three symbols (p_X, p_Y, p_Z), or as *Markovian*, giving the first order transition probabilities between each symbol and the next one (p_{IJ} where I and J are any combination of X, Y, and Z).

For each of the four cases below, provide each of the following answers:

(a). What are the equilibrium frequencies with which each of the three symbols occurs, p_X, p_Y, and p_Z ?

(b). What is the *entropy* of the source based on the *independent model* ?

(c). What is the *entropy rate* of the source based on the transition probabilities of the *Markov model* of the source ?

Case (1):

from	to X	to Y	to Z
X	.333..	.333..	.333..
Y	.333..	.333..	.333..
Z	.333..	.333..	.333..

(a). The probabilities are: ($p_X = p_Y = p_Z = 1/3$).

(b). The entropy of the source based on the *independent* model is $H(X_n) = H(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) = 1.585$.

(c). The entropy of the source based on the *1st order Markov* model is $H(X_n \mid X_{n-1}) = 3 \times \frac{1}{3} \times H(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) = 1.585$.

Case (2):

from	to X	to Y	to Z
X	.1	.9	.0
Y	.1	.0	.9
Z	.0	.1	.9

(a). The probabilities are: ($p_X = \frac{1}{91}$, $p_Y = \frac{9}{91}$, $p_Z = \frac{81}{91}$).

(b). The entropy of the source based on the *independent* model is $H(X_n) = H(\frac{1}{91}, \frac{9}{91}, \frac{81}{91}) = 0.5512$.

(c). The entropy of the source based on the *1st order Markov* model is $H(X_n \mid X_{n-1}) = H(.1, .9) = 0.4690$.

Case (3):

from	to X	to Y	to Z
X	.9	.1	.0
Y	.0	.9	.1
Z	.1	.0	.9

(a). The probabilities are: ($p_X = p_Y = p_Z = 1/3$).

(b). The entropy of the source based on the *independent* model is $H(X_n) = H(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) = 1.585$.

(c). The entropy of the source based on the *1st order Markov* model is $H(X_n \mid X_{n-1}) = H(.1, .9) = 0.4690$.

Case (4):

from	to X	to Y	to Z
X	.0	.5	.5
Y	.5	.0	.5
Z	.5	.5	.0

(a). The probabilities are: ($p_X = p_Y = p_Z = 1/3$).

(b). The entropy of the source based on the *independent* model is $H(X_n) = H(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) = 1.585$.

(c). The entropy of the source based on the *1st order Markov model* is $H(X_n \mid X_{n-1}) = H(.5, .5) = 1.000$.