Source Coding Homework #1 (Computer)

Lossless Compression

(H.-M. Hang, October 2009)

- You are asked to write your own programs (in C or C++) to compute entropy and to design and implement the Huffman codes and the LZW codes.
- You should be able to demonstrate your programs on a workstation or a PC
- Please attach the print out of your programs to your report.
- You may want consider submitting your homework using Internet emails. Please keep a copy of your own email and be sure to get **a confirmation** from the instructor. The date of your successful email is the date of your submission. Your report can be in the format of Microsoft Words, PDF, or postscript.

- *Reminder:* The organization, clarity, etc. of your report contributes to 30% of the report score.

There are three parts in this computer assignment. Part 0 is estimating the probability distributions from the given data files and computing the corresponding entropy values. Part 1 is Huffman coding, and Part 2 is LZW coding. There are two test data files. The first file is an i.i.d. sequence ("iid.dat") and the second one is a first-order two-state Markov sequence ("markov.dat"). Both of them contain binary data. However, they are stored in bytes. That is, eight (8) samples of binary data are stored in one byte. The least significant bit in a byte is the first bit (rightmost) in the data sequence. Each file contains 10,000 bytes (80,000 samples of data).

Part 0: Basics

(1) Use block size N=1, compute the first-order probability and the corresponding entropy values for these two data files.

(2) Use block size N=2, repeat the process in **(1)**; that is, compute the second-order joint probability and the corresponding entropy values for these two data files.

(3) Use block size N=4, repeat the process in **(1)**.

(4) Use block size N=8, repeat the process in **(1)**.

In your report, print the entropy values and draw the probability distributions (using figures) of all the above eight cases.

Part 1: Huffman codes

Design the Huffman codes for all the cases N=2, 4 and 8 described in **Part 0**. Use your designed Huffman codes to encode these two files. Essentially, you have six Huffman codebooks: three for the i.i.d. data with N=2, 4 and 8 and three for the Markov data with N=2, 4 and 8. In your report, list all your Huffman codebooks for N=2 and 4. Also shown in you reports are the numbers of bits for the following 12 cases: apply the above six Huffman codebooks to both the i.i.d. data and the Markov data.

Part 2: LZW codes

Design the LZW codes for the two given data files with two codebook sizes: 16 and 256. (Note: The first 2 entries in the codebooks are fixed. They are the alphabets: 0 and 1) That is, design a LZW codebook of size 16 and a LZW codebook of size 256 for the i.i.d. data and apply them to the i.i.d. data. Also, design a LZW codebook of size 16 and a LZW codebook of size 256 for the Markov data and apply them to the Markov data. In your report, list the first 64 (or 16) entries of your LZW codebooks.

Optional: If you like to know the conditional probabilities of these two data files, you can do so by estimating $P(0|0)$, $P(1|0)$, $P(0|1)$, and $P(1|1)$. The conditional probability may help you in explaining your results.

Remarks:

In doing research, it is important to interpret the results you obtained. At the end of your report, explain as much as you can the meaning of your results. *For example,* you should compare the entropy value and the computed compressed bit rate for each case and explain why they are similar (in values) or different. Explain the differences in compressing the i.i.d. data file and the Markov data file.

DUE: October 22, 2009