ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
School of Computer and Communication Sciences

Handout 12 Information Theory and Coding
Solutions to homework 6 November 4, 2007

PROBLEM 1.

(a) Since the lengths prescribed satisfy the Kraft inequality, an instantaneous code can be used for the final stage of encoding the intermediate digits to binary codewords. In this case, each stage of the encoding is uniquely decodable, and thus the overall code is uniquely decodable.

(b) The indicated source sequences have probabilities $0.1, (0.9)(0.1), (0.9)^2(0.1), (0.9)^3(0.1), \ldots, (0.9)^7(0.1), (0.9)^8$. Thus,

$$\bar{N} = \sum_{i=1}^{8} i(0.1)(0.9)^{i-1} + 8(0.9)^8 = 5.6953.$$

(c)

$$\bar{M} = 1(0.9)^8 + 4[1 - (0.9)^8] = 2.7086.$$

(d) Let $N(i)$ be the number of source digits giving rise to the first i intermediate digits. By the law of large numbers, with probability 1

$$\lim_{i\to\infty} \frac{N(i)}{i} = \bar{N}$$

Similarly, let $M(i)$ be the number of encoded bits corresponding the the first i intermediate digits. Then, with probability 1

$$\lim_{i\to\infty} \frac{M(i)}{i} = \bar{M}$$

From this, we have that with probability 1

$$\lim_{i\to\infty} \frac{M(i)}{N(i)} = \frac{\bar{M}}{\bar{N}}$$

and that for a long source sequence the number of encoded bits per source digit will be $\bar{M}/\bar{N} = 0.4756$.

A straightforward but tedious construction reveals that the average length of the Huffman code encoding 4 source digits at a time is 1.9702, yielding $1.9702/4 = 0.49255$ encoded bits per source digit.

For those of you puzzled by the fact that the 'optimum' Huffman code gives a worse result for this source than the run-length coding technique, observe that the Huffman code is the optimal solution to a mathematical problem with a given message set, but the choice of a message set can be more important than the choice of code words for a given message set.

PROBLEM 2.

(a) Since $\hat{p}(x) = \max_{1 \le k \le K} p_k(x)$, for any k, $p_k(x) \le \hat{p}(x)$, Also, since $p_k(x) \ge 0$,

$$\hat{p}(x) \le \sum_{k=1}^{K} p_k(x).$$

We thus have

$$1 = \sum_x p_k(x) \le \sum_x \hat{p}(x) \le \sum_k \sum_x p_k(x) = K.$$

(b) Since $l(x) \ge -\log_2 \hat{p}(x) + \log_2 A$,

$$\sum_x 2^{-l(x)} \le \sum_x \hat{p}(x)/A = 1,$$

We see that $l(x)$ satisfies the Kraft's inequality and we conclude that there exists a prefix free code with these lengths.

(c) We know that for any uniquely decodable code $\bar{L}_k = \sum_x p_k(x) l(x) \ge H_k$, so the left hand inequality follows. For the right hand, observe that

$$l(x) < -\log_2 \hat{p}(x) + \log_2 A + 1,$$

and since $\hat{p}(x) \ge p_k(x)$, we have

$$l(x) < -\log_2 p_k(x) + \log_2 A + 1,$$

and thus

$$\bar{L}_k < H_k + \log_2 A + 1.$$

The approach outlined here to universal coding is explored further in A. Orlitsky, N. P. Santhanam, J. Zhang "Universal compression of memoryless sources over unknown alphabets", *IEEE Transactions on Information Theory,* vol. 50, no. 7, pp. 1469–1481, July 2004. The quantity A is known as the "Shtarkov sum", and was introduced by Y. Shtarkov in "Universal sequential coding of single messages," *Problems of Information Transmission,* vol. 23 no. 3, pp. 3–17, 1987.