ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
School of Computer and Communication Sciences

**Handout 11** Information Theory and Coding
Homework 6 November 2, 2007

PROBLEM 1. The following problem concerns a technique known as run length coding. Along with being a useful technique, it should make you look carefully into the sense in which Huffman coding is optimal. A source produces a sequence of independent binary digits with probabilities $P(\mathbf{0}) = 0.9$ and $P(\mathbf{1}) = 0.1$. We shall encode this sequence in two stages, first counting the number of 0's between successive 1's in the source output, and then encoding these counts into binary code words. The first stage of encoding maps source sequences into intermediate digits by the following rule:

| Source Sequence | Intermediate Digits (# of zeros) |
|---|---|
| **1** | 0 |
| **01** | 1 |
| **001** | 2 |
| **0001** | 3 |
| ⋮ | ⋮ |
| **00000001** | 7 |
| **00000000** | 8 |

Thus the following sequence is encoded as follows:

$$\begin{array}{cccccccccccccccccccc} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \\ 0, & & & 2, & & & & & & & & 8, & & & 2, & 0, & & & & & 4 \end{array}$$

The final stage of encoding assigns a code word of length 1 to the intermediate digit 8 and codewords of length 4 to the other intermediate digits.

(a) Justify, in whatever detail you find convincing to yourself that the overall code is uniquely decodable.

(b) Find the average number $\bar{N}$ of source digits per intermediate digit.

(c) Find the average number $\bar{M}$ of encoded binary digits per intermediate digit.

(d) Show, by appeal to the law of large numbers, that for a very long source sequence of source digits, the ratio of the number of encoded binary digits to the number of source digits will with high probability be close to $\bar{M}/\bar{N}$. Compare this ratio to the average number number of code letters per source letter for a Huffman code encoding four source digits at a time.

PROBLEM 2. Suppose that a source has alphabet $\mathcal{X}$, and, it is known that its distribution is either $p_1(x)$ or $p_2(x)$, ..., or $p_K(x)$. Let $H_k = -\sum_x p_k(x) \log_2 p_k(x)$ denote the entropy of the distribution $p_k$, $k = 1, \dots, K$. Define $\hat{p}(x) = \max_{1 \le k \le K} p_k(x)$, and $A = \sum_x \hat{p}(x)$.

(a) (10 pts) Show that $1 \le A \le K$.

(b) (10 pts) Show that there is a prefix free source code for $X$ with codeword lenghts $l(x) = \lceil -\log_2 \hat{p}(x) + \log_2 A \rceil$.

(c) (10 pts) Show that, for a code as in (b), $\bar{L}_k = \sum_x p_k(x) l(x)$, (the average codeword length under distribution $p_k$), satisfies

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