ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
School of Computer and Communication Sciences

Handout 8 Information Theory and Coding
Homework 4 (Graded - Due on Oct. 16, 2013 - 6 PM) Oct. 08, 2013

PROBLEM 1. (25 pts) In this problem, we investigate what happens to the minimal average length of a code when we relax unique decodability while keeping non-singularity. Consider a random variable U which takes values in $\{a_1, \ldots, a_M\}$, and consider a non-singular code for the alphabet $\{a_1, \ldots, a_M\}$ with average length $\bar{l}$. Note that a non-singular code can contain the empty string $\o$ of length 0.

(a) Show that there exist unique integers $k > 0$ and $r \geq 0$ such that $M = 2^k - 1 + r$ and $r < 2^k$.

(b) Show that any non-singular code satisfies $\sum_i 2^{-l_i} \leq k + r2^{-k} \leq \lceil \log_2(M+1) \rceil$, where l_i is the length of the codeword associated to the symbol a_i.

(c) Show that any non-singular code satisfies $\bar{l} \geq H(U) - \log_2(k + r2^{-k}) \geq H(U) - \log_2 \lceil \log_2(M+1) \rceil$.

PROBLEM 2. (25 pts) This problem explains what can be done when there is an uncertainty about the true distribution of the source. Consider a source U with alphabet $\mathcal{U} = \{a_1, \ldots, a_M\}$ and suppose that we know that the true distribution of U is either P_1 or P_2 but we are not sure which.

(a) Show that there is a prefix-free code where the length of the codeword associated to a_i is $l_i = \lceil \log_2 \frac{2}{P_1(a_i) + P_2(a_i)} \rceil$.

(b) Show that the average (computed using the true distribution) length $\bar{l}$ of the code constructed in (a) satisfies $H(U) \leq \bar{l} < H(U) + 2$.

Now assume that the true distribution of U is one of k distributions $P_1, \ldots, P_k$ but we don't know which.

(c) Show that there exists a prefix-free code satisfying $H(U) \leq \bar{l} < H(U) + \log_2 k + 1$.

PROBLEM 3. (30 pts) A Huffman code is said to be maximally branched if it has at least one codeword of length l for every $1 \leq l \leq l_{\max}$ where $l_{\max}$ is the length of the longest codeword.

(a) Show that a maximally branched Huffman code has exactly one codeword of length l for every $1 \leq l < l_{\max}$ and exactly two codewords of length $l_{\max}$.

Now consider a source with input alphabet $\{a_1, \ldots, a_M\}$ where $P(a_1) \leq \ldots \leq P(a_M)$.

(b) Show that the source has a maximally branched Huffman code if and only if $P(a_i) \geq \sum_{j=1}^{i-2} P(a_j)$ for every $3 \leq i \leq M$.

(c) Show that if $P(a_i) > \sum_{j=1}^{i-2} P(a_j)$ for every $3 \le i \le M$, then every Huffman code is maximally branched.

(d) Consider the particular case in which $P(a_i)$ is proportional to φ^i, where φ is the golden ratio $\frac{1+\sqrt{5}}{2}$. Show that the source has a maximally branched Huffman code.

PROBLEM 4. (20 pts) Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be n pairs of random variables which may or may not be independent. For every $i \ge 1$ and $j \le n$, define X_i^j to be the sequence $X_i, \ldots, X_j$ if $i \le j$, and to be $\emptyset$ if $i > j$. Define Y_i^j similarly. Therefore, since $X_{n+1}^n = Y_1^0 = \emptyset$ we have $I(X_{n+1}^n; Y_n) = I(Y_1^0; X_1) = 0$ and $I(Y_1^{n-1}; X_n | X_{n+1}^n) = I(Y_1^{n-1}; X_n)$.

(a) Show that $I(Y_1^{n-1}; X_n) = \sum_{i=1}^{n-1} I(X_n; Y_i | Y_1^{i-1})$.

(b) Show that $\sum_{i=1}^{n} I(X_{i+1}^n; Y_i | Y_1^{i-1}) = \sum_{i=1}^{n} I(Y_1^{i-1}; X_i | X_{i+1}^n)$.