ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Handout 5	Information Theory and Coding
Solutions to homework 2	Oct. 03, 2022

Problem 1.

(a) Code I is prefix-free, Code II is not.

(b) Both codes are uniquely decodable: Code I because it is instantaneous, Code II because the 1's at the beginning of each code word act as markers that separates the codewords and the decoding can be performed by counting the 0's between the 1's.

(c) Since each codeword of code II begins with the letter 1 and since the letter 1 only appears at the beginning of codewords, this letter acts as an indicator of start of a codeword.

Problem 2.

(a) Recall that $\mathcal{C}$ is uniquely decodable means that $\mathcal{C}^*$ is injective, i.e., for any $u^n \neq v^m$ we have $\mathcal{C}^n(u^n) \neq \mathcal{C}^m(v^m)$. In particular, whenever $u^n \neq v^n$ we have $\mathcal{C}^n(u^n) \neq \mathcal{C}^n(v^n)$. The last statement is the definition of $\mathcal{C}^n$ being injective.

(b) Since we are supposed to show that $u_1 \neq v_1$, we may assume that $|\mathcal{U}| \geq 2$.

If $\mathcal{C}$ is not uniquely decodable, then there are $u^n \neq v^m$ such that $\mathcal{C}^n(u^n) = \mathcal{C}^m(v^m)$. Among all such (u^n, v^m) choose one for which $n+m$ is smallest, and assume (without loss of generality) that $m \leq n$. If $m \geq 1$ we are done, since in this case we must have $u_1 \neq v_1$ (because, if not, we can replace u^n by $\tilde{u}^{n-1} = u_2 \dots u_n$ and v^m by $\tilde{v}^{m-1} = v_2 \dots v_m$, contradicting $m+n$ being smallest).

Otherwise, $m = 0$ and $v^m = \lambda$ (the null string) with $\mathcal{C}(v^m) = \lambda$. Since $u^n \neq v^m = \lambda$ and $\mathcal{C}(u^n) = \lambda$, we have a letter $a = u_1 \in \mathcal{U}$ such that $\mathcal{C}(a) = \lambda$. Take now any letter $b \in \mathcal{U}$ with $b \neq a$, and note that $\mathcal{C}^2(ab) = \mathcal{C}^1(b)$, i.e., there are two source sequences that differ in their first letter and have the same representation.

(c) $\mathcal{C}$ is not uniquely decodable means that there is $u^n \neq v^m$ such that $\mathcal{C}^n(u^n) = \mathcal{C}^m(v^m)$. If $n = m$ then we are done: this would by definition mean that be $\mathcal{C}^n$ is not injective. If $n \neq m$, we could attempt the following reasoning: observe $\mathcal{C}^*(u^n v^m) = \mathcal{C}^*(v^m u^n)$ and conclude that $\mathcal{C}^{m+n}$ is not injective. However this reasoning fails because we can't be sure that $u^n v^m \neq v^m u^n$ just because $u^n \neq v^m$. (E.g., suppose $u^n = a$ and $v^m = aa$). This is the reason the problem has "part (b)":

As $\mathcal{C}$ is not uniquely decodable, we can find u^n and v^m as in part (b). Now observe that (i) $u^n v^m \neq v^m u^n$ (as they differ in their first letter), (ii) $u^n v^m$ and $v^m u^n$ have the same length $k = n + m$, and $\mathcal{C}^k(u^n v^m) = \mathcal{C}^k(v^m u^n)$, i.e., $\mathcal{C}^k$ is not singular.

Problem 3. Since the class of instantaneous codewords is a subset of the class of uniquely decodable codewords, it follows that $\bar{M}_2 \leq \bar{M}_1$. On the other hand, let $\{l_i\}$ be the codeword lengths of the uniquely decodable code for which $\bar{M} = \bar{M}_2$. Since $\{l_i\}$ satisfies the Kraft's inequality, there exists an instantaneous code with these codeword lengths. For this instantaneous code $\bar{M} = M_2$ and we see that $\bar{M}_1 \leq \bar{M} = M_2$, and we conclude that $\bar{M}_1 = \bar{M}_2$.

Problem 4.

(a) $\{00, 01, 100, 101, 1100, 1101, 1110, 1111\}$.

(b) First note that if any two number differ by 2^{-k}, their binary expansion will differ somewhere in the first k bits after the 'point'. (Think of the decimal case: if $a = 0.375\ldots$ and b differs by more than 10^{-3} by it, then b's expansion cannot start with 0.375.)

Next observe that that for $i > j$

$$Q_i - Q_j = \sum_{k=j}^{i-1} P(a_k) \geq P(a_j) \geq 2^{-l_j}.$$

So, the binary expansion of Q_i and Q_j must differ somewhere in the first l_j bits. Since codewords for i and j are at least l_j bits long, neither codeword can be a prefix of the other.

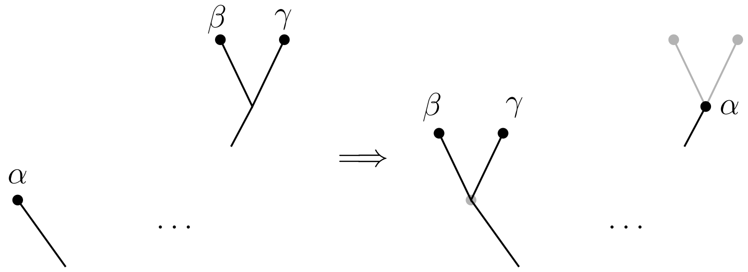
The bound on the average codeword length follows from

$$-\log_2 P(a_i) \leq l_i < -\log_2 P(a_i) + 1.$$

This method of coding is also known as Shannon coding and predates Huffman coding.

Problem 5.

(a) Consider the longest and the shortest codewords. We know that there are at least two longest codewords, suppose their length is l. Suppose the shortest codewords has length s. Suppose that s and l differ by 2 or more. To show that this cannot be the case for an optimal code, consider the transformation shown below:



We see that the transformation decreases the length of two codewords (for letters β and γ) by $l-(s+1)=l-s-1$, whereas it increases the length of one codeword (for the letter α) by $(l-1)-s=l-s-1$. But since $l-s-1>0$, and since all the codewords are equally likely, this would have decreased the average codeword length, contradicting the optimality of the Huffman code. Thus, the longest and shortest codeword lengths can differ by at most 1, and these lengths must be j and $j+1$. (If some other two consecutive depths were used we would either not have enough leaves, or have too many leaves).

(b) Let the number of codewords of length k be m_k, $k=j,j+1$. Since the Huffman procedure yields a complete tree (no leaf is unoccupied) all intermediate nodes have two children. Thus, the 2^j nodes at level j of the tree are either codewords (m_j of them) or each of their two children are codewords ($m_{j+1}/2$ of them). Thus

$$m_j + m_{j+1}/2 = 2^j,$$

and also $m_j + m_{j+1} = x2^j$. From these two equations we find

$$m_j = (2-x)2^j \quad \text{and} \quad m_{j+1} = (x-1)2^{j+1}.$$

(c) By the result of (b) the average codeword length is

$$[jm_j + (j+1)m_{j+1}]/(x2^j) = j + 2(x-1)/x.$$

Problem 6. *An* optimal set of codewords for the two sources are as follows:

Source I		Source II	
Binary	Ternary	Binary	Ternary
00	0	00	0
01	10	01	1
100	11	100	21
101	12	101	20
110	20	110	220
111	21	1110	221
		1111	222

with average codeword lengths 2.5, 1.7, 2.55, 1.65 digits/symbol, in the order the codes appear in the table.

Note that for the ternary code for Source I, we need to add to the symbols of the source an extra symbol of probability zero so that the number of symbols equal 1 modulo $D-1$.