Symbol codes

P1 (Kraft's inequality) Let each source symbol from the alphabet $A = \{a_1, \dots, a_n\}$ be encoded into a uniquely decodable code over an alphabet of size r with codeword lengths $l_1, l_2, \dots, l_n$ respectively. Then

$$\sum_{i=1}^{n} r^{-l_i} \leq 1 .$$

Sol: See wikipedia.

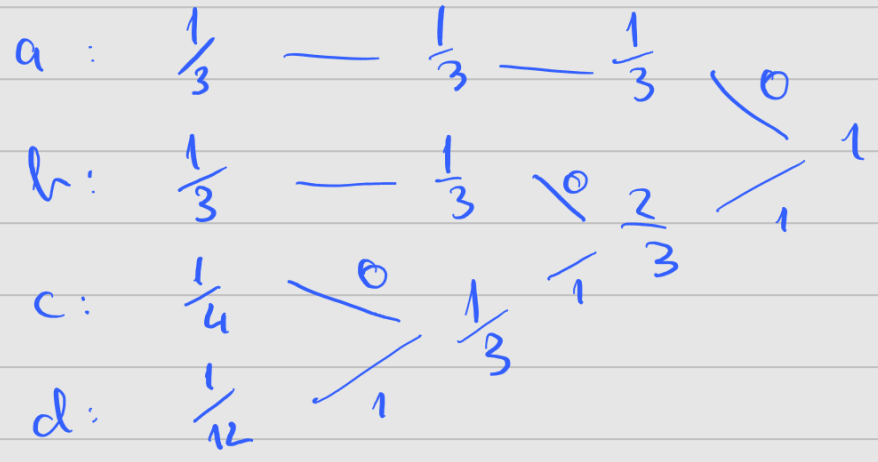
P2 Consider a random variable Y which takes on 4 values with probabilities $(1/3, 1/3, 1/4, 1/12)$.

a) Construct a Huffman code for Y.

b) Show that there exist two different sets of optimal lengths for the codewords.

c) Are there optimal codes with codewords lengths for some symbols that exceed the information content of the symbol $\lceil \log_2 \frac{1}{p(y)} \rceil$?

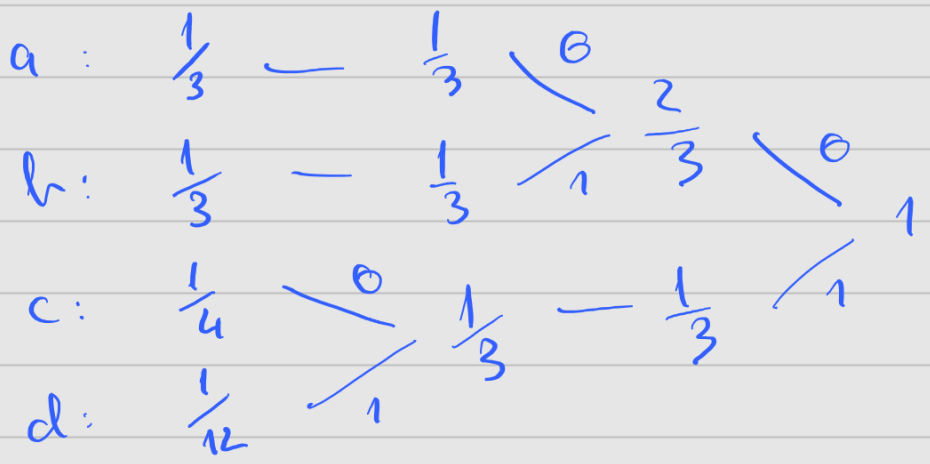
Sol: We construct a Huffman tree for Y as follows:



So the Huffman code for Y is:

a ↔ 0
b ↔ 10
c ↔ 110
d ↔ 111

b) We can also construct the following Huffman tree for Y:



and the corresponding Huffman code is

a ↔ 00
b ↔ 01
c ↔ 10
d ↔ 11

In a) lengths of the codewords are:

$(l_1, l_2, l_3, l_4) = (1, 2, 3, 3)$ and in b)

$(l_1, l_2, l_3, l_4) = (2, 2, 2, 2)$; so we have two different set of lengths.

c) In a) the length of the codeword for c is 3, but

$$\left\lceil \log_2 \frac{1}{P(Y=c)} \right\rceil = \left\lceil \log_2 \frac{1}{\frac{1}{4}} \right\rceil = \lceil \log_2 4 \rceil = \lceil 2 \rceil = 2 ,$$

so there are codes with codewords lengths for some symbols that exceed the information content of the symbol $\left\lceil \log_2 \frac{1}{P(y)} \right\rceil$, and one example is the code in a).

P4 Prove that for an optimal code with the maximum codeword length l_{max}, has at least two codewords with length l_{max}.

Pf: W.L.O.G. we can assume that the codeword corresponding c_{max} to l_{max} is the last with respect to the lexicographical order of codewords. If we have:

$$\sum_{i=1}^{n} 2^{-l_i} = 1 ,$$

and only one codeword corresponding to l_{max}, then:

$$2^{l_{max}-1} \cdot \sum_{i=1}^{n} 2^{-l_i} = 2^{l_{max}-1} , \text{ so}$$

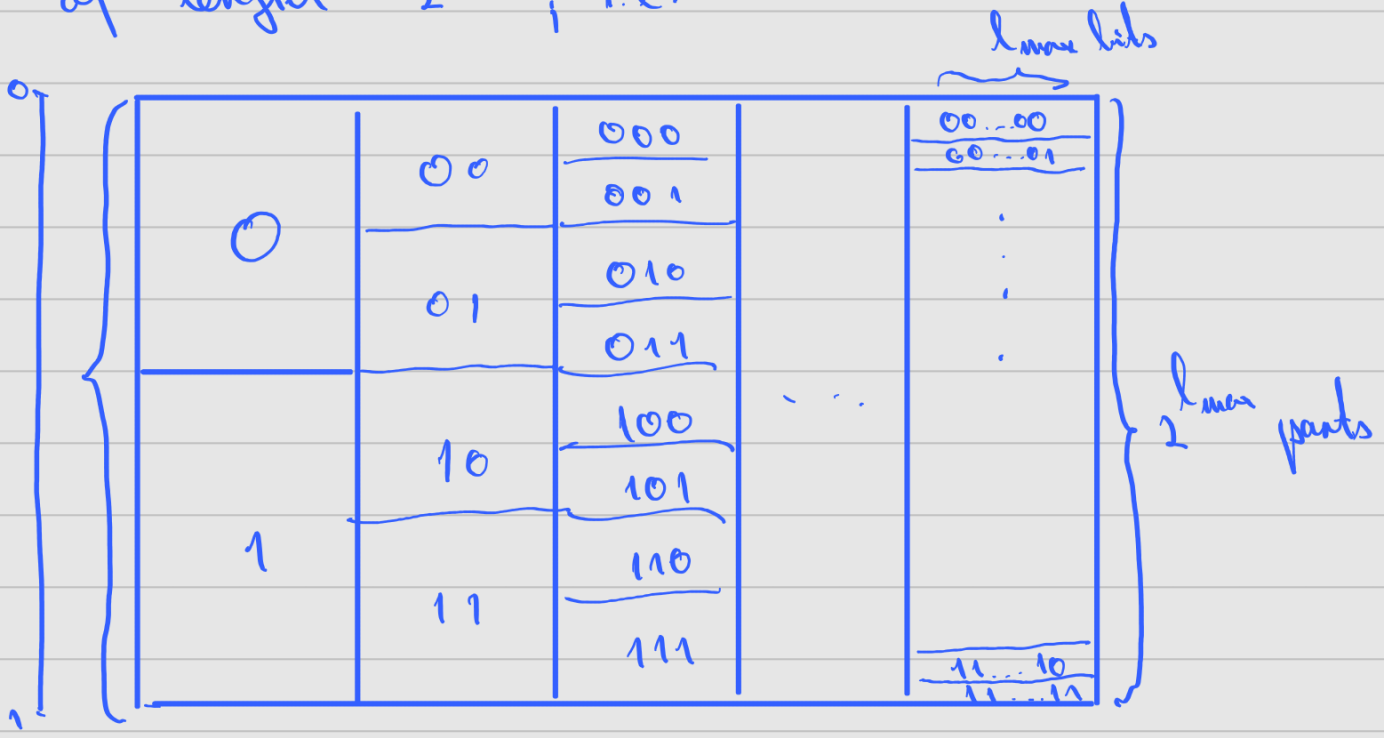
$$\underbrace{\frac{1}{2}}_{\text{not an integer}} + \underbrace{\sum_{l_i \leq l_{max}-1} \left(2^{-l_i} \cdot 2^{l_{max}-1} \right)}_{\text{integer}} = \underbrace{2^{l_{max}-1}}_{\text{integer}}$$

and thus would be a contradiction.

If $\sum_{i=1}^{} 2^{-l_i} < 1$, then we could assign to symbol a_{max} a shorter codeword by discarding the last bit of c_{max}, and this is not possible because the code is optimal. Hence we conclude that there are at least two codewords with length l_{max}.

P3 Prove that for any set of codewords $\{l_i\}$ satisfying the Kraft inequality there is a prefix code with those lengths.

Sol: Divide the interval $[0,1]$ in $2^{l_{max}}$ parts of length $2^{-l_{max}}$, i.e.



If we order the lengths $l_1, l_2, \ldots, l_k$ in an non-increasing way, the we start by assigning to symbol a_1 codeword c_1 of length l_1 containing $00\ldots0$ to the left, i.e. we assign it l_1 zeros. Then, starting where that interval ended i.e. 2^{l_1}, we assign to a_2 the

codeword c_2 of length l_2 containing on the left the next l_{max} bit word after c_1, i.e. 2^{l_1} in binary, and we continue doing that. Since $\sum 2^{-l_i} \leq 1$ we will construct the prefix code without running out of space.

P5 Prove that Huffman coding is optimal.

Sol: We know from P4 that there are at least two codewords with l_{max}. Among them choose the two with the smallest prob. Denote them by m_1 and m_2, and corr. prob. p_1 and p_2. If there is m_i with prob. $p_i < p_1$ and $l_i < l_1$, then by assigning to a_i m_i and to a_i m_1 instead, we would get a shorter avg. length, which is impossible. Similarly for p_2. Hence p_1 and p_2 are the smallest prob. in general. If like in P3 table we just delete the last bit of m_1 and look at the code obtained for $n-1$ symbols by setting a_{n-1} prob. to be $p_1 + p_2$, and if by inductive hypothesis for $n-1$ symb. Huffman is optimal, then the code that we started with will have the same avg. length as the Huffman extended to the n symbols. Hence Huffman coding is optimal.

2. a) Find the Huffman code for the random variable X given by:

$$X = \begin{pmatrix} a & b & c & d & e & f \\ \frac{1}{42} & \frac{5}{42} & \frac{3}{21} & \frac{4}{21} & \frac{5}{21} & \frac{6}{21} \end{pmatrix}$$

and calculate the expected length of the codewords.

4. To construct a source code for the variable using Fano coding, order the outcomes according to the probabilities, with the highest to the left, i.e. the vector becomes x_1, x_2, x_k, where $p_1 \geq p_2 \geq \ldots \geq p_k$. For each of the probability vectors below find the Fano code and state if it is also a Huffman code and/or if it is optimal.

a) $P_a = (\frac{4}{10}, \frac{3}{10}, \frac{2}{10}, \frac{1}{10})$

b) $P_b = (\frac{6}{21}, \frac{5}{21}, \frac{4}{21}, \frac{3}{21}, \frac{2}{21}, \frac{1}{21})$

c) $P_c = (\frac{15}{43}, \frac{7}{43}, \frac{7}{43}, \frac{7}{43}, \frac{7}{43})$