Practice in data compression.
Problem solving session
Instant coding

Boris Kudryashov

ITMO University / University of Tartu
1. Construct a Huffman code for the DMS with p.d. 
\( (0.3, 0.25, 0.15, 0.1, 0.1, 0.05, 0.05) \)
Compare the average codeword length with the entropy.
1. Construct a Huffman code for the DMS with p.d. 
\((0.3, 0.25, 0.15, 0.1, 0.1, 0.05, 0.05)\)
Compare the average codeword length with the entropy.

2. For DMS with alphabet \(X = \{a, b, c\}\) and p.d.
   \begin{enumerate}
   \item \(p(a) = 1/2, p(b) = 1/4, p(c) = 1/4;\)
   \item \(p(a) = p(b) = p(c) = 1/3.\)
   \end{enumerate}
construct Huffman codes for \(X, X^2, X^3.\)
Compare code rates with the entropy of the source.
3* Analyse behaviour of the Huffman code rate as a function of length $n$ of encoded blocks for the Binary Memoryless Source (BMS) with probability of 1 equal to 0.1.
4* Optimal coding for uniform distributions Find a Huffman code for source, which chooses letters from alphabet of size $M$ with uniform distribution. Calculate the average length of codewords and redundancy as a function of $M$, derive upper and lower bounds of redundancy.
Hints:
Intermediate steps Huffman code for this source contains
$D = 2 \cdot 2^{\lfloor \log M \rfloor} - M$ words of length $\lfloor \log M \rfloor$ and $M - D$ words of length $\lfloor \log M \rfloor + 1$. Average length of codewords and redundancy are equal to

$$\bar{l} = \lfloor \log M \rfloor + 2 - \frac{2}{M} 2^{\lfloor \log M \rfloor},$$

$$r = \bar{l} - \log M = 2 - d - 2 \cdot 2^{-d},$$

where $d = \log M - \lfloor \log M \rfloor$, $d \in [0, 1)$. Differentiation with respect to $d$ shows, that maximum of redundancy is achieved when $d = 1 - \log \log e$ and is equal to $1 + \log \log e - \log e \approx 0.0861$. 
5 Consider Markov chain with the probability transition matrix:

\[
P = \begin{bmatrix}
0 & \frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\
\frac{1}{4} & \frac{1}{2} & \frac{1}{4}
\end{bmatrix}.
\]

Calculate \( H(X) \), \( H_n(X) \), \( n = 1, 2, \ldots \), \( H(X/X^n) \), \( n = 1, 2, \ldots \), assuming that the initial distribution on letters is the stationary distribution. Find code rate for coding ensembles \( X \) and \( X^2 \) using Huffman coding. Suggest such a method for encoding, that code rate is equal to the entropy rate of the source.
6* Non-uniform run-length coding Sequence at the output of binary source is uniquely represented in the form of concatenation of subsequences 1, 01, ..., 0^{L-1}1, 0^L for some integer \( L \).

A Huffman code is used for encoding indices of subsequences from this set. For a BMS with probability of 1 equal to \( p=0,1 \), analyze behavior of the code rate as a function \( L \).

Code rate in this case is computed as

\[
R = \bar{n}/\bar{\tau}
\]

where \( \bar{\tau} \) is average length of sequence to be encoded; \( \bar{n} \) is average length of codewords.
**7** Variable-to-Variable (VV) coding. Tunstall code.
Recall that the redundancy of the Huffman code is determined by the maximum probability of letters.
Let $X = p(x)$ and $x_0 \in X$ is a most probable letter. Introduce a new ensemble $X_1$, which consists of all letters from $X \setminus x_0$ and of all pairs $(x_0, x)$, $x \in X$. Probabilities of pairs of letters are calculated like products of probabilities of corresponding letters. This procedure is applied recursively.
Compare the efficiency of Tunstall code for for BSS with other coding techniques based on the extension of the code alphabet.
8 Construct Shannon code for the source from Pr. 1:

\[ p = (0.3, 0.25, 0.15, 0.1, 0.1, 0.05, 0.05) \]

Compare its average length with the average length of Huffman code codewords and with entropy of the source.
8 Construct Shannon code for the source from Pr. 1:
\[ p = (0.3, 0.25, 0.15, 0.1, 0.1, 0.05, 0.05) \]
Compare its average length with the average length of Huffman code codewords and with entropy of the source.

9 Construct Gilbert-Moore code for the same source. Compare its average length with the average length of Huffman code codewords and with entropy of the source.
Problems

10 Use arithmetic coding to encode the sequence 01001 from the BMS with probability of 1 is 0.4. Compare the length of code sequence with the amount of information of the sequence.

11* What changes should be done in the algorithms and programs of arithmetic coding and decoding, to apply them to source, described by Markov chain.