LINKÖPING UNIVERSITY

Written Exam in
Data compression
TSBK08

27th August 2022 8:00 - 12:00

Location:	TER2
Examiner:	Harald Nautsch
Teacher:	Harald Nautsch, 0701718715
Department:	ISY
Module:	TEN1
Number of problems:	6
Number of pages:	4
Permitted equipment:	Calculator, general English dictionaries
Other:	Answers can be given in English or in Swedish. The teacher is only available by phone during the exam.
Grades:	0-13 U 14-19 3 20-25 4 26-30 5

1 a) Formulate Kraft-McMillan's inequality.

(1 p)

b) Explain what an instantaneous code is.

(1 p)

c) Explain how prediction with partial match (ppm) coding works.

(2 p)

d) Explain what a Golomb code is and what type of probability distribution it is good for.

(2 p)

e) Explain what the rate-distortion function is and how it is calculated for a stationary memoryless random source.

(2 p)

2 Let $H(X)$ be the entropy of the random variable X. Show that

$$0 \leq H(X) \leq \log L$$

where L is the size of the alphabet.

(4 p)

3 A memoryless source has the alphabet

$$\mathcal{A} = \{a,\ b,\ c,\ d,\ e,\ f,\ g,\ h\}$$

The symbol probabilities are

$$p(a) = 0.41,\ p(b) = 0.12,\ p(c) = 0.11,\ p(d) = 0.10$$

$$p(e) = 0.10,\ p(f) = 0.06,\ p(g) = 0.06,\ p(h) = 0.04$$

Construct a Huffman code for the source and calculate the resulting average data rate (in bits/symbol) of the code.

(2 p)

4 A stationary Markov source X_n of order 1, with alphabet $\mathcal{A} = \{a,\ b,\ c\}$, is given by the transition probabilities $p(x_n|x_{n-1})$ below

$$\begin{array}{lll} p(a|a) = 0.75 & p(b|a) = 0.15 & p(c|a) = 0.10 \\ p(a|b) = 0.25 & p(b|b) = 0.6 & p(c|b) = 0.15 \\ p(a|c) = 0.05 & p(b|c) = 0.15 & p(c|c) = 0.8 \end{array}$$

a) Calculate the entropies $H(X_n)$, $H(X_n|X_{n-1})$ and $H(X_n, X_{n+1}, X_{n+2})$ for the source.

(3 p)

b) Code the sequence

aabbcc

using arithmetic coding. The coding should utilize the memory of the source. Give both the interval and the corresponding codeword. You can assume that the source is in state a when the coding starts and that all calculations are performed with infinite precision.

(4 p)

5 A source has the alphabet $\mathcal{A} = \{m,\ n,\ o,\ p\}$. A symbol sequence of length 8 is coded using BWT and mtf. The resulting index is 5 and the mtf-coded sequence is 2,0,3,0,0,1,2,0. Decode the symbol sequence.

(3 p)

6 A source has the alphabet $\{a,\ b,\ c,\ d,\ e,\ f,\ g,\ h,\ i,\ j,\ k,\ l,\ m,\ n,\ o,\ p\}$.

a) Code the sequence that begins

$$kankankapankanbbbbapane\ldots$$

using LZSS. The length of the history buffer (search buffer) is chosen as 512. Matchlengths are coded using 4 bit fixed length codewords.

(3 p)

b) A sequence from the source is coded using LZW, giving the following index sequence:

$$6,\ 0,\ 16,\ 18,\ 0,\ 12,\ 8,\ 16,\ 21,\ 6,\ 14,\ 25,\ 27,\ 4,\ \ldots$$

The initial dictionary is:

index	symbol	index	symbol
0	a	8	i
1	b	9	j
2	c	10	k
3	d	11	l
4	e	12	m
5	f	13	n
6	g	14	o
7	h	15	p

Decode the index sequence as far as possible. Also give the resulting dictionary.

(3 p)