Solutions to Written Exam in

Data compression

TSBK08

21st March 2022

1 a) See the course literature.

b) See the course literature.

c) See the course literature.

d) See the course literature.

e) See the course literature.

f) See the course literature.

2 See the course literature.

3 Assume that we have M levels in our quantizer. The stepsize will then be $\Delta = \frac{2}{M}$. The mean square error will be $D = \frac{\Delta^2}{12} = \frac{1}{3M^2}$.

By coding sufficently long sequences into each codeword, the rate can be arbitrarily close to the entropy rate of the quantized signal $\hat{X}_k$. The quantized signal is memoryless, discrete uniformly distributed with alphabet size M. Thus the rate is given by

$$R = H(\hat{X}_k) = \log_2 M = \frac{1}{2}\log_2 M^2 = \frac{1}{2}\log_2 \frac{1}{3D}$$

4 From the given distribution we can easily find the marginal distributions for pairs of symbols $p(x_i, x_{i+1})$ and single symbols $p(x_i)$:

$$\begin{array}{ll} p(a,a) = 7/11 & p(a,b) = 1/11 \\ p(b,a) = 1/11 & p(b,b) = 2/11 \end{array}$$

$$p(a) = 8/11 \quad p(b) = 3/11$$

From these distributions we can calculate the entropies

$$\begin{aligned} H(X_i) &\approx 0.8454 \\ H(X_i, X_{i+1}) &\approx 1.4911 \\ H(X_i, X_{i+1}, X_{i+2}) &\approx 2.1182 \end{aligned}$$

$H(X_i)$ is the entropy rate of the memoryless model. For the two Markov models we can find the entropy rates using the chain rule:

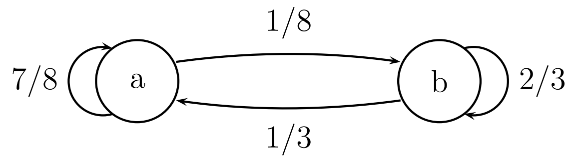
$$\begin{aligned} H(X_{i+1}|X_i) &= H(X_i, X_{i+1}) - H(X_i) \approx 0.6458 \\ H(X_{i+2}|X_i, X_{i+1}) &= H(X_i, X_{i+1}, X_{i+2}) - H(X_i, X_{i+1}) \approx 0.6271 \end{aligned}$$

The conditional probabilities are given by

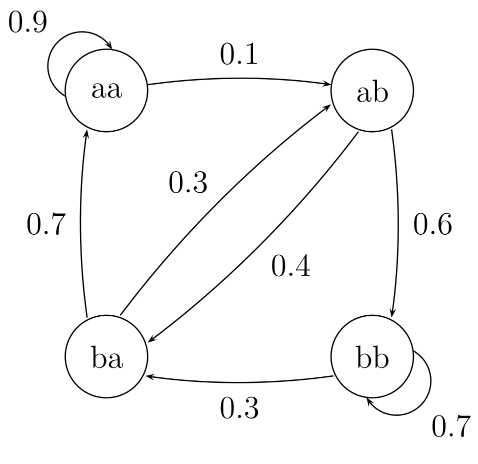
$$p(x_{i+1}|x_i) = \frac{p(x_i, x_{i+1})}{p(x_i)}$$

$$p(x_{i+2}|x_i, x_{i+1}) = \frac{p(x_i, x_{i+1}, x_{i+2})}{p(x_i, x_{i+1})}$$

The state model for the order 1 model looks like



Given states (x_i, x_{i+1}), the state diagram for the order 2 model looks like



5 A Huffman code for the distribution gives the mean codeword length $\bar{l} \approx 2.1545$ bits/codeword and average data rate $R = \frac{\bar{l}}{3} \approx 0.7182$ bits/symbol.

6 Under the assumption that the subintervals are always ordered in the same order as in the alphabet, we will get the interval $[0.8512\ 0.82992)$ with size 0.001792. (Simple check: $0.2 \cdot 0.8 \cdot 0.8 \cdot 0.1 \cdot 0.2 \cdot 0.7 = 0.001792$).

We will need at least $\lceil -\log_2 0.001792 \rceil = 10$ bits in our codeword, maybe one more.

Write the two limits as binary numbers:

$$\begin{aligned} 0.8512 &= 0.11011001111010000011\ldots \\ 0.852992 &= 0.11011010010111011010\ldots \end{aligned}$$

The smallest number with ten bits in the interval is 0.1101101000. All numbers that start with these bits are also inside the interval (ie smaller than the upper limit). Ten bits will therefore be enough.

The codeword is thus **1101101000**.

7 a) The history buffer size is $256 = 2^8$. The alphabet size is $16 = 2^4$, If we code a single symbol we will use a total of $1 + 4 = 5$ bits and if we code a match we will use a total of $1+8+4 = 13$ bits. This means that it is better to code matches of length 1 and 2 as single symbols. Since we use 4 bits for the lengths, we can then code match lengths between 3 and 18. The codeword for length l is the 4 bit binary representation of $l - 3$.

b) Resulting codewords. Offset 0 is the most recent position in the buffer.

f	o	l	c	codeword	sequence
0			b	0 0001	b
0			a	0 0000	a
0			d	0 0011	d
1	2	4		1 00000010 0001	badb
0			e	0 0100	e
0			p	0 1111	p
1	0	4		1 00000000 0001	pppp
1	8	7		1 00001000 0100	adbeppp
1	3	3		1 00000011 0000	epp
0			0	0 1110	o