LINKÖPING UNIVERSITY

Solutions to Written Exam in
Data compression
TSBK08

9th June 2022

1 a) See the course literature.

b) See the course literature.

c) See the course literature.

d) See the course literature.

2 See the course literature.

3 A Huffman code for the distribution gives the mean codeword length $\bar{l} = 2.99$ bits/codeword and average data rate $R = \bar{l} = 2.99$ bits/symbol.

For comparison, the entropy rate of the source is $H(X_i) \approx 2.9645$.

4 Inverse mtf gives the vector $L = [bbbddaaa]$.

Sort the sequence to get the vector $F = [aaabbbdd]$ which gives us the vector $T = [5\ 6\ 7\ 0\ 1\ 2\ 3\ 4]$ (the position in L where you find each symbol in F).

Inverse BWT gives the sequence *adbadbab*.

5 a) Stationary probabilities $p(x_i, x_{i-1})$ for the states (pairs of symbols):

$$p(aa) = \frac{12}{23}, \quad p(ab) = \frac{2}{23}, \quad p(ba) = \frac{2}{23}, \quad p(bb) = \frac{7}{23}$$

From this distribution we calculate $H(X_i, X_{i-1}) \approx 1.6248$.

The marginal distribution gives us the probabilities for single symbols

$$p(a) = p(aa) + p(ab) = \frac{14}{23}, \quad p(b) = p(ba) + p(bb) = \frac{9}{23}$$

From this distribution we calculate $H(X_i) \approx \underline{0.9656}$. We also get $H(X_i|X_{i-1}) = H(X_i, X_{i-1}) - H(X_{i-1}) \approx \underline{0.6592}$.

Probabilities for triples are given by
$p(x_i, x_{i-1}, x_{i-2}) = p(x_{i-1}, x_{i-2}) \cdot p(x_i|x_{i-1}x_{i-2})$

$$p(aaa) = \frac{108}{230}, \; p(baa) = \frac{12}{230}, \; p(aab) = \frac{12}{230}, \; p(bab) = \frac{8}{230}$$

$$p(aba) = \frac{6}{230}, \; p(bba) = \frac{14}{230}, \; p(abb) = \frac{14}{230}, \; p(bbb) = \frac{56}{230}$$

From this distribution we calculate $H(X_i, X_{i-1}, X_{i-2}) \approx 2.2503$.

We can rewrite

$$\begin{aligned} H(X_i, X_{i+1}, X_{i+2}, X_{i+3}) &= H(X_i, X_{i+1}, X_{i+2}) + H(X_{i+3}|X_i, X_{i+1}, x_{i+2}) = \\ &= H(X_i, X_{i+1}, X_{i+2}) + H(X_{i+3}|X_{i+1}, x_{i+2}) = \\ &= 2 \cdot H(X_i, X_{i+1}, X_{i+2}) - H(X_i, X_{i+1}) \approx \\ &\approx \underline{2.8758} \end{aligned}$$

where we used the fact that source is of order 2 and stationary.

b) Assuming that we start in state aa and that we always place the a interval closest to 0, the interval belonging to the sequence is $[0.67797 \; 0.688176)$. The interval has the size 0.010206.
We thus need at least $\lceil -\log_2 0.010206 \rceil = 7$ bits in our codeword.

The smallest seven bit number inside the interval is $(0.1010111)_2 = 0.6796875$. We check the largest number starting with these bits: $(0.1010111111111\ldots)_2 = (0.1011)_2 = 0.6875 < 0.688716$.
It is enough to use seven bits.

The codeword is 1010111.

6 The decoded sequence is

$$susssussurrrrtussrrrtu\ldots$$

and the dictionary looks like

index	word	index	word	index	word	index	word
0	r	4	su	8	uss	12	rt
1	s	5	us	9	sur	13	tu
2	t	6	ss	10	rr	14	$ussr$
3	u	7	ssu	11	rrr	15	$rrrt$

and the next word to add to position 16 is $tu*$, where $*$ will be the first symbol in the next decoded word.

7 The differential entropy is given by

$$\begin{aligned} h(X) &= -\int_{-\infty}^{\infty} f(x)\log f(x)dx \\ &= -\int_0^1 \frac{1}{2}\log\frac{1}{2}dx - \int_1^2 \frac{1}{3}\log\frac{1}{3}dx - \int_2^3 \frac{1}{6}\log\frac{1}{6}dx \\ &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{3}\log\frac{1}{3} - \frac{1}{6}\log\frac{1}{6} \\ &= \frac{\log 432}{6} \approx 1.4591 \quad \text{[bits]} \end{aligned}$$