LINKÖPING
UNIVERSITY

Solutions to Written Exam in
Data compression
TSBK08

26th August 2023

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.

2 If X takes values in the alphabet $\{a_1, a_2, \ldots, a_L\}$ the entropy is given by

$$H(X) = -\sum_{i=1}^{L} p(a_i) \cdot \log p(a_i)$$

The left inequality comes from

$$-p(a_i) \cdot \log p(a_i) \begin{cases} = 0, & p(a_i) = 0 \\ > 0, & 0 < p(x_i) < 1 \\ = 0, & p(a_i) = 1 \end{cases}$$

Thus $H(X) \geq 0$ with equality if and only if $p(a_i)$ is either 0 or 1 for every i, but then we must have that $p(a_i) = 1$ for exactly one i.

The right inequality comes from

$$\begin{aligned} H(X) - \log L &= -\sum_{i=1}^{L} p(a_i) \log p(a_i) - \log L \\ &= \sum_{i=1}^{L} p(a_i) \log \frac{1}{L \cdot p(a_i)} \\ &\leq \sum_{i=1}^{L} p(a_i) (\frac{1}{L \cdot p(a_i)} - 1) \log e \\ &= (\sum_{i=1}^{L} \frac{1}{L} - \sum_{i=1}^{L} p(a_i)) \log e \\ &= (1 - 1) \log e = 0 \end{aligned}$$

with equality if and only if $p(a_i) = \frac{1}{L}$ for all $i = 1, \ldots, L$. This inequality can also be proven by regular Lagrange minimization techniques.

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

$$p(aa) = 3/12, \quad p(ab) = 1/12, \quad p(ba) = 1/12, \quad p(bb) = 7/12$$

From this distribution we get $H(X_i, X_{i-1}) \approx 1.5511$.

The marginal distribution gives us probabilities for single symbols

$$p(a) = p(aa) + p(ab) = 1/3, \quad p(b) = p(ba) + p(bb) = 2/3$$

From this distribution we get $H(X_i) \approx \underline{0.9183}$. Use the chain rule to get $H(X_i|X_{i-1}) = H(X_i, X_{i-1}) - H(X_{i-1}) \approx \underline{0.6328}$.

Finally, $H(X_i|X_{i-1}, X_{i-2})$ can be calculated as

$$\begin{aligned} H(X_i|X_{i-1}, X_{i-2}) &= 3/12 \cdot (-0.8 \cdot \log 0.8 - 0.2 \cdot \log 0.2) + \\ & \quad 1/12 \cdot (-0.6 \cdot \log 0.6 - 0.4 \cdot \log 0.4) + \\ & \quad 1/12 \cdot (-0.3 \cdot \log 0.3 - 0.7 \cdot \log 0.7) + \\ & \quad 7/12 \cdot (-0.1 \cdot \log 0.1 - 0.9 \cdot \log 0.9) \\ &\approx 0.6084 \end{aligned}$$

Alternatively, calculate the entropy $H(X_i, X_{i-1}, X_{i-2})$ and then $H(X_i|X_{i-1}, X_{i-2}) = H(X_i, X_{i-1}, X_{i-2}) - H(X_{i-1}, X_{i-2})$.

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

$p(aaa) = 24/120$, $p(baa) = 6/120$, $p(aab) = 6/120$, $p(bab) = 4/120$

$p(aba) = 3/120$, $p(bba) = 7/120$, $p(abb) = 7/120$, $p(bbb) = 63/120$

A Huffman code (the codeword lengths are not unique) for triples can look like this

symbols	codeword	symbols	codeword
aaa	00	baa	01000
aab	01001	bab	01010
aba	01011	bba	0110
abb	0111	bbb	1

The average codeword length for this code is $\bar{l} = 262/120 \approx 2.1833$ bits/codeword, which gives the rate $R \approx 0.7278$ bits/symbol.

4 Assuming we use the alphabet order for the intervall order, with the x-interval closest to 0, the sequence corresponds to the intervall

$$[0.6972, \quad 0.701088)$$

with the interval size 0.003888. We will need at least $\lceil -\log_2 0.003888 \rceil = 9$ bits in our codeword, maybe one more.

Write the two interval limits as binary numbers:

$$\begin{aligned} 0.6972 &= 0.101100100111\ldots \\ 0.701088 &= 0.101100110111\ldots \end{aligned}$$

The smallest nine bit number inside the interval is 0.101100101, and all numbers starting with these bits are also inside the interval (ie smaller than the upper interval limit). Thus, nine bits are enough.

The codeword is **101100101**.

5 The decoded sequence is

$$bababababahehehehagdagda\ldots$$

and the dictionary looks like

index	word	index	word	index	word
0	*a*	8	*ba*	16	*heha*
1	*b*	9	*ab*	17	*ag*
2	*c*	10	*bab*	18	*gd*
3	*d*	11	*baba*	19	*da*
4	*e*	12	*ah*	20	*agd*
5	*f*	13	*he*	21	*da*∗
6	*g*	14	*eh*	22	
7	*h*	15	*heh*	23	

where ∗ in word 21 will be the first symbol in the next decoded word.

6 Create all cyclic shifts of the sequence and sort them

Cyclic shifts	Sorted shifts
cabcabbcad	*abbcadcabc*
abcabbcadc	*abcabbcadc*
bcabbcadca	*adcabcabbc*
cabbcadcab	*bbcadcabca*
abbcadcabc	*bcabbcadca*
bbcadcabca	*bcadcabcab*
bcadcabcab	*cabbcadcab*
cadcabcabb	*cabcabbcad*
adcabcabbc	*cadcabcabb*
dcabcabbca	*dcabcabbca*

The result of the transform is the sequence *cccaabbdba* (the last column of the sorted list) and the position in the sorted list where we find the original sequence, ie 7 (assuming we start counting at 0).

The mtf coding gives the sequence

$$2, 0, 0, 1, 0, 2, 0, 3, 1, 2$$

(assuming that the starting symbol list has the same order as the alphabet.)

7 The differential entropy is given by

$$\begin{aligned}
h(X) &= -\int_{-\infty}^{\infty} f(x)\log f(x)\ dx = -\int_0^1 (2-2x)\log(2-2x)\ dx = \\
&= \Big/ y = 2-2x\ ,\ dy = -2dx \Big/ = -\frac{1}{2}\log e \int_0^2 y \ln y\ dy = \\
&= -\frac{1}{2}\log e\big(2\ln 2 - 1\big) = \frac{1}{2}\log e - 1 \approx -0.2786 \text{ bits}
\end{aligned}$$