Information and Coding Theory CO349

Tutorial - Answers

Sheet 2

Exercise 1 (Biggs Ex 3.13) *Construct a binary Shannon-Feno Code for the source* $(S, \mathbf{p})$ *with* $S = \{s_1, s_2, s_3, s_4, s_5, s_6\}$ *and*

$$\mathbf{p} = (0.25, 0.10, 0.15, 0.05, 0.20, 0.25).$$

What is the entropy and what is the average word-length?

Solution

- For s_1 we have that y_1 is the least integer $2^{y_1} \geq \frac{1}{0.25} = 4$, i.e. $y_1 = 2$.
- For s_2 the least integer such that $2^{y_2} \geq \frac{1}{0.10} = 10$, i.e. $y_2 = 4$.
- For s_3 the least integer such that $2^{y_3} \geq \frac{1}{0.15} = \frac{100}{15} = 6\frac{2}{3}$, i.e. $y_3 = 3$.
- For s_4 the least integer such that $2^{y_4} \geq \frac{1}{0.05} = 20$, i.e. $y_4 = 5$.
- For s_5 the least integer such that $2^{y_5} \geq \frac{1}{0.20} = 5$, i.e. $y_5 = 3$.
- For s_6 we have that y_6 is the least integer $2^{y_6} \geq \frac{1}{0.25} = 4$, i.e. $y_6 = 2$.

A possible code is

$$\begin{array}{llll} s_1 & \mapsto 00 & s_4 & \mapsto 01011 \\ s_2 & \mapsto 0100 & s_5 & \mapsto 100 \\ s_3 & \mapsto 101 & s_6 & \mapsto 11 \end{array}$$

Average word-length $L = 2.7$. Entropy: $\mathbf{H}(\mathbf{p}) \approx 2.42$.

Exercise 2 (Biggs Example 4.1,Slide 26) *Consider the source* $(S, \mathbf{p})$ *and different block codes.*

- *single symbol code* $c : \mathbb{B} \to \mathbb{B}^*$ *with distribution:* $\mathbf{p}_1 = (0.9, 0.1)$ *on* $S = \mathbb{B} = \{0, 1\}$.

- *two symbol code* $c : \mathbb{B}^2 \to \mathbb{B}^*$ *with distribution* $\mathbf{p}_2 = (0.81, 0.09, 0.09, 0.01)$ *on* $S = \mathbb{B} \times \mathbb{B} = \{00, 01, 10, 11\}$.

- *three symbol code* $c : \mathbb{B}^3 \to \mathbb{B}^*$ *with distribution* $\mathbf{p}_3 = (0.729, 0.081, 0.081, 0.009, 0.081, 0.009, 0.009, 0.001)$ *on* $S = \mathbb{B} \times \mathbb{B} \times \mathbb{B}$.

- *etc.*

What is the entropy, the average code length and a Huffman code for each of these cases? What is the relationship between $\mathbf{p}_i$*'s?*

Solution

- single symbol code $c : \mathbb{B} \to \mathbb{B}^*$ with distribution: $\mathbf{p} = (0.9, 0.1)$ on $\{0, 1\}$ has entropy $\mathbf{H}(\mathbf{p}_1) \approx 0.469$ average codeword-length $L_1 = 1$ for code $c = \text{id}$.

- two symbol code $c : \mathbb{B}^2 \to \mathbb{B}^*$ with distribution $\mathbf{p}_2 = (0.81, 0.09, 0.09, 0.01)$ which allows an optimal Huffman code c

  $$00 \mapsto 0, \quad 01 \mapsto 10, \quad 10 \mapsto 110, \quad 11 \mapsto 111$$

  with $L_2 = 1.29$ or $\frac{L_2}{2} = 0.645$ and $\mathbf{H}(\mathbf{p}_2) \approx 0.938$.

- three symbol code $c : \mathbb{B}^3 \to \mathbb{B}^*$ with distribution $\mathbf{p}_3 = (0.729, 0.081, 0.081, 0.009, 0.081, 0.009, 0.009, 0.001)$ with optimal Huffman code

  $$\begin{array}{lll} 000 & \mapsto & 0 \\ 001 & \mapsto & 100 \\ 010 & \mapsto & 101 \\ 011 & \mapsto & 11100 \end{array} \qquad \begin{array}{lll} 100 & \mapsto & 110 \\ 101 & \mapsto & 11101 \\ 110 & \mapsto & 11110 \\ 111 & \mapsto & 11111 \end{array}$$

  we get $L_3 = 1.598$, or $\frac{L_3}{3} = 0.533$ and $\mathbf{H}(\mathbf{p}_3) \approx 1.407$. (NB there is a typo in Bigg's code).

Probabilities are obtained by "splitting", e.g. $\sum_{i=1}^{4}(\mathbf{p}_3)_i = \sum_{j=1}^{2}(\mathbf{p}_2)_j = (\mathbf{p}_1)_1$, etc. i.e. we have to consider marginal distributions.

Exercise 3 (Biggs, Ex 3.15) *Consider again the source from Example 1* $(S, \mathbf{p})$ *with* $S = \{s_1, s_2, s_3, s_4, s_5, s_6\}$ *and*

$$\mathbf{p} = (0.25, 0.10, 0.15, 0.05, 0.20, 0.25).$$

Construct the Huffman Code. What is the average word-length? How are entropy and the average word-length for the Shannon-Fano and Huffamn Code related?

Solution The Huffman construction is based on the following splitting

$p^{(1)}:$	$\frac{1}{4}:s_1$	$\frac{1}{10}:s_2$	$\frac{3}{20}:s_3$	$\frac{1}{20}:s_4$	$\frac{1}{5}:s_5$	$\frac{1}{4}:s_6$
$p^{(2)}:$	$\frac{1}{4}:s_1$	$\frac{3}{20}:[s_2s_4]$	$\frac{3}{20}:s_3$		$\frac{1}{5}:s_5$	$\frac{1}{4}:s_6$
$p^{(3)}:$	$\frac{1}{4}:s_1$		$\frac{3}{10}:[s_3[s_2s_4]]$		$\frac{1}{5}:s_5$	$\frac{1}{4}:s_6$
$p^{(4)}:$	$\frac{1}{4}:s_1$		$\frac{3}{10}:[s_3[s_2s_4]]$		$\frac{9}{20}:[s_5s_6]$	
$p^{(5)}:$			$\frac{11}{20}:[s_1[s_3[s_2s_4]]]$		$\frac{9}{20}:[s_5s_6]$	
$p^{(6)}:$			$1:[[s_1[s_3[s_2s_4]]][s_5s_6]]$			

The Huffman code (one option) is

$$
\begin{array}{llll}
s_1 & \mapsto 01 & s_4 & \mapsto 0010 \\
s_2 & \mapsto 0011 & s_5 & \mapsto 10 \\
s_3 & \mapsto 000 & s_6 & \mapsto 11
\end{array}
$$

The average code length is $L = 2.45$.

Exercise 4 (Cover&Thomas Ex 5.6.1) *Consider a source with*

$$\mathbf{p} = (0.25, 0.25, 0.2, 0.15, 0.15).$$

Construct a Huffman code and give the average word-length compared to the entropy.

Solution The presentation format of how to construct the Huffman code varies for different monographs/authors. In Cover&Thomas we have e.g.

Length	Code	s_i	Probabilities				
2	01	1	0.25	0.30	0.45	0.55	1
2	10	2	0.25	0.25	0.30	0.45	
2	11	3	0.20	0.25	0.25		
3	000	4	0.15	0.20			
3	001	5	0.15				

where in particular the order on probabilities is always monotone. Average word-length $L = 2.3$ and $\mathbf{H}(\mathbf{p}(\mathbf{p}) \approx 2.2855$.

Exercise 5 (Cover&Thomas Ex 5.6.2/3) *Try to generalise the Huffman rules for tenary codes, i.e.* $T = \{0, 1, 2\}$. *Construct the tenary Huffman code for*

$$\mathbf{p}_1 = (0.25, 0.25, 0.2, 0.15, 0.15)$$

and

$$\mathbf{p}_2 = (0.25, 0.25, 0.2, 0.1, 0.1, 0.1).$$

and their average word-length.

Solution For tenary code we could describe the procedure as follows:

H1 Given a source $(S, \mathbf{p})$ and be s', s'' and s''' be symbols with smallest probability. Construct a new source $(S^*, \mathbf{p}^*)$ by replacing s', s'' and s'' with a new (super-)symbol s^* with probability $p_{s^*} = p_{s'} + p_{p''} + p_{s'''}$.

Nothing changes for other symbols.

H2 If we are given a PF binary code h^* for $(S^*, \mathbf{p}^*)$ with $h^* : s^* \mapsto w$ then define a tenary code h for $(S, \mathbf{p})$ with $h : s' \mapsto w0$, $h : s'' \mapsto w1$ and $h : s''' \mapsto w2$.

Nothing changes for other symbols.

Length	Code	s_i	Probabilities		
1	1	1	0.25	0.50	1
1	2	2	0.25	0.25	
2	00	3	0.20	0.25	
2	01	4	0.15		
2	02	5	0.15		

Average (tenary) word-length is $L = 1.5$ while entropy is $\mathbf{H}_3(\mathbf{p}_1) \approx 1.4420$ (compared with $\mathbf{H}_2(\mathbf{p}_1) \approx 2.2855$).

It might be the case that there are not enough concrete symbols in the source, as is the case for $\mathbf{p}_2$. In this situation one can introduce"dummy" symbols with probability zero.

The number of concrete symbols should be $1 + k(b-1)$ for some k as in each stage we reduce the number of symbols in $(S^*, \mathbf{p}^*)$ by $b - 1$ symbols.

Length	Code	s_i	Probabilities			
1	1	1	0.25	0.25	0.50	1
1	2	2	0.25	0.25	0.25	
2	01	3	0.20	0.20	0.25	
2	02	4	0.10	0.20		
3	000	5	0.10	0.10		
3	001	5	0.10			
3	002	—	0.00			

Average (tenary) word-length is $L = 1.7$. The entropy is given by $\mathbf{H}_3(\mathbf{p}_2) \approx 1.5527$ (and $\mathbf{H}_2(\mathbf{p}_2) \approx 2.4610$).

Exercise 6 *Consider the alphabet* $S = \{U, V, W, X, Y, Z\}$ *and LZW coding (data compression). Use LZW to encode the following message:*

$$WWWWXWWWXWWXW$$

Specify the directory D_m *for each step.*

Solution

$\mathbf{W}WWWXWWWXWWXW$		U, V, W, X, Y, Z
$\mathbf{WW}WXWWWXWWXW$	3	U, V, W, X, Y, Z, WW
$\mathbf{W}XWWWXWWXW$	3 7	$U, V, W, X, Y, Z, WW, WWW$
$\mathbf{X}WWWXWWXW$	3 7 3	$U, V, W, X, Y, Z, WW, WWW, WX$
$\mathbf{WWW}XWWXW$	3 7 3 4	$U, V, W, X, Y, Z, WW, WWW, WX, XW$
$\mathbf{XW}WXW$	3 7 3 4 8	$U, V, W, X, Y, Z, WW, WWW, WX, XW, WWWX$
$\mathbf{WX}W$	3 7 3 4 8 10	$U, V, W, X, Y, Z, WW, WWW, WX, XW, WWWX, XWW$
$\mathbf{W}$	3 7 3 4 8 10 9	$U, V, W, X, Y, Z, WW, WWW, WX, XW, WWWX, XWW, WXW$
	3 7 3 4 8 10 9 3	