E4.40/SO20 – Information Theory

Problem Sheet 2

(Most questions are from Cover & Thomas, the corresponding question numbers are given in brackets at the start of the question)

Notation: We use a sans-serif font for random variables: x, $\mathbf{x}$, $\mathbf{X}$ are scalar, vector and matrix random variables respectively.

1. Use the Kraft inequality to show that it is possible to construct a 4-ary prefix code with lengths {1, 1, 2, 2, 2, 2, 2, 2, 2}. Construct such a code for symbols that take the values A, B, …, H, I with probabilities {.15, .15, .1, .1, .1, .1, .1, .1, .1}.

   Calculate the entropy of the input symbols and the expected length of the codewords.

2. The four symbols A, B, C, D are encoded using the following sets of codewords. In each case state whether the code is (i) non-singular, (ii) uniquely decodable and (iii) a prefix code.

   (a) {1, 01, 000, 001}

   (b) {0, 10, 000, 100}

   (c) {01, 01, 110, 100}

   (d) {0, 01, 011, 0111}

   (e) {10, 10, 0010, 0111}

3. In a complete *D*-ary code tree, the end of each branch is either a leaf node of else has *D* sub-branches. Show that the total number of leaf nodes is one more than a multiple of *D*–1.

4. For each value of *D* given below, find a *D*-ary Huffman code for the probability vector {0.25, 0.2, 0.15, 0.1, 0.1, 0.1, 0.1}. In each case calculate the expected code length. [Note that, to ensure a full code tree, you should add zero-probability symbols so that the total number of symbols is one more than a multiple of *D*–1]

   (a) D = 3

   (b) D = 4

   (c) D = 5

5. Suppose the input alphabet is $\mathcal{X}$ = {1, 2, 3, …, 100}. If *C* is a prefix code for $\mathcal{X}$, show that the sum of the 100 codeword lengths must exceed 664.

6. Find a binary prefix code with codeword lengths {2, 2, 3, 3, 3, 4, 4}. Find a probability mass vector for which the expected length of this code is equal to the entropy of the source.

7. For each of the following transition matrices, draw the state diagram and say whether the corresponding Markov process is (i) irreducible and (ii) aperiodic. Find all the possible stationary distributions for the process and determine the entropy rate corresponding to each of them.

   (a) $T = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$ (b) $T = \begin{pmatrix} 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 \end{pmatrix}$ (c) $T = \begin{pmatrix} 0.5 & 0 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$

   (d) $T = \begin{pmatrix} 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$ (e) $T = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$

8. [4.2] If $\{x_i\}$ is a stationary stochastic process, show that

   $$H(x_i \mid x_{i-1}, x_{i-2}, \cdots, x_{i-n}) = H(x_i \mid x_{i+1}, x_{i+2}, \cdots, x_{i+n})\,.$$

   In other words, the conditional entropy given the previous *n* samples is the same as the conditional entropy given the next *n* samples.

9. (a) [4.5] Determine the stationary distribution and the entropy rate, $H(\mathcal{X})$, of a Markov process with two states, 0 and 1 and transition matrix

$$T = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

(b) Find the values of p and q that maximize $H(\boldsymbol{X})$.

(c) If q=1,

(i) find the maximum value of $H(\boldsymbol{X})$ and the value of p that attains it.

(ii) define $s(n)$ to be the number of sequences of length n with non-zero probability. If $s_0(n)$ and $s_1(n)$ are the number of sequences starting with 0 and 1 respectively, show that $s_1(n) = s_0(n-1) = s(n-2)$. Hence show that $s(n) = s(n-1) + s(n-2)$ and derive an explicit expression for $s(n)$.

(iii) explain why we must have $H(\boldsymbol{X}) \le \lim_{n\to\infty}\left(n^{-1}\log s(n)\right)$ and calculate the value on the right hand side.

10. If $x_i \in \{A, B\}$ are i.i.d. with probability mass vector $\{0.9, 0.1\}$, determine $H(x_i)$.

Using binary Huffman codes, determine the coding redundancy (i.e. the difference between the entropy and the number of bits used per symbol) when (a) each x_i is encoded individually, (b) pairs of x_i are coded together (i.e. x_1x_2 followed by x_3x_4 etc) and (c) triplets of x_i are coded together.

11. Arithmetic coding is used to encode a sequence of Bernoulli symbols, x_i , with $\boldsymbol{X} = [\text{A, B}]^T$, $\mathbf{p} = [0.3, 0.7]^T$. For a sequence of length n , we define $Q_r^{(n)} = \sum_{j\le r} p(\mathbf{x}_j^{(n)})$ with $Q_0^{(n)} = 0$ where the $\mathbf{x}_j^{(n)}$ (with $j \ge 1$) are the 2^n possible sequences of length n arranged in lexical (alphabetical) order. A received code is decoded as $\mathbf{x}_j^{(n)}$ iff it lies in the range $Q_{j-1}^{(n)} \le \mathbf{x}_j^{(n)} < Q_j^{(n)}$.

[Note: To convert a decimal fraction to binary, multiply by 2 repeatedly and record the modulo-2 value of the integer part of each result as successive answer bits.]

(a) Determine the value of $Q_j^{(2)}$ for $0 \le j \le 4$ as decimal numbers and as binary fractions with five bits after the binary point.

(b) Calculate the codes transmitted for each of the four possible 2-symbol input sequences.

(c) Using the codes of part (b), calculate the expected number of transmitted bits per symbol and the value of $H(x_i)$.

(d) Determine how many output symbols can be decoded with certainty if the code 0111… is received.

E4.40/SO20 – Information Theory

Solution Sheet 2

1. Kraft Inequality: $2\times 4^{-1}+7\times 4^{-2}=0.9375\le 1$ so a prefix code is possible. A suitable code is {0, 1, 20, 21, 22, 23, 30, 31, 32}.

   Average length of code is 1.7 (equivalent to 3.4 bits since we are using a 4-ary code).

   Entropy of input is 3.1464 bits.

2. Note that Prefix ⇒ Uniquely Decodable ⇒ Non-singular

   (a) {1, 01, 000, 001} is a prefix code (and therefore uniquely decodable and non-singular).

   (b) {0, 10, 000, 100} is non-singular since no two codewords are the same. However it is not uniquely decodable since AAA and C are indistinguishable (it follows that it is not a prefix code either).

   (c) {01, 01, 110, 100} is singular since A and B have the same codeword (hence not uniquely decodable and not a a prefix code).

   (d) {0, 01, 011, 0111} is uniquely decodable: when you get a 0 it is the start of a new symbol and the previous symbol is given by counting how many 1's since the previous 0. It is not a prefix code since the first codeword is a prefix of all the others.

   (e) {10, 01, 0010, 0111} is not a prefix code since 01 is a prefix of 0111. Since all codewords have even length we can rewrite it as a 4-ary code {2, 1, 02, 13}. It is easy to see that this is uniquely decodable since 3 only ever appears as the second half of D. Thus if a 1 is followed by anything other than 3, it represents a B.

3. You can create the tree incrementally from an initial tree with only one node. At each stage, you replace an existing leaf node with a set of *D* sub-branches which end in leaf nodes. This process adds *D* new leaf nodes but removes one that existed previously. It therefore adds *D*–1 to the total. Initially, we only had one leaf node and so we will always have one more than a multiple of *D*–1.

4. (a) D = 3. We do not need to append any zero-probability symbols since 7 is one more that a multiple of D–1 already. The codes we get are {2, 00, 01, 02, 10, 11, 12}. The expected length is 1.75.

   (b) D = 4. Again 7 is one more than a multiple of D-1 so we do not need to add any symbols. The codes we get are {1, 2, 3, 00, 01, 02, 03}. The expected length is 1.4.

   (c) D = 5. We need to two zero-probability symbols to make the number of symbols one more than a multiple of 4. The codes we get are then {1, 2, 3, 4, 00, 01, 02, [03, 04]} where the last two codes are in brackets because they are not used. Not that adding the zero-probability symbols ensures that the unused codes are the longest ones. The expected length is now 1.3.

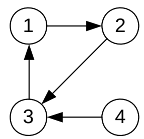
5. If all 100 symbols have an equal probability of 0.01, then $H(X)=\log_2 100=6.644\text{ bits}$. The average codeword length is then the sum of the lengths divided by 100 which must exceed 6.64. Hence the sum of the lengths must exceed 664. Note that choosing unequal probabilities will give a lower entropy and hence a reduced lower bound for the sum of the lengths; this is also valid but is a weaker constraint.

6. A suitable code is {00, 01, 100, 101, 110, 1110, 1111}.

   To attain the source entropy, all the symbol probabilities must equal 2^{-l}. This gives us {0.25, 0.25, 0.125, 0.125, 0.125, 0.0625, 0.0625} as the probabilities.

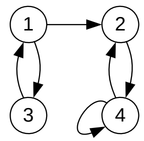
7.

   (a)



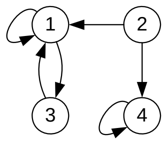
   Reducible since no path from 2 to 4.
   Periodic since 1 to 1 always takes a multiple of 3 steps.
   There is one stationary distribution: $[1\ 1\ 1\ 0]^T/3$.
   $H(\boldsymbol{X})$ = 0 bits (implying that knowledge of the previous state tells you the current state).

   (b)



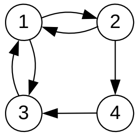
   Reducible since no path from 2 to 1.
   Periodic since 1 to 1 always takes a multiple of 2 steps.
   There is one stationary distribution: $[0\ 1\ 0\ 2]^T/3$.
   $H(\boldsymbol{X})$ = 0.6667 bits.

(c)



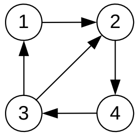
Reducible since no path from 1 to 2.
Aperiodic since, for example, 3 to 3 can take either 2 or 3 (or 4, …) steps and the highest common factor of these is 1.
There are two stationary distributions: $[2\ 0\ 1\ 0]^T/3$ and $[0\ 0\ 0\ 1]^T$. $H(\boldsymbol{X})$ = 0.6667 bits and 0 bits respectively.

(d)



Irreducible since you can to from anywhere to anywhere.
Periodic since 1 to 1 always takes an even number of steps.
There is one stationary distribution: $[4\ 2\ 3\ 1]^T/10$. $H(\boldsymbol{X})$ = 0.6 bits.

(e)



Irreducible since you can to from anywhere to anywhere.
Aperiodic since, for example, 3 to 3 can take either 3 or 4 steps and the highest common factor of these is 1.
There is one stationary distribution: $[1\ 2\ 2\ 2]^T/7$. $H(\boldsymbol{X})$ = 0.2857 bits.

8. We can write the following:

$$\begin{aligned} H(x_i \mid x_{i-1}, x_{i-2}, \cdots, x_{i-n}) &\overset{(a)}{=} H(x_i, x_{i-1}, x_{i-2}, \cdots, x_{i-n}) - H(x_{i-1}, x_{i-2}, \cdots, x_{i-n}) \\ &\overset{(b)}{=} H(x_{i+n}, x_{i+n-i}, x_{i+n-2}, \cdots, x_i) - H(x_{i+n}, x_{i+n-1}, \cdots, x_{i+1}) \\ &\overset{(c)}{=} H(x_i \mid x_{i+1}, x_{i+2}, \cdots, x_{i+n}) \end{aligned}$$

Where (a) and (c) follow from the definition of conditional entropy and (b) follows from stationarity: we add n onto the time index of the first term and $n+1$ onto the time index of the second.

9. (a) To find the stationary distribution, we write

$$T^T \begin{pmatrix} a \\ 1-a \end{pmatrix} = \begin{pmatrix} 1-p & q \\ p & 1-q \end{pmatrix} \begin{pmatrix} a \\ 1-a \end{pmatrix} = \begin{pmatrix} a \\ 1-a \end{pmatrix}$$
$$\Rightarrow \quad (1-p)a + q(1-a) = (1-p-q)a + q = a$$
$$\Rightarrow \quad a = q(p+q)^{-1}$$

It follows that the entropy rate is given by

$$H(\boldsymbol{X}) = aH(p) + (1-a)H(q) = (qH(p) + pH(q))(p+q)^{-1}$$

This is a weighted average of $H(p)$ and $H(q)$ with the weights equal to their frequencies in the stationary distribution.

(b) We know that $H(\boldsymbol{X}) = H(x_n \mid x_{n-1}) \le H(x_n) \le 1$

If $p = q = 0.5$ then $H(\boldsymbol{X}) = H(p) = H(q) = 1$ and this is the maximum attainable.

(c-i) Substituting $q = 1$ into the previous answer gives $H(\boldsymbol{X}) = (p+1)^{-1} H(p)$. Differentiating and setting this to zero gives $H(p) = (p+1)H'(p)$ from which we get $p = (1-p)^2$ and $p = 1.5 - \sqrt{1.25} = 0.382$. This gives $H(\boldsymbol{X}) = 0.694$ bits.

(c-ii) If the first bit is 0 the next can be either 0 or 1 which means that $s_0(n) = s(n-1)$. However $q = 1$, if the first bit of a sequence is 1, then the next must be 0. It follows that $s_1(n) = s_0(n-1) = s(n-2)$. Hence $s(n) = s_0(n) + s_1(n) = s(n-1) + s(n-2)$ with the initial conditions $s(1) = 2$ and $s(2) = 3$. The solutions to the recurrence relation (which is that of the Fibonacci series for those interested in such things) have the form $s(n) = a^n$ where a satisfies $a^2 = a + 1 \quad \Rightarrow \quad a = 0.5 \pm \sqrt{1.25} = \{-0.618, 1.618\}$. Imposing the initial conditions gives $s(n) = -0.17 \times -0.618^n + 1.17 \times 1.618^n$. For large n, the first term is insignificant and $s(n) \approx 1.17 \times 1.618^n$.

(c-iii) By definition $H(\boldsymbol{X}) = \lim_{n \to \infty} \left(n^{-1} H(X_{1:n}) \right)$. An upper bound on $H(x_1, x_2, \ldots x_n)$ is given by $\log s(n)$ which is attained if all $s(n)$ possible sequences have equal probability. Thus $\lim_{n \to \infty} \left(n^{-1} \log s(n) \right) = \log 1.618 = 0.694$ is an upper bound for $H(\boldsymbol{X})$.

10. $H(x_i) = 0.469$ bits.

(a) $p\{A, B\} = \{0.9, 0.1\}$ giving Huffman lengths of $\{1, 1\}$ with $E(L) = 1$. Redundancy = 0.531 bits.

(b) $p\{AA, AB, BA, BB\} = \{0.81, 0.09, 0.09, 0.01\}$ giving Huffman lengths of $\{1, 2, 3, 3\}$ with $E(L)/2 = 1.29/2 = 0.645$ bits per input symbol. Redundancy = 0.176 bits.

(c) $p\{AAA, AAB, ABA, BAA, ABB, BAB, BBA, BBB\} = \{0.729, 0.081 * 3, 0.009 * 3, 0.001\}$ giving Huffman lengths of $\{1, 3, 3, 3, 5, 5, 5, 5\}$ with $E(L)/3 = 1.598/3 = 0.533$ bits per input symbol. Redundancy = 0.064 bits.

11. (a) $Q_{0:4}^{(2)} = [0 \quad 0.09 \quad 0.3 \quad 0.51 \quad 1] = [0.\dot{0} \quad 0.00010 \quad 0.01001 \quad 0.10000 \quad 1.\dot{0}]$.

(b) The code for $\mathbf{x}_j^{(n)}$ must lie in the range $Q_{j-1}^{(n)} \le \mathbf{x}_j^{(n)} < Q_j^{(n)}$. To obtain the length, l_j, of the code, we must find the first bit position at which $Q_{j-1}^{(n)}$ and $Q_j^{(n)}$ differ and then, (i) if all remaining bits of $Q_{j-1}^{(n)}$ are zero stop there, else (ii) carry on to the next 0 in $Q_{j-1}^{(n)}$. The code then consists of the first l_j bits of the binary expansion of $Q_{j-1}^{(n)}$ incremented by 0 or 1 for cases (i) and (ii) respectively.

This gives the following codes: $\mathbf{x}_1^{(2)} = \text{"AA"} = 0000$, $\mathbf{x}_2^{(2)} = \text{"AB"} = 001$, $\mathbf{x}_3^{(2)} = \text{"BA"} = 011$, $\mathbf{x}_4^{(2)} = \text{"BB"} = 11$.

(c) The probabilities vector $p\left(\mathbf{x}_{1:4}^{(2)}\right) = [0.09 \quad 0.21 \quad 0.21 \quad 0.49]^T$ with lengths [4 3 3 2]. This gives an average length of 2.6 bits per symbol pair or 1.3 bits per symbol.

$H(x_i) = H(0.3) = 0.8813$ bits which is quite a lot less.

(d) $0.0111\ldots$ must correspond to $0.4375 \le \mathbf{x}_j^{(n)} < 0.5$. From part (a), the first two symbols are BA. Dividing the corresponding interval $[0.3 \quad 0.51]$ in the ratio 0.3:0.7 gives $[0.3 \quad 0.3 + 0.3 \times (0.51 - 0.3) \quad 0.51] = [0.3 \quad 0.363 \quad 0.51]$ and, since we are in the top half of this, the next symbol must be B.

Dividing the interval $[0.363 \quad 0.51]$ in the ratio 0.3:0.7 gives $[0.363 \quad 0.4071 \quad 0.51]$ and, since we are in the top half of this, the next symbol must also be B.

Dividing the interval $[0.4071 \quad 0.51]$ in the ratio 0.3:0.7 gives $[0.4071 \quad 0.43797 \quad 0.51]$ and, since neither half contains our entire interval, we cannot yet be sure of the next symbol.

The answer, therefore, is that four symbols, BABB, can be decoded with certainty.