Final version -
July 07.

Paper Number(s): **E4.40**
C5.27
SO20
ISE4.51

IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE
UNIVERSITY OF LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2007

MSc and EEE/ISE PART IV: MEng and ACGI

INFORMATION THEORY

Friday, 4 May 10:00 am

There are SIX questions on this paper.

Answer FOUR questions.

All questions carry equal marks

Time allowed: 3:00 hours

Examiners responsible:

First Marker(s): D.M. Brookes

Second Marker(s): J.A. Barria

Information for Candidates:

Notation: (a) Random variables are shown in a sans serif typeface. Thus $\mathsf{x}, \mathbf{\mathsf{x}}, \mathbf{\mathsf{X}}$ denote a random scalar, vector and matrix respectively. The alphabet of a discrete random scalar, x, is denoted by $\boldsymbol{\mathcal{X}}$ and its size by $|\boldsymbol{\mathcal{X}}|$.

(b) $\mathsf{x}_{1:n}$ denotes the sequence $\mathsf{x}_1, \mathsf{x}_2, \cdots, \mathsf{x}_n$.

(c) The normal distribution function is denoted by:
$N(x;\mu,\sigma^2) = \left(2\pi\sigma^2\right)^{-½} \exp(-½(\mathrm{x}-\mu)^2\sigma^{-2})$

(d) $\oplus$ denotes the exclusive-or operation or, equivalently, addition modulo 2.

(e) $\log x = \dfrac{\ln x}{\ln 2}$ denotes logarithm to base 2.

(f) $P(\bullet)$ denotes the probability of the discrete event $\bullet$.

(g) "i.i.d." denotes "independent identically distributed"

The Questions

1. In this question, all vectors and matrices have binary valued elements, i.e. 0 or 1, and all matrix or vector additions and multiplications are performed modulo 2. $\mathbf{z}$ is a vector of length n whose elements are i.i.d. Bernoulli random variables with probability $P(z_i = 1) = f$. The typical set, $T_\varepsilon^{(n)}$, is defined by $T_\varepsilon^{(n)} = \left\{\mathbf{z} : \left| -n^{-1} \log P(\mathbf{z}) - H(f) \right| < \varepsilon \right\}$.

   (a) Explain why:

   (i) there exists an N_ε such that $P(\mathbf{z} \notin T_\varepsilon^{(n)}) < \varepsilon$ for $n > N_\varepsilon$. [2]

   (ii) the size of the typical set satisfies $\left|T_\varepsilon^{(n)}\right| \le 2^{n(H(f)+\varepsilon)}$ [2]

   (b) $\mathbf{B}$ is a matrix of dimension $(n-m) \times n$ whose elements are i.i.d. Bernoulli random variables with $P(b_{i,j} = 1) = 0.5$.

   (i) If $\mathbf{b}_1^T$ is the first row of $\mathbf{B}$, show that for any non-zero vector $\mathbf{d}$, $P(\mathbf{b}_1^T \mathbf{d} = 0) = 0.5$ where, as with all vector arithmetic in this question, the product is performed modulo 2 and all vector elements are either 0 or 1. [2]

   (ii) Hence show that $P(\mathbf{Bd} = \mathbf{0}) = 2^{m-n}$. [2]

   (iii) Explain why this implies that if $\mathbf{d}_1$ and $\mathbf{d}_2$ are distinct vectors, $P(\mathbf{Bd}_1 = \mathbf{Bd}_2) = 2^{m-n}$. [2]

   (c) For any matrix $\mathbf{B}$ as defined above, we select 2^m distinct codewords $\mathbf{x}_i$ of length n satisfying $\mathbf{Bx}_i = \mathbf{0}$. Codewords are transmitted through a binary symmetric channel with error probability f whose output is $\mathbf{y} = \mathbf{x} + \mathbf{z}$ where $\mathbf{z}$ represents the channel noise. The decoder estimates the noise by searching for $\hat{\mathbf{z}} \in T_\varepsilon^{(n)}$ such that $\mathbf{B}\hat{\mathbf{z}} = \mathbf{By}$ and then estimates the input codeword as $\hat{\mathbf{x}} = \mathbf{y} - \hat{\mathbf{z}}$.

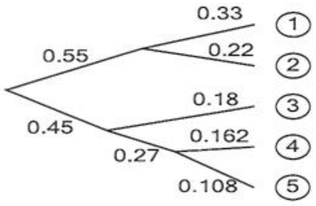
   (i) Show that the probability that the true noise vector, $\mathbf{z}$, satisfies the requirements for $\hat{\mathbf{z}}$ is greater than $1-\varepsilon$ for $n > N_\varepsilon$. [3]

   (ii) Show that the probability of a $\hat{\mathbf{z}} \ne \mathbf{z}$ satisfying the requirements is less than $2^{m-n} \times \left|T_\varepsilon^{(n)}\right|$. [3]

   (d) Determine a bound on $R = m/n$, below which the probability that $\hat{\mathbf{x}} \ne \mathbf{x}$ can be made arbitrarily small by taking n sufficiently large. [4]

2. In the diagram of *Figure 2.1*, x, y and $\hat{x}$ all lie in the set $\{0,1,2,3,4\}$ and $\hat{x}$ is a deterministic function of y. The Bernoulli variable e equals 1 if $\hat{x} \neq x$ and 0 otherwise.

(a) If $p_e = P(e = 1)$, justify each step in the derivation below and hence derive a lower bound on the probability that $\hat{x} \neq x$:



[12]

(b) The channel input takes one of five values: $x \in \{0,1,2,3,4\}$ with probabilities $[4,1,1,1,1]/8$ respectively. The channel output is given by $y = x + z$ modulo 5 where $z \in \{-2,-1,0,1,2\}$ with probabilities $[1,4,6,4,1]/16$ respectively.

(i) Show that $H(x \mid y) = 1.7463$ bits. [5]

(ii) Using the result of part (a), determine a bound on the decoder error probability. [1]

(iii) Define the operation of the decoder $\hat{x}(y)$ such that the error probability is minimized. Calculate the error probability of this optimum decoder. [2]

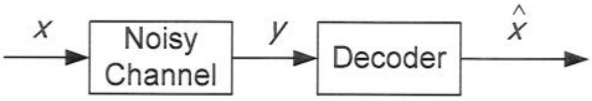


Figure 2.1

3. (a) *Figure 3.1* shows a binary tree used to construct a Fano code for the symbol set $\{1,2,3,4,5\}$ with probabilities $\mathbf{p}=[0.33, 0.22, 0.18, 0.162, 0.108]^T$. At each node in the tree, the symbols are split into two consecutive groups with the splitting point chosen to minimize the difference in total probability between the groups. Each branch is labelled with the total probability of its group. Thus the first split divides the symbol set into the two groups $\{1,2\}$ and $\{3,4,5\}$ with total probabilities 0.55 and 0.45 respectively.

(i) Give the Fano codeword for each symbol and the expected code length. [3]

(ii) Find the entropy $H(\mathbf{p})$. [3]

(b) (i) Show that if the initial symbol probabilities are in descending order, then the difference between the probabilities of the two groups at each split cannot exceed the probability of the lowest symbol in the upper group. [3]

(ii) Explain why each symbol except one will provide this bound for precisely one of the splitting operations. [2]

(c) At a typical internal node in the tree a consecutive set of symbols, $i:k$, is divided into the two groups $i:j$ and $j+1:k$. We define $H_{i:k}$ to be the entropy of a random variable $x_{i:k}\in\{i,i+1,\cdots,k\}$ having probability vector $[p_i, p_{i+1},\cdots,p_k]Q_{1:k}^{-1}$ where $Q_{i:k}=\sum_{r=i}^{k} p_r$. We define $L_{i:k}$ to be the expected length of a Fano code for $x_{i:k}$. For the leaf nodes, we define $H_{i:i}=L_{i:i}=0$.

For each non-leaf node,

(i) show that $L_{i:k}=1+L_{i:j}Q_{i:j}Q_{1:k}^{-1}+L_{j+1:k}Q_{j+1:k}Q_{1:k}^{-1}$ [3]

(ii) show that $H_{i:k}=H([Q_{i:j}Q_{1:k}^{-1},\quad Q_{j+1:k}Q_{1:k}^{-1}])+H_{i:j}Q_{i:j}Q_{1:k}^{-1}+H_{j+1:k}Q_{j+1:k}Q_{1:k}^{-1}$ [2]

(iii) hence show that

$$Q_{i:k}(L_{i:k}-H_{i:k})\le|Q_{i:j}-Q_{j+1:k}|+Q_{i:j}(L_{i:j}-H_{i:j})+Q_{j+1:k}(L_{j+1:k}-H_{j+1:k}).$$ [2]

You may assume without proof that $H([p,q])\ge 1-|p-q|$.

(d) By combining the answers to (b) and (c) show that that if the initial symbol probabilities are in descending order, the expected length of the Fano code for an alphabet size of n is bounded by $L_{1:n}\le H(\mathbf{p})+1-p_n$. [2]

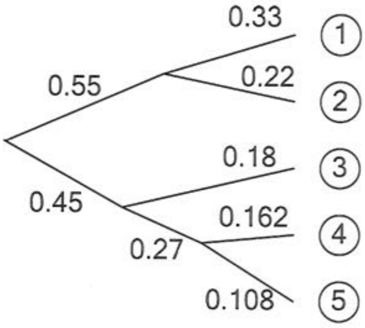


Figure 3.1

4. (a) Determine the differential entropy $h(V) = -E\log u_a(V)$ when V is uniformly distributed in the range $(-a,+a)$ with probability density function $u_a(v) = 0.5a^{-1}$. [3]

(b) By considering the relative entropy $D(f \| u_a) = E_f \log(f(X)/u_a(X))$, show that if X is restricted to the range $(-a,+a)$, its differential entropy is maximized when its distribution is uniform. [3]

(c) *Figure 4.1* shows a communications channel whose additive noise, Z, is uniformly distributed in the range $(-1,+1)$. Justifying each step in your argument, determine the distribution $f_1(x)$ that maximizes $I(X;Y) = h(Y) - h(Y|X)$ subject to the restriction that $|X| \le 1$. Give coding and decoding schemes that achieve the channel capacity for this case. [5]

(d) If X is instead subject to the restriction that $|X| \le 2$, determine the distribution $f_2(x)$ that maximizes $I(X;Y)$. Give coding and decoding schemes that achieve the channel capacity for this case. [4]

(e) Derive an expression for $I(X;Y)$ in terms of a when X is uniformly distributed in the range $(-a,+a)$ with $a \ge 1$. [6]

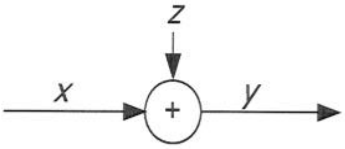


Figure 4.1

5. X is a continuous, possibly non-Gaussian, random variable with zero mean and variance σ^2. $\hat{X}$ is a random variable that is correlated with X and that satisfies $E(X-\hat{X})^2 \le D$. You may assume without proof that $h(X) \le ½\log(2\pi e\sigma^2)$ with equality if and only if X is Gaussian.

(a) Justify each step in the following sequence:

$$\begin{aligned}
I(X;\hat{X}) &\overset{\text{(i)}}{=} h(X) - h(X \mid \hat{X}) \\
&\overset{\text{(ii)}}{=} h(X) - h(X-\hat{X} \mid \hat{X}) \\
&\overset{\text{(iii)}}{\ge} h(X) - h(X-\hat{X}) \\
&\overset{\text{(iv)}}{\ge} h(X) - ½\log\left(2\pi e\,\mathrm{Var}\left(X-\hat{X}\right)\right) \\
&\overset{\text{(v)}}{\ge} h(X) - ½\log\left(2\pi e E\left(X-\hat{X}\right)^2\right) \\
&\overset{\text{(vi)}}{\ge} h(X) - ½\log 2\pi e D
\end{aligned}$$

[6]

(b) Suppose that $\hat{X}$ is generated as $\hat{X} = k(X+Z)$ where $k = 1 - D\sigma^{-2}$ and Z is zero-mean Gaussian with variance $k^{-1}D$ and is independent of X. Show that

(i) $E\,\hat{X}^2 = \sigma^2 - D$

(ii) $E\,(X-\hat{X})^2 = D$ [6]

(c) When $\hat{X}$ is generated as in part (b) above, justify the steps in the following

$$I(X;\hat{X}) \overset{\text{(i)}}{=} h(\hat{X}) - h(kZ) \overset{\text{(ii)}}{\le} ½\log(2\pi e(\sigma^2 - D)) - h(kZ).$$

[4]

Hence show that $I(X;\hat{X}) \le ½\log(\sigma^2 D^{-1})$.

(d) Explain the significance of the bounds proved in (a) and (c) on the rate at which it is possible to code with maximum squared error of D, a sequence of i.i.d. random variables drawn from the distribution of X. [4]

6. The stationary Bernoulli Markov process $\{x_i\}$ has $x_i \in \{0, 1\}$ and a transition matrix given by

$$\mathbf{Q} = \begin{pmatrix} 0.5 & 0.5 \\ 1 & 0 \end{pmatrix}$$

where $q_{a+1,b+1} = P(x_i = b \mid x_{i-1} = a)$ for $a, b \in \{0,1\}$.

(a) Determine

(i) The stationary distribution of the process.

(ii) The entropy of x_i. [6]

(iii) The entropy rate of the process.

(b) A coder processes the x_i in pairs, i.e. the first codeword encodes $\{x_1, x_2\}$, the next encodes $\{x_3, x_4\}$, etc. Calculate the probabilities of all possible pairs and hence design a Huffman coder. Determine the average code length per sample of $\{x_i\}$. [4]

(c) The Bernoulli stochastic process $\{y_i\}$ has a probability distribution that depends on $\{x_i\}$ with the following conditional probability matrix

$$\mathbf{R} = \begin{pmatrix} 0.75 & 0.25 \\ 0 & 1 \end{pmatrix}$$

where $r_{a+1,b+1} = P(y_i = b \mid x_i = a)$ for $a, b \in \{0,1\}$.

(i) Determine the distribution of y_i. [2]

(ii) Determine the entropy of y_i.

(d) A coder processes the y_i in pairs as in part (b). Calculate the probabilities of all possible pairs and hence design a Huffman coder. Determine the average code length per sample of $\{y_i\}$. [4]

(e) Determine the values of $H(y_i \mid y_{i-1})$ and $H(y_i \mid x_{i-1}, y_{i-1})$ and say how they relate to the entropy rate of the process $\{y_i\}$. [4]

Final version -
July 07.

2007 E4.40/SO20 Solutions

Key to letters on mark scheme: B=Bookwork, C=New computed example, A=New analysis

1. (a) (i) From the weak law of large numbers, $-n^{-1}\log p(\mathbf{z}) \xrightarrow{prob} H(f)$ since the elements $\mathbf{z}$ are i.i.d. The result follows directly. [2B]

(ii) $1=\sum_{\mathbf{z}} p(\mathbf{z}) \ge \sum_{\mathbf{z}\in T_\varepsilon^{(n)}} p(\mathbf{z}) \ge \sum_{\mathbf{z}\in T_\varepsilon^{(n)}} 2^{-n(H(Z)+\varepsilon)} = 2^{-n(H(f)+\varepsilon)}\left|T_\varepsilon^{(n)}\right|$. The result follows by rearranging. [2B]

(b) (i) $\mathbf{b}_1^T\mathbf{d} = \sum_{d_j\neq 0} b_{1,j}$ where the $b_{1,j}$ are i.i.d. Bernoulli random variables with Bernoulli probability of 0.5. This equals 0 with probability of 0.5 provided that the sum is non-empty, i.e. that $\mathbf{d}\neq\mathbf{0}$. [2A]

(ii) Since each row of $\mathbf{B}$ is independent, the $m-n$ elements of $\mathbf{Bd}$ are also independent and the probability that they all equal 0 is 2^{m-n}. [2A]

(iii) $\mathbf{Bd}_1=\mathbf{Bd}_2$ iff $\mathbf{B}(\mathbf{d}_1-\mathbf{d}_2)=\mathbf{0}$. Since $\mathbf{d}_1$ and $\mathbf{d}_2$ are distinct vectors, $\mathbf{d}_1-\mathbf{d}_2\neq\mathbf{0}$ and the result of part (ii) holds. [2A]

(c) (i) $\mathbf{z}$ definitely satisfies $\mathbf{Bz}=\mathbf{B}(\mathbf{y}-\mathbf{x})=\mathbf{By}-\mathbf{Bx}=\mathbf{By}$ and the probability that $\mathbf{z}\in T_\varepsilon^{(n)}$ is greater than $1-\varepsilon$ for $n>N_\varepsilon$. [3A]

(ii) From part (b)(iii), the probability of any specific member, $\hat{\mathbf{z}}\neq\mathbf{z}$, of $T_\varepsilon^{(n)}$ satisfying $\mathbf{B}\hat{\mathbf{z}}=\mathbf{Bz}$ is 2^{m-n} so the probability of any one of them satisfying it is less than $2^{m-n}\times\left|T_\varepsilon^{(n)}\right|$. [3A]

(d) The transmission can result in an error if either of the conditions in (c)(i) and (c)(ii) arise. The probability of an error is therefore less than their sum. I.e.

$$P(\hat{\mathbf{x}}\neq\mathbf{x}) \le \varepsilon + 2^{m-n}\times|T_\varepsilon^{(n)}| \le \varepsilon + 2^{m-n}\times 2^{n(H(f)+\varepsilon)} = \varepsilon + 2^{n(H(f)+\varepsilon+R-1)}$$

For this to become arbitrarily small for large n, we need the exponent to be negative:

$$H(f)+\varepsilon+R-1<0 \quad\Rightarrow\quad R<1-H(f)-\varepsilon$$ [4A]

The right hand side is in fact the capacity of the channel though this was not requested in the question.

2. (a) (i) From the chain rule, $H(e,x\,|\,y)=H(x\,|\,y)+H(e\,|\,x,y)$. However the second term is zero because e is completely determined by x and y. [3B]

(ii) $H(e,x\,|\,y)=H(e\,|\,y)+H(x\,|\,e,y)\le H(e)+H(x\,|\,e,y)$ where the first step follows from the chain rule and the second because conditioning reduces entropy. [3B]

(iii) We can split up a conditional entropy into a weighted sum of row entropies. [2B]

(iv) If $e=0$ then we know there is no error so x is completely determined by y so $H(x\,|\,y,e=0)=0$. [2B]

(v) e is a Bernoulli variable so $H(e)\le 1$. [1B]

Hence $p_e \ge \dfrac{H(x\,|\,y)-1}{\log(|\mathrm{X}|-1)}$ [1B]

(b) (i) We have $H(x\,|\,y)=H(x,y)-H(y)=H(x)+H(y\,|\,x)-H(y)$ and we can use either the first of the second expression. The joint probability distribution of x and y is

$$\frac{1}{8\times16}\begin{pmatrix}4&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{pmatrix}\begin{pmatrix}6&4&1&1&4\\4&6&4&1&1\\1&4&6&4&1\\1&1&4&6&4\\4&1&1&4&6\end{pmatrix}=\frac{1}{128}\begin{pmatrix}24&16&4&4&16\\4&6&4&1&1\\1&4&6&4&1\\1&1&4&6&4\\4&1&1&4&6\end{pmatrix}$$

From the column sums

$P(y)=[34,28,19,19,28]/128=[0.266,0.219,0.148,0.148,0.219]$ from which

$$H(y)=7-(34\times5.0875+2\times28\times4.8074+2\times19\times4.2480)/128$$
$$=2.2843 \text{ bits}$$

Method (1) To get $H(x,y)$ directly:

Including the scale factor, 2^{-7}, we can calculate

$$\log P(x,y)+7=\begin{pmatrix}4.585&4&2&2&4\\2&2.585&2&0&0\\0&2&2.585&2&0\\0&0&2&2.585&2\\2&0&0&2&2.585\end{pmatrix}$$

from which,

$$H(x,y)=7-(24\times4.585+2\times16\times4+4\times6\times2.585+10\times4\times2)/128$$
$$=4.0306 \text{ bits}$$ [5A]

Hence $H(x\,|\,y)=H(x,y)-H(y)=4.0306-2.2843=1.7463$ bits.

Method (2) To calculate $H(x,y)=H(x)+H(y\,|\,x)$

$\mathbf{p}(x)=[4,1,1,1,1]/8 \;\Rightarrow\; -\log\mathbf{p}(x)=[1,3,3,3,3] \;\Rightarrow\; H(x)=2$

$$\mathbf{p}(y \mid x) = [1,4,6,4,1]/16 \quad \Rightarrow \quad -\log \mathbf{p}(y \mid x) = [4,2,1.415,2,4]$$
$$\Rightarrow \quad H(y \mid x) = H(z) = 2.0306$$

Hence $H(x,y) = H(x) + H(y \mid x) = 2 + 2.0306 = 4.0306$

(ii) From (a), $p_e \geq \dfrac{H(x \mid y) - 1}{\log(|\mathrm{X}| - 1)} = \dfrac{0.7463}{2} = 0.3732$ [1C]

(iii) Taking the highest entry in each column of the above matrix, we see that the optimum decoder maps $y = [0,1,2,3,4]$ to $\hat{x} = [0,0,2,3,0]$. Summing the maximum probabilities in each column gives an error probability of $1 - (24 + 16 + 6 + 6 + 16)/128 = 0.4688$ which does indeed exceed the upper bound in (ii). [2A]

3. (a) (i) The codewords are $[00,01,10,110,111]$. Its expected length is given by $[0.33,0.22,0.18,0.162,0.108][2,2,2,3,3]^T = 2.27$ bits. [3C]

(ii) The entropy is given by

$$[0.33,0.22,0.18,0.162,0.108][1.60,2.18,2.47,2.63,3.21]^T = 2.23 \text{ bits}$$ [3C]

(b) (i) Suppose that $q \le p$ are the probabilities of the first symbol in the lower group and the last in the upper group respectively. If the probability of the upper group exceeds that of the lower group by more than p then transferring one symbol into the lower group will subtract $2p$ from this difference thereby reducing its absolute value. Similarly, if the probability of the lower group is larger by more than p, the difference also exceeds $q \le p$. and we can reduce its absolute value by transferring one symbol into the upper group. [3A]

(ii) The symbol providing the bound is the last one in the upper group. For all subsequent divisions of that group, this symbol is bound to be in the lower group and so can never again form the bound. After $n-1$ divisions, each symbol lies in its own group. Each symbol except the last must therefore have formed the bounding value for the division that separated it from its lower neighbour. [2A]

(c) (i) The Fano code uses one bit to select between $i:j$ and $j+1:k$ so the expected length is given by:

$$\begin{aligned} L_{1:k} &= 1 + L_{1:j}P(x \in 1:j) + L_{j+1:k}P(x \in j+1:k) \\ &= 1 + L_{1:j}Q_{1:j}Q_{1:k}^{-1} + L_{j+1:k}Q_{j+1:k}Q_{1:k}^{-1} \end{aligned}$$ [3A]

(ii) If we define u to be a Boolean variable that equals 1 if x is in the lower group, then since $H(u\,|\,x) = 0$,

$$\begin{aligned} H(x) &= H(x,u) = H(u) + H(x\,|\,u) \\ &= H(u) + H(x\,|\,u=1)P(u=1) + H(x\,|\,u=0)P(u=0) \\ &= H([Q_{1:j} \quad Q_{j+1:k}]Q_{1:k}^{-1}) + H_{i:j}Q_{1:j}Q_{1:k}^{-1} + H_{j+1:k}Q_{j+1:k}Q_{1:k}^{-1} \end{aligned}$$ [2A]

(iii) $H([Q_{1:j} \quad Q_{j+1:k}]Q_{1:k}^{-1}) \ge 1 - |Q_{1:j}Q_{1:k}^{-1} - Q_{j+1:k}Q_{1:k}^{-1}| = 1 - |Q_{1:j} - Q_{j+1:k}|\,Q_{1:k}^{-1}$

Hence, subtracting the previous two results gives

$$L_{1:k} - H_{1:k} \le 1 - 1 + |Q_{1:j} - Q_{j+1:k}|\,Q_{1:k}^{-1} + (L_{1:j} - H_{1:j})Q_{1:j}Q_{1:k}^{-1} + (L_{j+1:k} - H_{j+1:k})Q_{j+1:k}Q_{1:k}^{-1}$$ [2A]

from which the result follows by multiplying by $Q_{1:k}$.

(d) From (b), the quantity in (c)(iii) satisfies $|Q_{1:j} - Q_{j+1:k}| \le p_j$. Hence (c)(iii) becomes

$$(L_{1:k} - H_{1:k})Q_{1:k} \le p_j + (L_{1:j} - H_{1:j})Q_{1:j} + (L_{j+1:k} - H_{j+1:k})Q_{j+1:k}$$

[2A]

At the leaf nodes, the quantities $(L_{i,i} - H_{i,i})Q_{i,i} = 0$. As we go up the tree, we accumulate p_j terms and, from (b)(ii) eventually include them all except p_n. Thus they sum to $1 - p_n$ and, since, $Q_{1:n} = 1$, the result follows.

4. (a) $h(v) = -E \log u_a(v) = -\log(0.5a^{-1}) = \log 2a$ [2B]

(b) If $x \sim f(x)$, we have

$$0 \le D(f \| u_a) = E_f \log(f(x)/u_a(x)) = -h_f(x) + \log 2a$$ [3B]

and the result follows.

(c) We have:

$$I(x;y) = h(y) - h(y \mid x) = h(y) - h(x + z \mid x) = h(y) - h(z)$$
$$= h(y) - \log 2 = h(y) - 1$$

Thus we need to find the input distribution that maximizes $h(y)$. Since $|x| \le 1$, we must have $|y| \le 2$ and the optimum distribution for y is $u_2(y)$ giving $h(y) = \log 4 = 2 \;\Rightarrow\; I(x;y) = 1$. [5A]

We can achieve capacity by making $x = \pm 1$ with equal probabilities. We can then send one bit per channel use and the decoder just detects the sign of y.

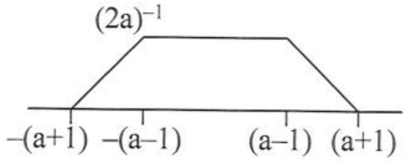
(d) Now we have

$$|x| \le 2 \;\Rightarrow\; |y| \le 3 \;\Rightarrow\; h(y) \le \log 6 = 2.585 \;\Rightarrow\; I(x;y) \le 1.585$$

We can make y uniform by choosing $x \in \{-1 \quad 0 \quad 1\}$ with equal probabilities and detecting at thresholds of $y = \pm 0.5$. [4A]

(e) The pdf of $y = x + z$ is trapezoidal in the range $\pm(a+1)$ since it is the convolution of two rectangular distributions. More precisely:

$$p(y) = \begin{cases} 0.5a^{-1} & |y| \le a-1 \\ 0.25a^{-1}(1 + a - |y|) & a-1 < |y| \le a+1 \end{cases}$$



From symmetry and shift-invariance, we need integrate only half the distribution and can shift it so the sloping portion goes through the origin:

$$h(\mathcal{Y}) = -2\int_0^{a+1} p \log p \, dy$$

$$= -2\int_0^{2} p \log p \, dy - 2(a-1) \times (2a)^{-1} \times -\log 2a$$

(substitute $p = (4a)^{-1} y \quad \Rightarrow \quad dy = 4a \, dp$)

$$= (1-a^{-1})\log 2a - 8a \log e \times \int_0^{(2a)^{-1}} p \log p \, dp$$

$$= (1-a^{-1})\log 2a - 8a \log e \left[0.5 p^2(-0.5 + \ln p)\right]_0^{(2a)^{-1}}$$

$$= (1-a^{-1})\log 2a + 0.5a^{-1}\log e + a^{-1}\log(2a)$$

$$= \log 2a + 0.5a^{-1}\log e$$

Hence [6A]

$$I(\mathcal{X};\mathcal{Y}) = h(\mathcal{Y}) - 1 = \log 2a + 0.5a^{-1}\log e - 1 = \log a + 0.5a^{-1}\log e$$

5. (a) (i) Definition of mutual information

(ii) Translation invariance of differential entropy + $\hat{X}$ is conditionally constant.

(iii) Removing conditioning increases entropy (decrease because of – sign).

(iv) Gaussian bound on differential entropy for a given variance

(v) Mean square value is ≥ Variance (note $\hat{X}$ may not be zero mean) [6B]

(vi) Mean square deviation bounded by D and log is monotonic increasing.

(b) (i) Since X and Z are xero mean and independent, we have

$$E\,\hat{X}^2 = k^2 EX^2 + k^2 EZ^2 = k^2\sigma^2 + kD = k(k\sigma^2 + D)$$
$$= (1 - D\sigma^{-2})((1 - D\sigma^{-2})\sigma^2 + D) = (1 - D\sigma^{-2})\sigma^2 = \sigma^2 - D$$

[3A]

(ii) For the same reason,

$$E(X - \hat{X})^2 = E(X - k(X + Z))^2 = E((1-k)X - kZ)^2 = (1-k)^2 E\,X^2 + k^2 E\,Z^2$$
$$= (1-k)^2\sigma^2 + k^2k^{-1}D = (1-k)^2\sigma^2 + kD$$
$$= (D\sigma^{-2})^2\sigma^2 + (1 - D\sigma^{-2})D = D^2\sigma^{-2} + D + D^2\sigma^{-2} = D$$

[3A]

(c) (i) This is a contraction of steps (i), (ii) and (iii) from part (a).

(ii) This is the Gaussian upper bound using the variance calculated in (b)(i).

Using this as the starting point [4A]

$$I(X;\hat{X}) \le ½\log(2\pi e(\sigma^2 - D)) - h(kZ)$$
$$= ½\log(2\pi e(\sigma^2 - D)) - ½\log(2\pi e(k^2k^{-1}D))$$
$$= ½\log(2\pi e(k\sigma^2)) - ½\log(2\pi e(kD)) = ½\log(\sigma^2 D^{-1})$$

(d) We have shown in (a) that it is always true that $I(X;\hat{X}) \ge h(X) - ½\log 2\pi e D$ and in (c) that there exists an $\hat{X}$ satisfying the distortion constraint with $I(X;\hat{X}) \le ½\log(2\pi e\sigma^2 D^{-1})$. The rate distortion function $R(D)$ is the minimum value of $I(X;\hat{X})$ with $\hat{X}$ satisfying the distortion constraint and must therefore lie between these bounds. Note that if X happens to be Gaussian then the two bounds coincide. [4A]

6. (a) (i) The stationary distribution satisfies $\mathbf{Q}^T\mathbf{p} = \mathbf{p}$. If $\mathbf{p} = [p, q]^T$ we have $p = 0.5p + q \;\Rightarrow\; p = 2q \;\Rightarrow\; [p, q] = [2/3, 1/3]$.

(ii) The entropy of $\{x_i\}$ is $H([2/3, 1/3]) = 0.918 \text{ bits}$.

(iii) The entropy rate of a stationary Markov process is [6C]

$$H(x_i \mid x_{i-1}) = 2/3 \times H(0.5) + 1/3 \times H(1) = 2/3 \times 1 + 1/3 \times 0 = 0.667 \text{ bits}.$$

(b) We have $P(ab) = P(x_1 = a)P(x_2 = b \mid x_1 = a)$ which gives:

$$P([00, 01, 10, 11]) = [0.667 \times 0.5,\ 0.667 \times 0.5,\ 0.333 \times 1,\ 0.333 \times 0] = [0.333,\ 0.333,\ 0.333,\ 0]$$

Huffman codes are therefore 0, 10, 11 with no code for the impossible pair 11. The average code length is 1.667 bits per pair or 0.833 bits per sample: this lies between the entropy rate and the entropy. [4A]

(c) (i) $\mathbf{p}_y = \mathbf{R}^T\mathbf{p}_x = [0.5,\ 0.5]^T$. [2A]

(ii) $H(y_i) = H(\mathbf{p}_y) = 1 \text{ bit}$.

(d) We can take the pair probabilities calculated in (b) and determine the possible $(y_i\}$ pairs that can result. We have the following conditional probabilities:

x pair	p(x)	y pair			
		00	01	10	11
00	1/3	9/16	3/16	3/16	1/16
01	1/3	0	12/16	0	4/16
10	1/3	0	0	12/16	4/16

[4A]

Summing the columns and multiplying by $p(x_i) = 1/3$ gives

$$P([00, 01, 10, 11]) = [3/16,\ 5/16,\ 5/16,\ 3/16] = [0.1875,\ 0.3125,\ 0.3125,\ 0.1875]$$

The Huffman codes are 10, 00, 01, 11 and the average code length is 2.

(e) From the pair probabilities in part (d) we get

$$H(y_i \mid y_{i-1}) = H([3/8,\ 5/8]) = 0.9544 \text{ bits}$$

Since y_i does not depend directly on y_{i-1} we have

$$\begin{aligned} H(y_i \mid x_{i-1}, y_{i-1}) &= H(y_i \mid x_{i-1}) \\ &= H(y_i \mid x_{i-1} = 0)P(x_{i-1} = 0) + H(y_i \mid x_{i-1} = 0)P(x_{i-1} = 0) \\ &= H([12,\ 20]/32) \times 2/3 + H([12,\ 4]/16) \times 1/3 \\ &= 0.9544 \times 2/3 + 0.8113 \times 1/3 = 0.9067 \end{aligned}$$

[4A]

The entropy rate of $\{y_i\}$ must lie between these two values.