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 2006

MSc and EEE/ISE PART IV: MEng and ACGI

**INFORMATION THEORY**

Wednesday, 26 April 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, $\mathsf{x}$, 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.

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

## The Questions

1. (a) Show that, for a discrete random variable $X$ with alphabet size $|\mathcal{X}| = N$, the maximum value of $H(X)$ is obtained when all members of $\mathcal{X}$ have equal probability. You may assume without proof that for any probability mass vectors $\mathbf{p}$ and $\mathbf{q}$, $D(\mathbf{p} \| \mathbf{q}) = E_{\mathbf{p}}\{-\log q(X)\} - H(\mathbf{p}) \geq 0$. [6]

(b) The value, $X$, obtained when a biased die is thrown can take the possible values $\mathbf{v} = [1, 2, \cdots, 6]$ with probabilities $\mathbf{p} = [p_1, p_2, \cdots, p_6]$. The expected value of $X$ is known and is equal to $q$.

(i) Using the method of Lagrange multipliers or otherwise, show that the $\mathbf{p}$ that maximizes $H(X)$ subject to the constraints $\mathbf{p}^T \mathbf{v} = q$ and $\sum \mathbf{p} = 1$, is of the form $\mathbf{p} = [ab, ab^2, \cdots, ab^6]^T$. [6]

(ii) Show that $b$ satisfies $(q-6)b^6 + (q-5)b^5 + \ldots + (q-1)b = 0$ [2]

(c) Give a simplified expression for $H(X)$ in terms of $a$, $b$ and $q$ and determine its value when $b = 2$. [4]

(d) Determine the values of $a$, $b$ and $H(X)$ when $q = 3.5$. [2]

2. The sequence $\{\mathsf{x}_i\}$ arises from a stationary binary Markov process with transition matrix

$$\mathbf{Q} = \begin{pmatrix} 1-r & r \\ q & 1-q \end{pmatrix}$$

*Table 2.1* gives the probabilities of all possible 4 symbol sequences for the particular case $q = 0.06$ and $r = 0.04$.

(a) Show that the only probability mass vector satisfying $\mathbf{Q}^T\mathbf{p} = \mathbf{p}$ is $\mathbf{p} = \left[ q(r+q)^{-1} \quad r(r+q)^{-1} \right]^T$. Explain the significance of this uniqueness. [5]

(b) For the particular case $q = 0.06$ and $r = 0.04$, calculate the entropy $H(\mathsf{x}_i)$ and the entropy rate of the process $H(\boldsymbol{\mathcal{X}})$. Explain why these two quantities are not equal in value and explain the significance of the difference for coding the sequence efficiently. [5]

(c) The sequence $\{\mathsf{x}_i\}$ represents the $2^{16}$ pixels of a binary image for which $q = 0.06$ and $r = 0.04$. Determine the expected number of bits needed to code the entire image for each of the following codes:

(i) Each pixel is encoded individually using 1 bit. [2]

(ii) $\{\mathsf{x}_i\}$ is divided into groups of 3 consecutive values and a Huffman code used to encode each group. [4]

(iii) The sequence $\mathsf{y}_i = \mathsf{x}_i \oplus \mathsf{x}_{i-1}$ is formed by exclusive-oring each pixel with the previous one (with the initial value $\mathsf{x}_0 = 0$). Then $\{\mathsf{y}_i\}$ is divided into groups of 3 consecutive values and a Huffman code used to encode each group. [3]

(d) Explain how your answers to (c) related to those of (b). [1]

| **x** | $p(\mathbf{x})$ | **x** | $p(\mathbf{x})$ |
|---|---|---|---|
| 0000 | 0.5308416 | 1000 | 0.0221184 |
| 0001 | 0.0221184 | 1001 | 0.0009216 |
| 0010 | 0.0013824 | 1010 | 0.0000576 |
| 0011 | 0.0216576 | 1011 | 0.0009024 |
| 0100 | 0.0013824 | 1100 | 0.0216576 |
| 0101 | 0.0000576 | 1101 | 0.0009024 |
| 0110 | 0.0013536 | 1110 | 0.0212064 |
| 0111 | 0.0212064 | 1111 | 0.3322336 |

*Table 2.1*

3. (a) Prove that

(i) $H(x \mid z) = \sum_{z \in \mathbf{Z}} H(x \mid z = z)\, p(z = z)$ [4]

(ii) $I(x;y \mid z) = \sum_{z \in \mathbf{Z}} I(x;y \mid z = z)\, p(z = z)$ [4]

(b) *Figure 3.1* shows a communications system containing a binary symmetric channel. The channel is shown in *Figure 3.2* and has an error probability $f(z)$ whose value is a known function of a binary random variable $z$ which takes a value 0 or 1 with equal probability.

For each of the following cases, determine the capacity of the channel by deriving the maximum value of $I(w;\hat{w})$. You may assume without proof that the capacity of a binary symmetric channel with error probability $q$ is equal to $1 - H(q)$.

(i) Neither the encoder nor the decoder knows the value of $z$. [2]

(ii) Both the encoder and decoder know the value of $z$. [2]

(iii) The decoder knows the value of $z$ but the encoder does not. [2]

(iv) The encoder knows the value of $z$ but the decoder does not. [2]

(c) For the specific case in which $f(0) = 0$ and $f(1) = 1$, give encoding and decoding algorithms that allow binary information to be sent through the system at its full capacity for each of the situations (i), (ii), (iii) and (iv) of part (b). [4]

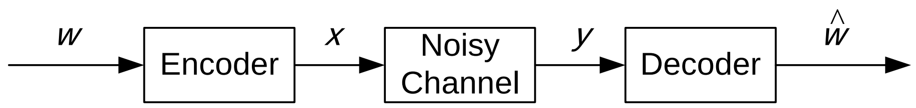



*Figure 3.1*

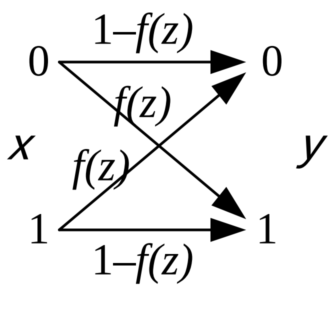



*Figure 3.2*

4. For a real-valued continuous random variable $x$ with a known probability density function $p(x)$, we define the *Information Rate Distortion* function, $R(D)$, as

$$R(D) = \min I(x;\hat{x}) \text{ over all } p(\hat{x} \mid x) \text{ such that } E_{x,\hat{x}}\left( (x-\hat{x})^2 \right) \le D$$

(a) If $p(x) = N(x;0,\sigma^2)$, justify each of the following steps in deriving a lower bound for $R(D)$ and state the conditions for equality in steps (iii), (iv), (v) and (vi). You may assume without proof that $h(x) = ½\log 2\pi e\sigma^2$. [7]

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

$$\text{Hence}\quad R(D) \overset{\text{(vi)}}{\ge} \max\left( ½\log\frac{\sigma^2}{D}, 0 \right)$$

(b) $z$ and $\hat{z}$ are complex-valued random variables and the real and imaginary parts of $z$, $z_R$ and $z_I$, are independent. Show that $I(z;\hat{z}) \ge I(z_R;\hat{z}_R) + I(z_I;\hat{z}_I)$ and state the conditions for equality. [5]

(c) We wish to quantise a sequence of independent identically-distributed complex random numbers, $\{z_i\}$. The real and imaginary parts of $z_i$ are independent and follow zero-mean Gaussian distributions with variances of 4 and 1 respectively. The quantisation distortion is defined by $D = E\left( \left| z-\hat{z} \right|^2 \right)$.

(i) Assuming that the bounds of parts (a) and (b) are attainable, prove that $R(D) = 2 - \log D$ provided that $D \le 2$. [5]

(ii) Determine expressions for $R(D)$ for all values of $D$ and in each case, state the range of $D$ for which they apply. [3]

5. (a) $x$ and $y$ are correlated discrete random variables and we form an estimate $\hat{x} = f(y)$ using a deterministic function $f()$. We define a binary random variable $e$ to equal 0 if $\hat{x} = x$ and 1 if $\hat{x} \neq x$. Justify each of the steps in the following where $p_e$ is the probability that $\hat{x} \neq x$:

$$\begin{aligned} H(x\,|\,y) + H(e\,|\,y,x) &\overset{\text{(i)}}{=} H(e\,|\,y) + H(x\,|\,y,e) \\ \Rightarrow \quad H(x\,|\,y) &\overset{\text{(ii)}}{\le} H(e) + H(x\,|\,y,e) \\ &\overset{\text{(iii)}}{=} H(e) + H(x\,|\,y,e=0)(1-p_e) + H(x\,|\,y,e=1)\,p_e \\ &\overset{\text{(iv)}}{\le} 1 + \log(|\boldsymbol{\mathcal{X}}|-1)\,p_e \end{aligned}$$

[7]

(b) *Figure 5.1* shows a communications system whose input $w$ takes one of $2^{nR}$ equiprobable values and is converted into a sequence $x_{1:n}$ of $n$ values for transmission through the Gaussian channel. The channel output $y_{1:n}$ is decoded to generate an estimate $\hat{w}$ of $w$ with an error probability $p_e^{(n)}$ that tends to 0 as $n \to \infty$. Assuming that the capacity of a Gaussian channel with noise variance $N$ and average power constraint $P$ is $½\log\left(1+PN^{-1}\right)$, justify each step in the following argument

$$\begin{aligned} nR \overset{\text{(i)}}{=} H(w) &\overset{\text{(ii)}}{=} I(w;y_{1:n}) + H(w\,|\,y_{1:n}) \\ &\overset{\text{(iii)}}{\le} I(x_{1:n};y_{1:n}) + H(w\,|\,y_{1:n}) \\ &\overset{\text{(iv)}}{\le} \sum_{i=1}^{n} I(x_i;y_i) + H(w\,|\,y_{1:n}) \\ &\overset{\text{(v)}}{\le} ½n\log\left(1+PN^{-1}\right) + 1 + nRp_e^n \end{aligned}$$

[7]

Explain carefully why this implies for both large and small $n$ that if $R > ½\log\left(1+PN^{-1}\right)$ there is a non-zero lower bound for $p_e^{(n)}$. [3]

(c) (i) The bandwidth of a DAB (Digital Audio Broadcasting) digital radio channel is 1.537 MHz. If the total signal to in-band noise ratio is 14 dB, calculate the theoretical maximum data rate. You may assume that a continuous time Gaussian channel with bandwidth $B$ is equivalent to a discrete time channel with $2B$ transmissions per second and the same in-band signal-to-noise ratio. [2]

(ii) The raw data rate of the DAB channel is 2.4 Mbit/s. Determine the lowest possible signal-to-noise ratio in dB at which the system could operate. [1]

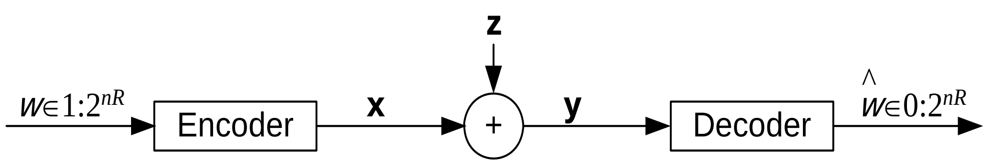



*Figure 5.1*

6. (a) *Figure 6.1* shows which shows a channel with input and output alphabets $\mathcal{X} = \mathcal{Y} = \{1, 2, 3, 4\}$. Determine the capacity of the channel for the special cases [4]

(i) $f = g = 0$, (ii) $f = g = ½$ and (iii) $f = g = 1$.

(b) Show that the input distribution to the channel may always be written in the form $\mathbf{p}_x = [(1-a)(1-b) \quad (1-a)b \quad a(1-c) \quad ac]$ where $0 \le a, b, c \le 1$. Hence show that $I(x;y) = H(a) + (1-a)I(x;y \mid x \le 2) + aI(x;y \mid x \ge 3)$. [6]

(c) Using the result of (b), express $I(x;y)$ in terms of $a$, $f$ and $g$ and hence find the value of $a$ that maximises the channel capacity. You may assume without proof that the capacity of a binary symmetric channel with error probability $q$ is equal to $1 - H(q)$. [5]

(d) For each of the following cases, determine the channel capacity and the input distribution $\mathbf{p}_x$ that attains it. In each case, comment briefly on the significance of the result.

(i) $f = g = 0.25$ [5]

(ii) $f = 0, \quad g = 0.5$

(iii) $f = 0, \quad g = 0.25$

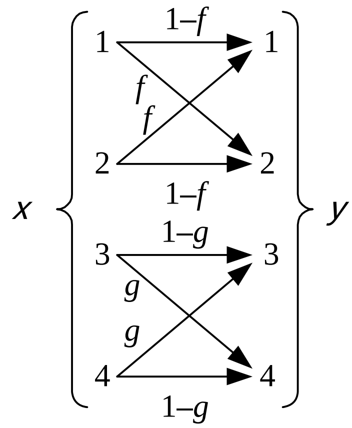



*Figure 6.1*

# 2006 E4.40/SO20 Solutions

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

1. (a) If we set all elements of $\mathbf{q}$ to $N^{-1}$, then for any $\mathbf{p}$ we have:

$$0 \le D(\mathbf{p} \| \mathbf{q}) = E_{\mathbf{p}}\{-\log q(X)\} - H(\mathbf{p}) = -\log N^{-1} - H(\mathbf{p})$$
$$\Rightarrow \quad H(\mathbf{p}) \le \log N = H(\mathbf{q})$$

[6B]

(b) (i) Using a Lagrange multiplier, we wish to maximize

$$J = H(X) + \lambda \mathbf{p}\mathbf{v}^T = -\mathbf{p}^T \log(\mathbf{p}) + \lambda \mathbf{p}^T \mathbf{v} + \mu \mathbf{p}^T \mathbf{1} = \mathbf{p}^T (\lambda \mathbf{v} - \log(\mathbf{p}) + \mu)$$
$$\frac{dJ}{d\mathbf{p}} = (\lambda \mathbf{v} - \log(\mathbf{p}) + \mu) - \log e = 0$$

[6A]

$$\Rightarrow \quad \log(e\mathbf{p}) = \lambda \mathbf{v} + \mu \quad \Rightarrow \quad \mathbf{p} = e^{-1} 2^{\mu} 2^{\lambda \mathbf{v}} = a 2^{\lambda \mathbf{v}}$$

Hence $a = e^{-1} 2^{\mu}$ and $\mathrm{b} = 2^{\lambda}$

Where $\mu$ and $\lambda$ are chosen to ensure that $\mathbf{p}^T \mathbf{1} = 1$ and $\mathbf{p}^T \mathbf{v} = q$.

(ii) $0 = q - \mathbf{p}^T \mathbf{v} = \mathbf{p}^T \mathbf{1} q - \mathbf{p}^T \mathbf{v} = \mathbf{p}^T (\mathbf{1} q - \mathbf{v})$ from which the result follows. [2A]

(c) Since $\log a = \mu - \log e$ and $\log b = \lambda$, we can write

$$\log \mathbf{p} = \log(b)\mathbf{v} + \log(a)\mathbf{1}$$
$$H(X) = -\mathbf{p}^T \log \mathbf{p} = -\log(b)\mathbf{p}^T \mathbf{v} - \log(a)\mathbf{p}^T \mathbf{1} = -q \log b - \log a = -\log\left(ab^q\right)$$

If $b = 2$ then $\mathbf{p}^T \mathbf{1} = 1 \Rightarrow a^{-1} = 2(1 + 2 + \ldots + 32) = 126$ and $q = a(2 + 2 \times 4 + 3 \times 8 + \ldots + 6 \times 64) = 642/126 = 5.0952$. Hence $H(X) = 1.882$ bits. [4A]

(d) From (a), the unconstrained entropy maximum is when all elements of $\mathbf{p}$ are equal and this gives $q = 3.5$. Thus the entropy maximum under the constraint that $q = 3.5$ will also have all elements of $\mathbf{p}$ equal. Thus $a = 1/6$ and $b = 1$. Hence $H(X) = \log 6 = 2.585$ bits. [2A]

2. (a) If $\mathbf{p}=\begin{bmatrix}a & b\end{bmatrix}^T$, then we require $\left(\mathbf{Q}^T-\mathbf{I}\right)\mathbf{p}=\mathbf{0} \Leftrightarrow ra=qb \Leftrightarrow \frac{a}{b}=\frac{q}{r}$. Since we also require $a+b=1$ for a valid probability mass vector, the only solution is $\mathbf{p}=\begin{bmatrix}q(r+q)^{-1} & r(r+q)^{-1}\end{bmatrix}^T$. This means that the Markov process has a unique steady-state distribution and that it will converge to this regardless of the distribution of $x_1$. [5A]

(b) For these values of $q$ and $r$, we have $\mathbf{p}=\begin{bmatrix}0.6 & 0.4\end{bmatrix}$ giving $H(\mathsf{x}_i)=0.971\,\text{bits}$.

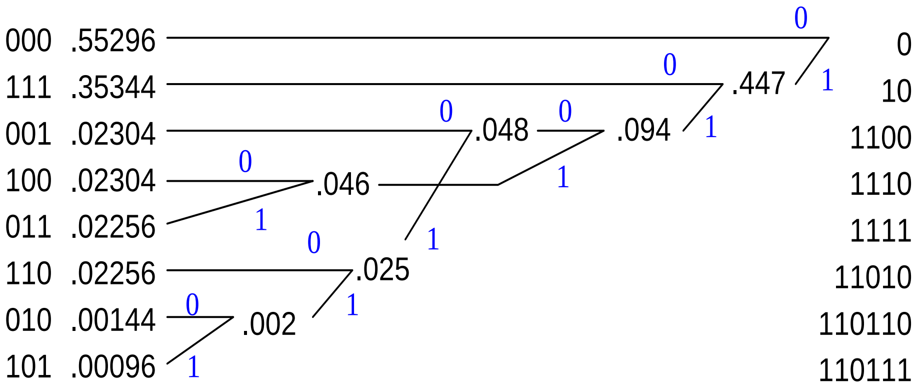
On the other hand,

$$\begin{aligned} H(\boldsymbol{X}) &= H(x_i \mid x_{i-1}) \\ &= H(x_i \mid x_{i-1}=0)\,p(x_{i-1}=0)+H(x_i \mid x_{i-1}=1)\,p(x_{i-1}=1) \\ &= H(r)\times 0.6+H(q)\times 0.4 \\ &= 0.2423\times 0.6+0.3274\times 0.4=0.2764\text{ bits} \end{aligned}$$ [5C]

Thus, if we treat the $\{\mathsf{x}_i\}$ as independent when coding, we will require at least 0.971 bits per symbol. If however we take advantage of the inter-symbol correlations, we can reduce this to 0.2764 bits per pixel with a large enough block size.

(c) (i) We need 1 bit per pixel or $2^{16}=65536\,\text{bits}$ in total. [2A]

(ii) We create a Huffman tree in which the probabilities come from summing the rows of table 2.1:

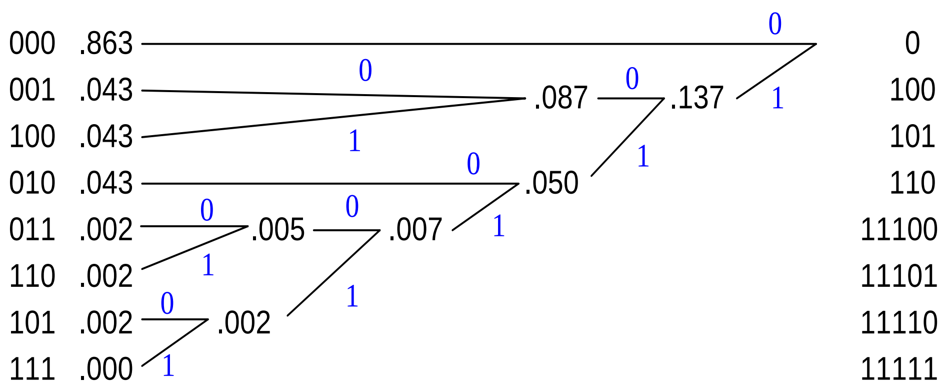



The average length is 1.662 bits for every 3 pixels giving $1+21845\times 1.662=36298.7$ bits for the entire image. (the 1 arises because $2^{16}$ is not a multiple of 3) [4C]

(iii) From the table of probabilities for 4 consecutive bits, we can them in pairs to get the probabilities for the eight possible y triples:

| y | x\|0 | p(x\|0) | x\|1 | p(x\|1) | p |
|---|---|---|---|---|---|
| 000 | 0000 | 0.5308416 | 1111 | 0.3322336 | 0.8630752 |
| 001 | 0001 | 0.0221184 | 1110 | 0.0212064 | 0.0433248 |
| 100 | 0111 | 0.0212064 | 1000 | 0.0221184 | 0.0433248 |
| 010 | 0011 | 0.0216576 | 1100 | 0.0216576 | 0.0433152 |
| 011 | 0010 | 0.0013824 | 1101 | 0.0009024 | 0.0022848 |
| 110 | 0100 | 0.0013824 | 1011 | 0.0009024 | 0.0022848 |
| 101 | 0110 | 0.0013536 | 1001 | 0.0009216 | 0.0022752 |
| 111 | 0101 | 0.0000576 | 1010 | 0.0000576 | 0.0001152 |

Hence we can construct the Huffman tree:




[3A]

This gives an average of 1.288 bits for every 3 pixels to give a total of $1+21845\times1.288=28132.3$ bits for the entire image.

(d) Both approaches would achieve the entropy limit of $1+65535\times0.2764=18115$ if the block size is made large enough. However, by performing the exclusive or operation, we squeeze more of the probability into the most common case (y=000) and allow the Huffman code to perform better. Even so, the bit requirement in (iii) is substantially larger than optimum. [1A]

3. (a) (i)

[4B]

$$
\begin{aligned}
H(x \mid z) &= E - \log p(x \mid z) = \sum_{x,z} - p(x,z)\log p(x \mid z) \\
&= \sum_{x,z} - p(z)p(x \mid z)\log p(x \mid z) = \sum_{z \in \mathbf{Z}} p(z) \sum_{x \in \mathbf{X}} - p(x \mid z)\log p(x \mid z) \\
&= \sum_{z \in \mathbf{Z}} p(z) H\left(x \mid z\right)
\end{aligned}
$$

(ii)

[4B]

$$
\begin{aligned}
I(x;y \mid z) &= H(x \mid z) + H(y \mid z) - H(x,y \mid z) \\
&= \sum_{z \in \mathbf{Z}} \left(H(x \mid z = z) + H(y \mid z = z) - H(x,y \mid z = z)\right) p(z = z) \\
&= \sum_{z \in \mathbf{Z}} I(x;y \mid z = z) p(z = z)
\end{aligned}
$$

(b) (i) From the data processing theorem, $I(w;\hat{w}) \le I(x;y)$. If neither the transmitter or receiver know the value of $z$, then we just have a binary symmetric channel with an error probability of $f = ½ f(0) + ½ f(1)$. [2A]

$$
I(x;y) = H(y) - H(y \mid x) = H(y) - H(f) \le 1 - H(f)
$$

with equality iff $y$ is uniform which in turn is iff $x$ is uniform.

(ii) If both transmitter and receiver know $z$, then we want to maximize $I(x;y \mid z)$. From part (a)(i):

[2A]

$$
\begin{aligned}
I(x;y \mid z) &= \sum_{z \in \mathbf{Z}} I(x;y \mid z = z) p(z = z) \\
&\le \left(1 - H(f(0))\right) p(z = 0) + \left(1 - H(f(1))\right) p(z = 1) \\
&= 1 - ½ H(f(0)) - ½ H(f(1))
\end{aligned}
$$

with equality, as before, when $x$ is uniform. This will in general be greater than the value in part (i) due to the concavity of $H(f)$.

(iii) If only the receiver knows the value of $z$, then $\hat{w} = \hat{w}(y,z)$ and we need to maximize $I(x;y,z)$, that is, the information that $y$ and $z$ together give you about $x$. From the chain rule for mutual information: [2A]

$$
\begin{aligned}
I(x;y,z) &= I(x;z) + I(x;y \mid z) \\
&\le 0 + 1 - ½ H(f(0)) - ½ H(f(1))
\end{aligned}
$$

where $I(x;z) = 0$ since they are independent. As before, equality arises when $x$ is uniform.

(iv) If only the transmitter knows the value of $z$, then $x = x(w,z)$ and we have a markov chain $(w,z) \to (x,z) \to y$ thus [2A]

$$
\begin{aligned}
I(w,z;y) &\le I(x,z;y) \\
&= H(y) - H(y \mid x,z) \\
&= H(y) - H(y \mid x, z = 0) p(z = 0) - H(y \mid x, z = 1) p(z = 1) \\
&\le 1 - ½ H(f(0)) - ½ H(f(1))
\end{aligned}
$$

(c) We have in all cases $y = x \oplus z$

(i) In this case $f = ½ f(0) + ½ f(1) = ½ \Rightarrow H(f) = 1$. Hence channel capacity is 0 and the encoding/decoding scheme doesn't matter.

(ii) Set $x = w$ and $\hat{w} = y \oplus z = w$ for error-free transmission.

(iii) Set $x = w$ and $\hat{w} = y \oplus z = w$ for error-free transmission.

(iv) Set $x = w \oplus z$ and $\hat{w} = y = w$ for error-free transmission. [4A]

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

(ii) $h(x) = ½\log 2\pi e\sigma^2$ given in question. $h(x \mid \hat{x}) = h(x - \hat{x} \mid \hat{x})$ because of translational invariance.

(iii) Conditioning reduces entropy. We have equality if the error, $x - \hat{x}$, is independent of $\hat{x}$.

(iv) A Gaussian has the maximum entropy for a given variance. Thus it is always true that $h(x - \hat{x}) \le ½\log\left(2\pi e \operatorname{Var}\left(x - \hat{x}\right)\right)$. We have equality if the error, $x - \hat{x}$, is Gaussian.

(v) $\operatorname{Var}\left(x - \hat{x}\right) = E\left(\left(x - \hat{x}\right)^2\right) - \left(E\left(x - \hat{x}\right)\right)^2 \le E\left(\left(x - \hat{x}\right)^2\right) \le D$ from the constraint on $p(\hat{x} \mid x)$. We have equality if the error, $x - \hat{x}$, is zero-mean with variance $D$.

(vi) $R(D) = \min\left(I(x;\hat{x})\right) \ge \min\left(½\log\frac{\sigma^2}{D}\right)$ but also $I(x;\hat{x}) \ge 0$ so the result follows. We have equality if we can find a $p(\hat{x} \mid x)$ that gives equality in steps (iii), (iv) and (v) or alternatively, if the variance of $\operatorname{Var}(x) \le D$ when we can just set $\hat{x} \equiv E(x)$. [7B]

(b)

$$
\begin{aligned}
I(z;\hat{z}) &= h(z_R, z_I) - h(z_R, z_I \mid \hat{z}_R, \hat{z}_I) \\
&= h(z_R) + h(z_I) - h(z_R \mid \hat{z}_R, \hat{z}_I) - h(z_I \mid \hat{z}_I, \hat{z}_R, z_R) \\
&\ge h(z_R) + H(z_I) - h(z_R \mid \hat{z}_R) - h(z_I \mid \hat{z}_I) \\
&= I(z_R \mid \hat{z}_R) + I(z_I \mid \hat{z}_I)
\end{aligned}
$$

[5A]

We have equality provided that $z_R$ is independent of $\hat{z}_I$ and also $z_I$ is independent of $\hat{z}_R$ (true for most reasonable encodings).

(c) We have $D = D_R + D_I = E\,(z_R - \hat{z}_R)^2 + E\,(z_I - \hat{z}_I)^2$

(i)

$$
\begin{aligned}
I &= \max_{D_R}\left(½\log\frac{4}{D_R} + ½\log\frac{1}{D - D_R}\right) \\
&= ½\log e\left(\ln 4 - \ln D_R - \ln(D - D_R)\right)
\end{aligned}
$$

$$
\frac{dI}{dD_R} = ½\log e\left(\frac{-1}{D_R} + \frac{1}{D - D_R}\right) = 0 \text{ when } D_R = ½D
$$

[5C]

$$
\begin{aligned}
\text{Hence}\quad I &= ½\log\frac{8}{D} + ½\log\frac{2}{D} \\
&= ½\,(3 - \log D + 1 - \log D) = 2 - \log D
\end{aligned}
$$

This expression is valid provided that $½D = D_R, D_I \le \min\left(\sigma_R^2, \sigma_I^2\right) = \min\left(4,1\right) = 1$, that is when $D \le 2$.

(ii) If $D \ge 5$ then we set $\hat{z} = 0$ and $R(D) = 0$. If $2 \le D \le 5$, then we set $\hat{z}_I = 0$ and just encode $z_R$ with distortion $D-1$. Hence

$$R(D) = \begin{cases} 2 - \log D & D \le 2 \\ 1 - ½\log(D-1) & 2 < D \le 5 \\ 0 & D > 5 \end{cases}$$ [3C]

5. (a) (i) Both expressions are equal to $H(x, e \mid y)$ decomposed in alternative ways.

(ii) This follows from the previous line since $H(e \mid y, x) = 0$ and $H(e \mid y) \le H(e)$.

(iii) Decomposition of conditional entropy:

$$\begin{aligned} E_{x,y,e} - \log p(x \mid y, e) &= \sum_e p(e = e) E_{x,y} - \log p(x \mid y, e) \\ &= \sum_e p(e = e) H(x \mid y, e) \\ &= H(x \mid y, e = 0)(1 - P_\varepsilon) + H(x \mid y, e = 1) p_e \end{aligned}$$

(iv) $H(e) \le 1$ since binary variable, $H(x \mid y, e = 0) = 0$ since when $e = 0$, $x = \hat{x} = f(y)$, $H(x \mid y, e = 1) \le |\boldsymbol{X}| - 1$ since $x$ could have any value except $\hat{x} = f(y)$. [7B]

(b) (i) Since the members of $\boldsymbol{W}$ are equiprobable, $H(w) = \log|\mathbf{W}| = nR$

(ii) Definition of mutual information

(iii) Data processing inequality since $\mathbf{x} = f(w)$

(iv) Independence bound [7B]

(v) Capacity of single-use Gaussian channel

If we divide through by $n$, then as $n \to \infty$ the second and third terms tend to zero implying that if $p_e^{(n)}$ that tends to 0, then we must have $R \le ½\log\left(1 + PN^{-1}\right)$.

Suppose that for some $n_0$ there is no lower bound to $p_e^{(n_0)}$. Then for any $\varepsilon$ and $n$ we can find a code with $p_e^{(n_0)} < n_0 n^{-1} \varepsilon$. By dividing a block of length $n$ into $n_0^{-1} n$ sub blocks of length $n_0$ and encoding each of them separately, we have a total error probability that is less than $n_0^{-1} n \times p_e^{(n0)} = \varepsilon$. Thus we can make a code with arbitrarily long block size with total error probability less than $\varepsilon$. [3A]

(c) (i) $C = 2B \times ½\log(1 + P/N) = B\log(1 + P/N)$

In our case, $P/N = 14\,\text{dB} = 25.1$ giving a capacity of 7.23 MBit/s [2C]

(ii) $P/N = 2^{C/B} - 1$

In our case $C/B = 1.56$ giving $P/N = 2.9\,\text{dB}$ [1C]

6. (a) (i) Here there are no errors so the capacity is 2 bits per transmission.

(ii) Here we cannot distinguish between inputs 1 and 2 or between 3 and 4. Thus the channel capacity is reduced to 1 bit per transmission.

(iii) Now crossovers are certain, so the capacity reverts to 2 bits. [4A]

(b) If $p_x = [p_1 \quad p_2 \quad p_3 \quad p_4]$, then we can set $a = p_3 + p_4$, $b = p_2(p_1 + p_2)^{-1}$ and $c = p_4(p_3 + p_4)^{-1}$. Clearly all these lie in the range 0 to 1. [3A]

For convenience, we define two new binary random variables: $u = \{x \text{ is even}\}$ and $v = \{x \ge 3\}$. Then

$$\begin{aligned} I(x;y) &= I(u,v;y) = I(v;y) + I(u;y\,|\,v) \\ &= H(v) - H(v\,|\,y) + (1-a)I(u;y\,|\,v=0) + aI(u;y\,|\,v=1) \\ &= H(a) - 0 + (1-a)I(x;y\,|\,v=0) + aI(x;y\,|\,v=1) \end{aligned}$$ [3A]

Note that $H(v\,|\,y) = 0$ since knowledge of $y$ implies knowledge of $v$. Also $I(u;y\,|\,v) = I(x;y\,|\,v)$ since if $v$ is known then $u$ and $x$ are equivalent.

(c) We are told to assume $I(x;y\,|\,v=0) \le 1 - H(f)$. Hence

$$\begin{aligned} I(x;y) &\le H(a) + (1-a)(1-H(f)) + a(1-H(g)) \\ &= 1 + H(a) - H(f) + a\big(H(f) - H(g)\big) = C \end{aligned}$$

with equality if $b = c = ½$.

$$\frac{dC-1}{da} = 0 \quad \Rightarrow \quad \frac{dH(a)}{da} = \log\frac{1-a}{a} = H(g) - H(f)$$
$$\Rightarrow \quad a^{-1} - 1 = 2^{H(g)-H(f)}$$
$$\Rightarrow \quad a = \left(1 + 2^{H(g)-H(f)}\right)^{-1}$$ [5A]

(d) (i) If $f = g$ then we take $a = ½$ giving a uniform $p_x$ and a capacity of $C = 2 - H(f)$ from the formula in part (d). This makes sense since we can transmit one error-free bit as $v$ and the chosen subchannel has a capacity of $1 - H(f)$. For $f = g = 0.25$ this gives $C = 1.189 \text{ bits}$.

(ii) Here $a = \left(1 + 2^{1-0}\right)^{-1} = 1/3$ and so $p_x = [2 \quad 2 \quad 1 \quad 1]/6$. The capacity is $C = 1 + H(a) - a = 1 + 0.918 - 0.333 = \log 3 = 1.585 \text{ bits}$. This is effectively a noiseless channel with ternary inputs and outputs since output 3 and 4 are indistinguishable.

(iii) This time $a = \left(1 + 2^{H(0.25)-0}\right)^{-1} = \left(1 + 2^{0.811}\right)^{-1} = 0.363$ which gives:

$$p_x = [0.318 \quad 0.318 \quad 0.182 \quad 0.182]$$
$$C = 1 + H(a) - aH(g) = 1 + 0.945 - 0.363 \times 0.811 = 1.651 \text{ bits}$$ [5A]

Thus we get a little bit more capacity than in part (ii).