# 2005 Paper E4.40, ISE4.51, SO20: Information Theory

Answer FOUR questions.

There are SIX questions on the paper.

45 minutes per question

**Information for Candidates:**

**Notation:** (a) $\log x = \dfrac{\ln x}{\ln 2}$ denotes logarithm to base 2.

(b) $\oplus$ denotes the exclusive-or operation.

(c) $X_{1:n}$ denotes the sequence $X_1, X_2, \cdots, X_n$

## The Questions

1. $X$ and $\hat{X}$ are correlated discrete Bernoulli random variables with $p(X=1)=p$ and $p(\hat{X}=1)=r$. The Hamming distance between $X$ and $\hat{X}$ is denoted by $d(X,\hat{X})=X\oplus\hat{X}$.

(a) If $E\,d(X,\hat{X})\le D$ and $D\le p\le ½$, justify each step in the following proof: [7]

$$I(X;\hat{X})\overset{\text{(i)}}{=}H(X)-H(X\mid\hat{X})\overset{\text{(ii)}}{=}H(p)-H(X\oplus\hat{X}\mid\hat{X})$$
$$\overset{\text{(iii)}}{\ge}H(p)-H(X\oplus\hat{X})\overset{\text{(iv)}}{\ge}H(p)-H(D)$$

(b) *Figure 1.1* shows a binary symmetric channel whose input is $\hat{X}$ and whose error probability is $D\le ½$. Determine the value of $r$ needed to ensure that $p(X=1)=p$. Show that for this value of *r*, $I(X;\hat{X})=H(p)-H(D)$. [5]

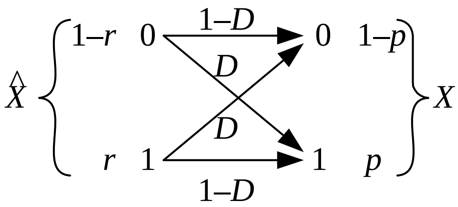



*Figure 1.1*

(c) $X$ is encoded using a rate $R=½$ code using a block length of 2, i.e. a single bit is used to encode each pair of $X$ values. If $p=0.2$, devise a code and a corresponding decoding rule that minimizes the bit-error probability. Determine the bit-error probability for your coder/decoder. [5]

(d) Determine the minimum code rate that can achieve the same bit-error probability for large block sizes. [3]

2. (a) The input and output signals of a discrete memoryless channel are given by the random variables $X$ and $Y$. Explain what is meant by the transition probability matrix, $\mathbf{Q}_{Y|X}$ , of the channel. [2]

(b) The channel is *symmetric* if all the rows and columns of $\mathbf{Q}_{Y|X}$ are permutations of the first row and first column respectively. Show that in this case [4]

(i) $H(Y|X)$ is equal to the entropy of the first row of $\mathbf{Q}_{Y|X}$ .

(ii) If $X$ is uniformly distributed then so is $Y$.

(c) Information is transmitted using an $N$-phase modem and decoded using decision boundaries placed mid-way between the uniformly constellation points. The constellation points and decision boundaries are illustrated in *Figure 2.1* for the case $N$=8.

The channel adds phase noise onto the signal in the range ±30° with the triangular probability distribution function illustrated in *Figure 2.2* and given by $p(z) = (30 - |z|)/900$. [6]

Calculate the capacity of the channel when (i) $N$=6 and (ii) $N$=8.

(d) The discrete channel is used to transmit a stream of binary bits using a block size of two symbols. Determine the maximum rate for error-free data transmission when (i) $N$=6 and (ii) $N$=8 and describe, in each case, a suitable encoding scheme for the data. [4]

(e) Calculate the differential entropy of the phase noise and hence determine the discrete-time capacity of the channel when $X$ and $Y$ take continuous values. [4]

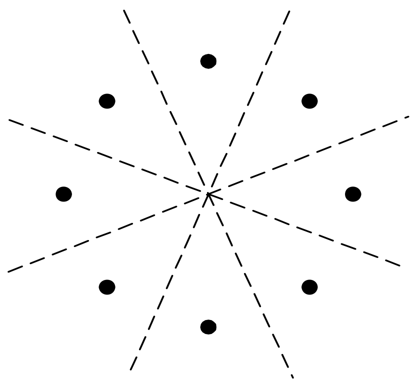


*Figure 2.1*

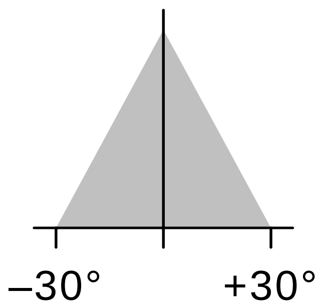



*Figure 2.2*

3. (a) (i) Explain what is meant by saying that a symbol code is (1) uniquely decodable and (2) a prefix code. [2]

(ii) Suppose that a uniquely decodable binary symbol code for an alphabet, $\boldsymbol{X}$, of size *K*, has code lengths $l_1, l_2, \cdots, l_K$ and define $M = \max(l_i)$ and $S = \sum_{i=1}^{K} 2^{-l_i}$ .

If $\mathbf{x}$ is a sequence of *N* values taken from $\boldsymbol{X}$, we define $l(\mathbf{x})$ to be the sum of the code lengths for the individual elements of $\mathbf{x}$.

Justify each of the steps in the following: [6]

$$S^N \overset{\text{(i)}}{=} \sum_{i_1=1}^{K} \sum_{i_2=1}^{K} \cdots \sum_{i_N=1}^{K} 2^{-\left(l_{i_1}+l_{i_2}+\cdots+l_{i_N}\right)} \overset{\text{(ii)}}{=} \sum_{\mathbf{x}\in\boldsymbol{X}^N} 2^{-l(\mathbf{x})}$$
$$\overset{\text{(iii)}}{=} \sum_{t=1}^{NM} 2^{-t} \,|\, \mathbf{x} : l(\mathbf{x}) = t \,| \overset{\text{(iv)}}{\leq} \sum_{t=1}^{NM} 2^{-t} 2^{t} = \sum_{l=1}^{NM} 1 = NM$$

in which |•| denotes the number of elements in a set and explain why the argument fails if the code is not uniquely decodable.

(iii) Explain why the result of part (ii) implies that $S \leq 1$. [3]

(b) (i) We wish to record the outcome of each set of a “best of 3-set” tennis match between two players A and B. The match ends when one of the players has won two sets and so there are six possible outcome sequences: AA, ABA, ABB, BAA, BAB and BB. The probability of either player winning the first set is ½, but for subsequent sets, the probability is ¾ that the player who won the previous set wins again. Calculate the entropy of the outcome sequence. [3]

(ii) Determine the expected length for each of the following methods of coding the outcome sequence of part (b) as a sequence of bits: [6]

(1) A code containing one bit for each set played that equals 0 or 1 according to whether the winner was A or B.

(2) A Shannon code in which the code length for outcome sequence *x* is equal to $p(x)^{-1}$ rounded up if necessary to an exact integer.

(3) A Huffman code.

4. *Figure 4.1* shows a diagram of a communications system in which a message $w$ is encoded as one of $2^{nR}$ possible $n$-bit code vectors $\{\mathbf{x}_i\}$ and transmitted over a symmetric binary channel. The received vector, $\mathbf{y}$, is decoded as $\hat{w}$. We define $C = I(X;Y)$ when $X$ is taken from a uniform distribution.

For any $\varepsilon > 0$, we define the joint typical set, $J_\varepsilon^{(n)}$, to be the set of vector pairs, $\{\mathbf{x}, \mathbf{y}\}$ that satisfy:

$$J_\varepsilon^{(n)} = \left\{\mathbf{x}, \mathbf{y} : \left| n^{-1} \log p(\mathbf{x}, \mathbf{y}) + H(X,Y) \right| < \varepsilon \right\}$$

where $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}) p(\mathbf{y} \mid \mathbf{x})$ is the joint probability of $\mathbf{x}$ and $\mathbf{y}$ when $p(\mathbf{x})$ is uniform.

You may assume without proof that $|J_\varepsilon^{(n)}| \le 2^{n(H(X,Y)+\varepsilon)}$ and that $n$ is large enough to ensure that $p\left(\mathbf{x}, \mathbf{y} \in J_\varepsilon^{(n)}\right) > 1 - \varepsilon$.

(a) Show that if $\mathbf{x}'$ and $\mathbf{y}'$ are chosen independently from a uniform distribution, then $p(\mathbf{x}', \mathbf{y}') = 2^{-2n}$. Hence show that the probability that they lie in $J_\varepsilon^{(n)}$ is at most $2^{-n(C-\varepsilon)}$. [6]

(b) Suppose that the encoder is formed by choosing $2^{nR}$ vectors $\{\mathbf{x}_i\}$ independently from a uniform distribution. A received vector, $\mathbf{y}$, is decoded as $\hat{w} = i$ if $\mathbf{x}_i$ is the only codeword with which it is jointly typical and as $\hat{w} = 0$ if it is jointly typical with either no codewords or more than one codeword. [7]

If the vector $\mathbf{x}_1$ is transmitted through the channel, determine upper bounds on the average probability over all possible codes that the received vector, $\mathbf{y}$, is

(i) not jointly typical with $\mathbf{x}_1$,

(ii) jointly typical with $\mathbf{x}_2$,

(iii) jointly typical with at least one of the vectors $\mathbf{x}_i$, $i \in \{2, 3, \ldots, 2^{nR}\}$.

Hence show that the average probability of a transmission error is bounded by $\varepsilon + 2^{-n(C-R-\varepsilon)}$ and demonstrate the condition under which this can be made $< 2\varepsilon$ for large enough $n$.

(c) Show that if a $(2^{nR}, n)$ code exists whose error probability averaged over all codewords is $< 2\varepsilon$, it is possible to make a $(2^{nR-1}, n)$ code in which every codeword has an error probability $< 4\varepsilon$. [7]

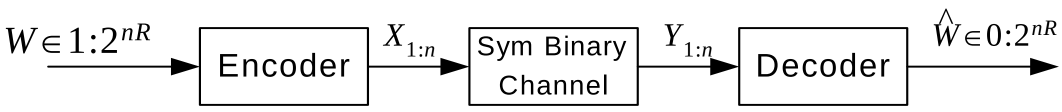



*Figure 4.1*

5. The input to a source coder is a sequence, $X_i$, of independent, identically distributed Bernoulli random variables, each with a probability $p$ of equalling 1. The vector $\mathbf{x}$, denotes the sequence $x_1, x_2, \cdots, x_n$. The notation "$\mathbf{x} < \mathbf{y}$" denotes a lexical comparison in which $x_1$ is the most significant bit and $x_n$ the least.

(a) In an Arithmetic Coder, **x** is encoded as the first $k_\mathbf{x}$ binary places of the fractional value $m_\mathbf{x} 2^{-k_\mathbf{x}}$ where $k_\mathbf{x}$ is the smallest integer for which an $m_\mathbf{x}$ exists that satisfies

$$\sum_{\mathbf{y}<\mathbf{x}} p(\mathbf{y}) \le m_\mathbf{x} 2^{-k_\mathbf{x}} < (m_\mathbf{x}+1)2^{-k_\mathbf{x}} \le \sum_{\mathbf{y}\le\mathbf{x}} p(\mathbf{y})$$

(i) Show that the expected length of the codewords is less than $H(X_{1:n})+2$. [4]

(ii) Using the table of binary values given below, determine the eight codewords for the case $p = 0.1$ and $n = 3$ and calculate the expected codeword length. Calculate also the entropy $H(X_{1:n})$. [7]

The binary values in the table have been truncated to 11 binary places. Except for values that are an exact multiple of 0.25, the truncated value is slightly less than the true value.

| Decimal | Binary | Decimal | Binary | Decimal | Binary |
|---|---|---|---|---|---|
| 0.000 | 0.0 | 0.583 | 0.10010101010 | 0.891 | 0.11100100000 |
| 0.250 | 0.01 | 0.667 | 0.10101010101 | 0.900 | 0.11100110011 |
| 0.333 | 0.01010101010 | 0.729 | 0.10111010100 | 0.981 | 0.11111011001 |
| 0.417 | 0.01101010101 | 0.750 | 0.11 | 0.990 | 0.11111101011 |
| 0.500 | 0.1 | 0.810 | 0.11001111010 | 0.999 | 0.11111111101 |

(b) If the value of $p$ is not known in advance, an adaptive arithmetic coder can be used in which $p(\mathbf{x})$ is assumed to be $\dfrac{|\mathbf{x}|!(n-|\mathbf{x}|)!}{(n+1)!}$ where $|\mathbf{x}|$ denotes the number of bits in the sequence that equal 1.

Determine the codewords for the case $n = 3$ and calculate the expected codeword length when the true value of $p = 0.1$. [7]

(c) If, unknown to the code designer, the value of $p$, changes to 0.9, calculate the expected codeword lengths of the codes from parts (b) and (c). [2]

6. *Figure 6.1* shows two parallel Gaussian channels whose noise inputs, $Z_1, Z_2$ are independent of $X_1, X_2$ and are zero-mean Gaussian with variances $N_1 = 1$ and $N_2 = 2$ respectively. The mean transmission power of each channel is given by $P_i = E\,X_i^2$. You may assume without proof that $h(Z_1) = ½\log(2\pi e N_1)$.

(a) Carefully justify the steps in the following argument and explain and justify the conditions on $X_1$ to give equality in step (vi). [6]

$$
\begin{aligned}
I(X_1;Y_1) &\overset{\text{(i)}}{=} h(Y_1) - h(Y_1 \mid X_1) \overset{\text{(ii)}}{=} h(Y_1) - h(Z_1 + X_1 \mid X_1) \\
&\overset{\text{(iii)}}{=} h(Y_1) - h(Z_1 \mid X_1) \overset{\text{(iv)}}{=} h(Y_1) - h(Z_1) \\
&\overset{\text{(v)}}{=} h(Y_1) - ½\log(2\pi e N_1) \overset{\text{(vi)}}{\le} ½\log(2\pi e(P_1 + N_1)) - ½\log(2\pi e N_1) \\
&\overset{\text{(vii)}}{=} ½\log(1 + P_1 N_1^{-1})
\end{aligned}
$$

(b) Prove that the capacity of the parallel channel is given by $\sum_{i=1}^{2} ½\log\left(1 + P_i N_i^{-1}\right)$ and determine the conditions on $X_1$ and $X_2$ for this capacity to be reached. [5]

(c) Find the maximum capacity of the parallel channel if (i) $P_1 + P_2 = 2$ and (ii) $P_1 + P_2 = 0.5$. In each case give the values of $P_1$ and $P_2$ that achieve this capacity. [5]

(d) In Figure 6.2, the two channels are reconfigured as a multi-hop transmission path with processing in the middle. Determine the maximum capacity from $X_1$ to $Y_2$ when (i) $P_1 + P_2 = 2$ and (ii) $P_1 + P_2 = 0.5$. In each case give the values of $P_1$ and $P_2$ that achieve this capacity. [4]

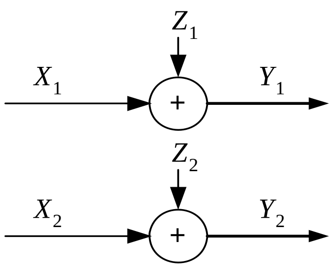



*Figure 6.1*

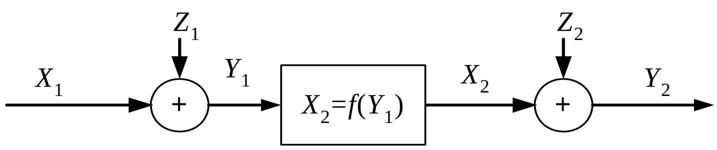



*Figure 6.2*

# 2005 E4.40/SO20 Solutions

1. (a) (i) $I(X;\hat{X}) = H(X) - X(X \mid \hat{X})$ is the definition of mutual information.

(ii) $H(X \mid \hat{X}) = H(X \oplus \hat{X} \mid \hat{X})$ since $X \oplus \hat{X}$ is a one-to-one function of $X$ for fixed $\hat{X}$ and hence has the same entropy.

(iii) $H(X \oplus \hat{X} \mid \hat{X}) \le H(X \oplus \hat{X})$ since conditioning reduces entropy [7]

(iv) $H(X \oplus \hat{X}) = H(p(X \oplus \hat{X}))$ since $X \oplus \hat{X}$ is a Bernoulli variable and $H(p(X \oplus \hat{X})) \le H(D)$ since $p(X \oplus \hat{X}) \le D \le ½$ and $H(x)$ is monotonic for $x \le ½$.

(b) We require $r(1-D) + (1-r)D = p$ from which $r = (p-D)(1-2D)^{-1}$.

$I(X;\hat{X}) = H(X) - X(X \mid \hat{X}) = H(p) - H(D)$. [5]

(c) Since we use a single bit to represent a pair of input values, we can only correctly decode two of the four possible input sequences. The other two input sequences are bound to have a single bit error. We therefore choose the code so that the wrongly coded input values have the two smallest probabilities. [5]

| **Prob** | **Input** | **Code** | **Decoded** | ***p*(error)** |
|---|---|---|---|---|
| 0.64 | 00 | 0 | 00 | 0 |
| 0.16 | 01 | 0 | 00 | ½ |
| 0.16 | 10 | 1 | 10 | 0 |
| 0.04 | 11 | 1 | 10 | ½ |

The bit error probability is $0.16 \times ½ + 0.04 \times ½ = 0.1$.

(d) From parts (a) and (b), the minimum code rate is given by: [3]

$$R = H(p) - H(D) = H(0.2) - H(0.1) = 0.7219 - 0.4690 = 0.2529$$

2. (a) The $i, j$ entry of $\mathbf{Q}_{Y|X}$ gives the conditional probability $p(Y = y_j \mid X = x_i)$. [2]

(b) (i) $$H(Y \mid X) = \sum_i p(x_i) H(Y \mid x_i) = \sum_i p(x_i) H(q_{i,:}) \overset{\text{(a)}}{=} H(q_{1,:}) \sum_i p(x_i) \overset{\text{(b)}}{=} H(q_{1,:})$$

where $q_{i,:}$ denotes the $i^{\text{th}}$ row of $\mathbf{Q}$. (a) follows because the rows of $\mathbf{Q}_{Y|X}$ are permutations of each other and (b) follows because the $p(x_i)$ sum to 1. [4]

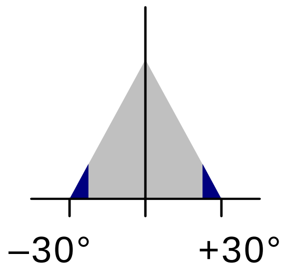
(ii) $$p(y_j) = \sum_i p(x_i) q_{i,j} \overset{\text{(a)}}{=} p(x_i) \sum_i q_{i,j} \overset{\text{(b)}}{=} p(x_i) \sum_i q_{i,1}$$

where (a) follows because $p(x_i)$ is independent of $i$, and (b) follows because the columns of $\mathbf{Q}_{Y|X}$ are permutations of each other and therefore have the same sum.

(c) (i) If $N$=6, then the constellation points are separated by 60° and the decision boundaries are at ±30° relative to them. Therefore we get perfect decoding and the capacity is $\log 6 = 2.585\ \text{bits}$.

[6]

(ii) If $N$=8, then the constellation points are separated by 45° and the decision boundaries are at ±22.5° relative to them. The dark shaded regions each have an area of 1/32, so the non-zero entries in each row of $\mathbf{Q}_{Y|X}$ are equal to [1/32 30/32 1/32] giving a conditional entropy $H(Y \mid X) = 0.3998\ \text{bits}$. Thus the capacity is $C = H(Y) - H(Y \mid X) = 3 - 0.3998 = 2.6002\ \text{bits}$, marginally more than the previous case.




(d) (i) With $N$=6, we have error-free transmission and with a block size of two, we get 36 possibilities which we can use to encode 5 bits. Thus we get an error-free capacity of 2.5 bits. [4]

(ii) With $N$=8, we can only get error free transmission by ignoring half the constellation points and transmitting 2 bits per symbol. Thus the error-free capacity is 2 bits.

(e) The differential entropy of the phase noise is given by [4]

$$h(Z) = -2 \log e \int_{-30}^{0} \left(\frac{30+z}{900}\right) \ln\left(\frac{30+z}{900}\right) dz = -1800 \log e \int_{0}^{1/30} y \ln y \, dy$$

$$= -1800 \log e \left[ \tfrac{1}{4} t^2 (2\ln(t) - 1) \right]_0^{1/30} = 1800 \log e \, \frac{2\ln(30) - 1}{3600}$$

$$= \tfrac{1}{2} \times 1.4427 \times (6.8024 - 1) = 4.1855$$

Hence the channel capacity is

$$I(X;Y) = h(Y) - h(Y \mid X) = h(Y) - h(Z) = \log 360 - h(Z) = 8.4919 - 4.1855 = 3.0876\ \text{bits}$$

3. (a) (i) Uniquely decodable means that no pair of distinct finite length sequences map onto the same code. In other every finite-length sequence of symbols maps onto a distinct code sequence. [2]

(ii) A prefix code is one in which no code word is a prefix of another. This implies that the end of a codeword can be determined as soon as it occurs. [2]

(b) (i)

$$S^N = \left(\sum_{i_1} 2^{-l_{i1}}\right)\left(\sum_{i_2} 2^{-l_{i2}}\right)\cdots\left(\sum_{i_N} 2^{-l_{iN}}\right) = \sum_{i_1}\sum_{i_2}\cdots\sum_{i_N} 2^{-l_{i1}}2^{-l_{i2}}\cdots 2^{-l_{iN}}$$
$$= \sum_{i_1}\sum_{i_2}\cdots\sum_{i_N} 2^{-(l_{i1}+l_{i2}+\ldots+l_{iN})}$$

[6]

(ii) The multiple sum exactly includes all possible sequences from $\boldsymbol{X}$ of length $N$.

(iii) We can divide the sequences of length $N$ into subsets having the same total code length $l(\mathbf{x})$. The maximum code length is $NM$ since the maximum symbol code length is $M$.

(iv) Provided that the code is uniquely decodable, no two code sequences of length $t$ can be identical and there are a maximum of $2^t$ such sequences. It follows that $|\mathbf{x}: l(\mathbf{x}) = t| \le 2^t$.

(v) The remaining two steps are just algebra.

(c) Part (b) must be true for all $N$. If $S > 1$, then $S^N$ increases exponentially with $N$ and must, for large enough $N$ exceed $NM$ which increases only linearly. It follows that $S$ must be $\le 1$. [3]

(d,e) The table gives the lengths of the different codes:

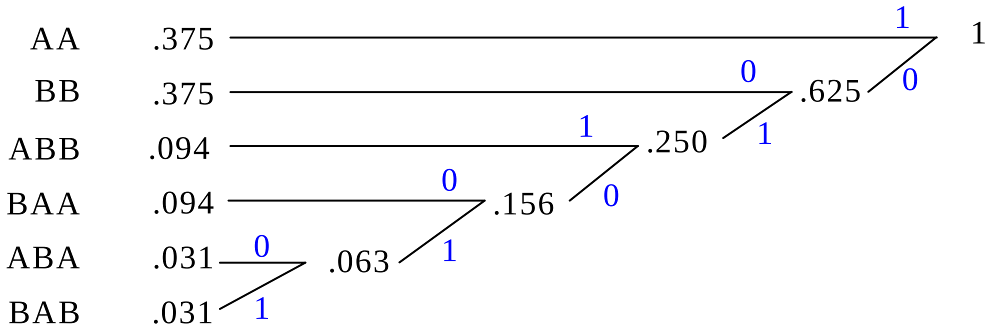
[3]

| Outcome | Prob×32 | Direct | Shannon | Huffman |
|---|---|---|---|---|
| AA | 12 | 2 | 2 | 1 |
| ABA | 1 | 3 | 5 | 3 |
| ABB | 3 | 3 | 4 | 4 |
| BAA | 3 | 3 | 4 | 5 |
| BAB | 1 | 3 | 5 | 5 |
| BB | 12 | 2 | 2 | 2 |
| Ent/Len | 2.0141 | 2.25 | 2.5625 | 2.2188 |

We note that the direct code is better than the Shannon code and only slightly worse than the Huffman code.

(e) See table for direct and Shannon code lengths [6]

Huffman derivation:




[5]

4. (a) Since $X'$ and $Y'$ are uniform, we have $H(X') = H(Y') = H(X) = H(Y) = 1$. Since they are independent, we have

$$\log p(\mathbf{x}', \mathbf{y}') = \log p(\mathbf{x}') + \log p(\mathbf{y}') = -n(H(X') + H(Y'))\,.$$

$$p(\mathbf{x}', \mathbf{y}' \in J_\varepsilon^{(n)}) \le \max(|J_\varepsilon^{(n)}|)\, p(\mathbf{x}', \mathbf{y}') = 2^{n(H(X,Y)+\varepsilon)} 2^{-n(H(X')+H(Y'))}$$
$$= 2^{n(H(X,Y)-H(X)-H(Y)+\varepsilon)} = 2^{-n(C-\varepsilon)}$$ [6]

(b) (i) We are given that $p\left(\mathbf{x}, \mathbf{y} \in J_\varepsilon^{(n)}\right) > 1 - \varepsilon$. Hence the probability that $\mathbf{x}_1$ and $\mathbf{y}$ are not typical is $< \varepsilon$.

(ii) Since $\mathbf{x}_2$ is independent of $\mathbf{x}_1$, it is also independent of $\mathbf{y}$. Hence the probability that they are jointly typical is $\le 2^{-n(C-\varepsilon)}$. [7]

(iii) There are $2^{nR} - 1$ possible wrong codes, so that probability that $\mathbf{y}$ is jointly typical with any one of them is at most $2^{nR} 2^{-n(C-\varepsilon)} = 2^{-n(C-R-\varepsilon)}$

The probability that either $\mathbf{y}$ is not typical with $\mathbf{x}_1$ or that it is typical with an incorrect $\mathbf{x}_i$ is less than $\varepsilon + 2^{-n(C-R-\varepsilon)}$. Provided that $R < C - \varepsilon$, the exponent in the second term is negative and for large enough *n* the term can be made $< \varepsilon$, thus making the whole probability $< 2\varepsilon$.

(c) To do this, we delete half the codewords, leaving behind only those with the lowest error rates. Assume that we re-order the codewords to have decreasing error probabilities and that the error probability of codeword $i$ is $\lambda_i$. We have [7]

$$2\varepsilon \ge 2^{-nR} \sum_{i=1}^{2^{nR}} \lambda_i \ge 2^{-nR} \sum_{i=1}^{2^{nR-1}} \lambda_i \ge 2^{-nR} \sum_{i=1}^{2^{nR-1}} \lambda_{2^{nR-1}} \ge ½ \lambda_{2^{nR-1}}$$
$$\Rightarrow \quad \lambda_{2^{nR-1}} \le 4\varepsilon \quad \Rightarrow \quad \lambda_i \le 4\varepsilon \quad \forall\, i > 2^{nR-1}$$

Thus we have only half as many codewords ($2^{nR-1}$), but each one has an error probability less than $4\varepsilon$.

5. (a) (i) We can set $k_{\mathbf{x}} = \lceil -\log p(\mathbf{x}) \rceil + 1$ and $m_{\mathbf{x}} = \left\lceil 2^{k_{\mathbf{x}}} \sum_{\mathbf{y}<\mathbf{x}} p(\mathbf{y}) \right\rceil$. We must have $m_{\mathbf{x}} < 1 + 2^{k_{\mathbf{x}}} \sum_{\mathbf{y}<\mathbf{x}} p(\mathbf{y})$ and the required inequality follows directly. Since $k_{\mathbf{x}} = \lceil -\log p(\mathbf{x}) \rceil + 1 < 2 - \log p(\mathbf{x})$, we must have $E\, k_{\mathbf{x}} < 2 + H(X_{1:n})$. Note that it may be possible to reduce $k_{\mathbf{x}}$ by 1 for some values of **x**, but this can only reduce the expected codeword length. [4]

(ii) We can make the following table: [7]

| **x** | $p(\mathbf{x})$ | $\sum_{\mathbf{y}<\mathbf{x}} p(y)$ | Binary | Code | Len |
|---|---|---|---|---|---|
| 000 | 0.729 | 0.0 | 0.0 | 0 | 1 |
| 001 | 0.081 | 0.729 | 0.10111010100 | 11000 | 5 |
| 010 | 0.081 | 0.810 | 0.11001111010 | 1101 | 4 |
| 011 | 0.009 | 0.891 | 0.11100100000 | 11100101 | 8 |
| 100 | 0.081 | 0.900 | 0.11100110011 | 11101 | 5 |
| 101 | 0.009 | 0.981 | 0.11111011001 | 11111100 | 8 |
| 110 | 0.009 | 0.990 | 0.11111101011 | 11111110 | 8 |
| 111 | 0.001 | 0.999 | 0.11111111101 | 1111111111 | 10 |

The expected codeword length is 2.089 bits. Since the bits are i.i.d., we can calculate $H(X_{1:n}) = nH(0.1) = 3 \times 0.469 = 1.407$ which is quite a bit less than our arithmetic coder but well within the limit calculated in part (i).

(b) Our table now looks like: [7]

| **x** | " $p(\mathbf{x})$ " | $\sum_{\mathbf{y}<\mathbf{x}} p(y)$ | Binary | Code | Len |
|---|---|---|---|---|---|
| 000 | 0.25 | 0.0 | 0.0 | 00 | 2 |
| 001 | 0.083 | 0.25 | 0.01000000000 | 0100 | 4 |
| 010 | 0.083 | 0.333 | 0.01010101010 | 01011 | 5 |
| 011 | 0.083 | 0.417 | 0.01101010101 | 01110 | 5 |
| 100 | 0.083 | 0.500 | 0.10000000000 | 1000 | 4 |
| 101 | 0.083 | 0.583 | 0.10010101010 | 10011 | 5 |
| 110 | 0.083 | 0.667 | 0.10101010101 | 1011 | 4 |
| 111 | 0.25 | 0.750 | 0.11000000000 | 11 | 2 |

The expected codeword length is now 2.639 bits (using the correct values of $p(\mathbf{x})$ from part (b)).

(c) The list of probabilities is now reversed and the expected lengths become 9.361 and 2.711 respectively. Thus, the adaptive coder is much more robust to changes in *p*. [2]

6. (a)

$$\begin{aligned} I(X_1;Y_1) &\overset{\text{(i)}}{=} h(Y_1) - h(Y_1 \mid X_1) \overset{\text{(ii)}}{=} h(Y_1) - h(Z_1 + X_1 \mid X_1) \\ &\overset{\text{(iii)}}{=} h(Y_1) - h(Z_1 \mid X_1) \overset{\text{(iv)}}{=} h(Y_1) - h(Z_1) \\ &\overset{\text{(v)}}{=} h(Y_1) - ½\log(2\pi e N_1) \overset{\text{(vi)}}{\le} ½\log(2\pi e(P_1 + N_1)) - ½\log(2\pi e N_1) \\ &\overset{\text{(vii)}}{=} ½\log(1 + P_1 N_1^{-1}) \end{aligned}$$

[6]

Reasons: (i) definition of mutual information, (ii) action of channel, (iii) translation independence, (iv) $Z_1$ independent of $X_1$, (v) entropy of $Z_1$, (vi) Gaussian has maximum entropy for a given variance, (vii) log of quotient.

We have equality at (vi) if $Y_1$ is Gaussian which in turn is true iff $X_1$ is Gaussian.

(b) Taking **x** to be the vector $X_{1:2}$,

$$\begin{aligned} I(\mathbf{x};\mathbf{y}) &\overset{(h)}{=} h(\mathbf{y}) - h(\mathbf{y} \mid \mathbf{x}) \overset{(i)}{=} h(\mathbf{y}) - h(\mathbf{z}) \\ &\overset{(j)}{=} h(Y_1) + h(Y_2 \mid Y_1) - h(Z_1) - h(Z_2) \\ &\overset{(k)}{\le} h(Y_1) + h(Y_2) - h(Z_1) - h(Z_2) \\ &\overset{(l)}{=} I(X_1;Y_1) + I(X_2;Y_2) \\ &\overset{(m)}{\le} ½\log(1 + P_1 N_1^{-1}) + ½\log(1 + P_2 N_2^{-1}) \end{aligned}$$

[5]

Reasons: (h) definition of mutual information, (i) same as (b),(c),(d) above, (j) chain rule and independence of $Z_i$, (k) Conditioning reduces entropy, (l) definition of mutual information, (m) same as (a) to (g) above.

Equality at (k) if $Y_1$ and $Y_2$ are independent which in turn is iff $X_1$ and $X_2$ are independent. Equality at (m) if $X_1$ and $X_2$ are Gaussian.

(c) (i) We use waterfilling to make $P_1 + N_1 = P_2 + N_2$. This gives $P_1 = 1.5, P_2 = 0.5$ and a total capacity of $½\log\big((1+1.5)(1+0.25)\big) = 0.822\text{ bits}$. [5]

(ii) In this case we only use $X_1$ and the capacity is $½\log\big(1+0.5\big) = 0.293\text{ bits}$.

(d) Our optimum strategy is here to make the capacity of each hop the same which means that they must have the same SNR. Thus $P_2 = 2P_1$ and the capacity is $½\log(1 + P_1 N_1^{-1})$. [4]

(i) $P_1 = 0.667$ giving a capacity of 0.369 bits.

(ii) $P_1 = 0.167$ giving a capacity of 0.111 bits.