

Examination: Information Theory

MODEL ANSWER and MARKING SCHEME

2003

First Examiner: Professor L F Turner

Paper Code: E4.40 / S020

Second Marker: Dr J A Barria

Question

Page 1 of 23

1

The transition matrix is as follows

$$\begin{array}{c}
 \text{Following Digit} \\
 \begin{array}{cc}
 0 & 1 \\
 \text{Preceding Digit} & \begin{bmatrix} .9 & .1 \\ .7 & .3 \end{bmatrix}
 \end{array}
 \end{array}$$

Given that the source starts in the 0 state the following sequence of events occurs leading to the state probabilities

$$1^{\text{st}} \text{ Digit Generated } \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} .9 & .1 \\ .7 & .3 \end{bmatrix} \Rightarrow \begin{bmatrix} .9 & .1 \end{bmatrix} \text{ state after } 1^{\text{st}}$$

$$2^{\text{nd}} \text{ Digit Generated } \begin{bmatrix} .9 & .1 \end{bmatrix} \begin{bmatrix} .9 & .1 \\ .7 & .3 \end{bmatrix} \Rightarrow \begin{bmatrix} .81 & .12 \end{bmatrix} \text{ state after } 2^{\text{nd}}$$

$$3^{\text{rd}} \text{ Digit Generated } \begin{bmatrix} .81 & .12 \end{bmatrix} \begin{bmatrix} .9 & .1 \\ .7 & .3 \end{bmatrix} \Rightarrow \begin{bmatrix} .876 & .124 \end{bmatrix} \text{ state after } 3^{\text{rd}}$$

$$4^{\text{th}} \text{ Digit Generated } \begin{bmatrix} .876 & .124 \end{bmatrix} \begin{bmatrix} .9 & .1 \\ .7 & .3 \end{bmatrix} \Rightarrow \begin{bmatrix} .8752 & .1248 \end{bmatrix} \text{ state after } 4^{\text{th}}$$

$$5^{\text{th}} \text{ Digit Generated } \begin{bmatrix} .8752 & .1248 \end{bmatrix} \begin{bmatrix} .9 & .1 \\ .7 & .3 \end{bmatrix} \Rightarrow \begin{bmatrix} .87504 & .12496 \end{bmatrix}$$

This agrees with the steady-state probabilities which from Markov theory are $\frac{.7}{.8} = .875$ and $\frac{.1}{.8} = .125$

Examination: Information Theory

MODEL ANSWER and MARKING SCHEME

First Examiner: Professor L F Turner

Paper Code: E4.40 / S020

Second Marker: Dr J A Barria

Question

Page 2 of 23

Now the entropy of the source when in state s_k is given by

$$H(x/s_k) = - \sum_i p(x_i/s_k) \log_2 p(x_i/s_k)$$

and the average entropy

$$= \sum_k P(s_k) H(x/s_k) = - \sum_k P(s_k) \sum_i p(x_i/s_k) \log_2 p(x_i/s_k)$$

We have two states $s_0 (= 0)$; $s_1 (= 1)$ and the conditional entropies are

$$-p(x_0/s_0) \log_2 p(x_0/s_0) - p(x_1/s_0) \log_2 p(x_1/s_0) \quad (1)$$

$$\text{and } -p(x_0/s_1) \log_2 p(x_0/s_1) - p(x_1/s_1) \log_2 p(x_1/s_1) \quad (2)$$

These are determined solely by the transition probabilities, so we have

$$H(x/s_0) = -(.9 \log_2 .9 + .1 \log_2 .1) = 0.4684$$

$$H(x/s_1) = -(.7 \log_2 .7 + .3 \log_2 .3) = 0.9812.$$

The entropy as the source generates its symbols are thus the above, weighted by the appropriate $P(s_0)$ and $P(s_1)$ as calculated on previous page.

$$\text{We thus get } H_1 = 1 \times 0.4684 + 0 \times 0.9812 = 0.4684 \text{ bits/symbol}$$

$$H_2 = .88 \times 0.4684 + .12 \times 0.9812 = 0.5179 \quad \dots$$

Examination: Information Theory
MODEL ANSWER and MARKING SCHEME

First Examiner: Professor L F Turner

Paper Code: E4.40 / S020

Second Marker: Dr J A Barria

Question

Page 3 of 23

$$H_3 = .876 \times .4684 + .124 \times .8812 = 0.5196$$

$$H_4 = .8752 \times .4684 + .1248 \times .8812 = 0.52$$

$$H_5 = .875 \times .4684 + .125 \times .8812 = 0.52$$

$H_5 = .52$ bits/binary digit in the steady-state entropy.

16 marks

If source is considered as memoryless then the probability of generating 0 and 1 will be $P(0) = .875$ and $P(1) = .125$
Hence the zero-order entropy is

$$H(X)_{\text{memoryless}} = -.875 \log_2 .875 - .125 \log_2 .125 = 0.545 \text{ bits/binary digit}$$

3 marks

If we encode using a Shannon/Fano code and we encode pairs of digits we get

	prob	code
$X_0 (00)$	$49/64$	1
$X_1 (01)$	$1/64$	0 1
$X_2 (10)$	$7/64$	0 0 1
$X_3 (11)$	$1/64$	0 0 0

3 marks

So average codeword length is $\frac{49}{64} \times 1 + \frac{1}{64} + \frac{21}{64} + \frac{3}{64} = \frac{87}{64}$

Hence average/digit = 0.68 binary digits/digit

Examination: Information Theory

MODEL ANSWER and MARKING SCHEME

First Examiner: Professor L F Turner

Paper Code: E4.40 / S020

Second Marker: Dr J A Barria

Question

Page 4 of 23

Conclusions :

- (1) Memoryless entropies close to actual entropy
- (2) If encode with s/p then need to encode more than 2 digits at a time to approach memoryless entropies
- (3) If encode long enough block then will approach true entropy

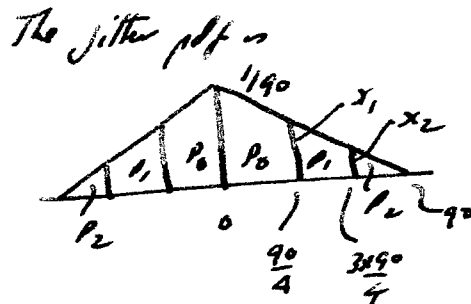
(3 marks)

Total mark (25)

MODEL ANSWER and MARKING SCHEME

Paper Code: E4.40 / S020

Page 5 of 23



The equation is $P(\theta) = -\frac{1}{8100}\theta + \frac{1}{90}$ for $0 \leq \theta \leq 120$.

So we have $x_1 = \frac{3}{4 \times 90}$; $x_2 = \frac{1}{4 \times 90}$

Hence $P_0 = 7/32$, $P_1 = 2/32$; $P_2 = 1/32$

Now 2% is the probability that jitter will not be sufficient to cause symbol error, i.e. in the probability that symbol will be decoded as an adjacent symbol and P_2 is probability that symbol will be decoded as a symbol two symbols away so we have a channel transition matrix as follows

Examination: Information Theory
MODEL ANSWER and MARKING SCHEME

First Examiner: Professor L F Turner

Paper Code: E4.40 / S020

Second Marker: Dr J A Barria

Question

Page 6 of 23

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
x_1	$2P_0$	P_1	P_2	0	0	0	P_2	P_1
x_2	P_1	$2P_0$	P_1	P_2	0	0	0	P_2
x_3	P_2	P_1	$2P_0$	P_1	P_2	0	0	0
x_4	0	P_2	P_1	$2P_0$	P_1	P_2	0	0
x_5	0	0	P_2	P_1	$2P_0$	P_1	P_2	0
x_6	0	0	0	P_2	P_1	$2P_0$	P_1	P_2
x_7	P_2	0	0	0	P_2	P_1	$2P_0$	P_1
x_8	P_1	P_2	0	0	0	P_2	P_1	$2P_0$

So, channel is doubly symmetric and hence capacity is

$$C = \log_2 K + \sum_{i=1}^K q_i \log_2 q_i$$

where K is number of output symbols and q_i 's are elements of row of matrix.

Hence we obtain

$$C = \log_2 8 + 2P_0 \log_2 2P_0 + 2P_1 \log_2 P_1 + 2P_2 \log_2 P_2$$

on substituting $P_0 = 1/32$, $P_1 = 3/32$

and $P_2 = 1/32$ we arrive at

$$C = 3 - .522 - 1 - .3125$$

$$= 1.165 \text{ bits/transmitted symbol}$$

20 marks

Examination: Information Theory

MODEL ANSWER and MARKING SCHEME

First Examiner: Professor L F Turner

Paper Code: E4.40 / S020

Second Marker: Dr J A Barria

Question

Page 8 of 23

3

Part 1 $I = H(X) - H(X|Y)$

This can be interpreted as the difference between the a priori and a posteriori uncertainty. We are $H(X)$ uncertain about the input to the channel before receiving Y , and are $H(X|Y)$ uncertain about the input after having received Y .

$H(X|Y)$ could be interpreted as the amount of information still needed in order to be certain as to the input.

$$I = H(X) - H(X|YZ) \\ = H(X) - H(X|Y) + \{H(X|Y) - H(X|YZ)\} \quad (*)$$

This can be related to the situation in which we send x and receive Y followed by Z . The sequence pairs in \mathcal{D} can be interpreted in terms of successive reductions in our uncertainty about the input.

i.e. $H(X) - H(X|Y)$ is the reduction in uncertainty about x once we have received Y . We remain $H(X|Y)$ uncertain about x .

Limitations. Although we can argue in terms

Examination: Information Theory

MODEL ANSWER and MARKING SCHEME

First Examiner: Professor L F Turner

Paper Code: E4.40 / S020

Second Marker: Dr J A Barria

Question

Page 9 of 23

of the reduction in uncertainty, and we can argue through the capacity theorem that $I = H(X) - H(X|Y)$ (or its maximum) can be shown to be the rate at which information can be communicated error free, we cannot argue anything about certainty of information reception through an argument relating to the reduction in uncertainty. To achieve error-free communication we have to employ channel coding.

8 marks

Part 2 Consider $I = H(X) - H(X|Y)$,

$$\text{where } H(X) = - \sum_i P(x_i) \log P(x_i)$$

$$H(X|Y) = - \sum_i \sum_j P(x_i, y_j) \log P(x_i|y_j)$$

$$\therefore I = \sum_i \sum_j P(x_i, y_j) \log P(x_i|y_j) / P(x_i)$$

$$= \sum_i \sum_j P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(y_j) \cdot P(x_i)}$$

$$= \sum_i \sum_j P(x_i|y_j) P(y_j) \log P(y_j|x_i) / P(y_j)$$

$$= \sum_j \sum_i P(y_j) P(x_i|y_j) (\log 1/P(y_j) - \sum_i P(y_j) P(x_i|y_j) \log 1/P(x_i|y_j))$$

Examination: Information Theory

MODEL ANSWER and MARKING SCHEME

First Examiner: Professor L F Turner

Paper Code: E4.40 / S020

Second Marker: Dr J A Barria

Question

Page 10 of 23

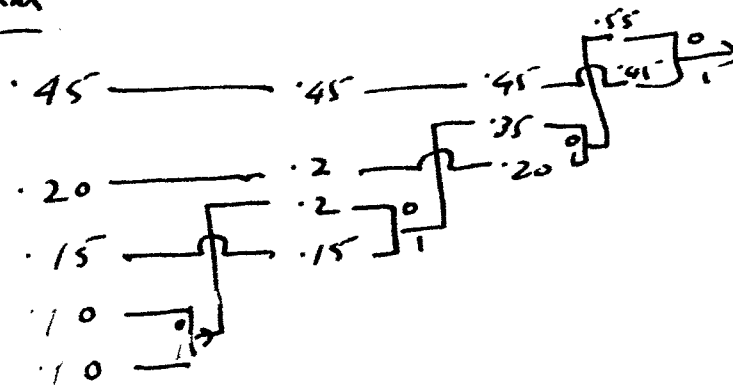
$$= H(Y) - H(Y/X)$$

5 marks

Part 3 Consider by way of example
(will accept any sensible example)

	Prob	S/F
x_1	.45	0
x_2	.20	1 0
x_3	.15	1 1 0
x_4	.10	1 1 1 0
x_5	.10	1 1 1 1

Huffman



Which yields the Huffman code

$x_1 \rightarrow 1$
 $x_2 \rightarrow 01$
 $x_3 \rightarrow 001$
 $x_4 \rightarrow 0000$
 $x_5 \rightarrow 0001$

Both are non-prefix codes (instantaneous). The codewords in this case are of the same length and hence the codes are of the same efficiency. The codes have different code words but that does not matter.

The Huffman technique is the optimum, although more often than not there is no difference in performance between the SFT and H techniques.

The most general and easiest way of implementing is to use ROM look-ups for both encoding and decoding.

With this technique the inputs x_i are used to address the ROM and the codewords are stored in the ROM locations. At the receiver the code words are used as addresses and the symbols x_i are located in the corresponding address locations.



5 marks

First Examiner: Professor L F Turner

Paper Code: E4.40 / S020

Second Marker: Dr J A Barria

Question

Page 12 of 23

Part 4

Let the source symbols be x_1, \dots, x_K and let the associated probabilities be p_1, \dots, p_K in which $p_1 \geq p_2 \geq p_3 \dots \geq p_{K-1} \geq p_K$

Let the corresponding codewords be of length n_1, \dots, n_K respectively.

Now suppose some $n_i \neq n_K$ is longer than n_K . It is immediately clear that by interchanging n_i and n_K

(1) no change in code structure occurs so it remains instantaneous and hence uniquely decodable

(2) the average codeword length is $\sum p_i n_i$ and that this will be reduced by the change, so the code becomes more efficient i.e. closer to optimum.

Thus the code can always be re-arranged so that the longer codewords are associated with the smaller source symbol probabilities and hence after re-arrangement the two least-probable codewords will be the longest.

Examination: Information Theory
MODEL ANSWER and MARKING SCHEME

First Examiner: Professor L F Turner

Paper Code: E4.40 / S020

Second Marker: Dr J A Barria

Question

Page 13 of 23

Now consider N_{k-1} and N_k (the re-arranged & longest) (7 marks)

If N_k is longer than N_{k-1} then on account of the prefix property of the instantaneous code the extra digit in N_k can be dropped.

That 25 marks

Examination: Information Theory
MODEL ANSWER and MARKING SCHEME

First Examiner: Professor L F Turner

Paper Code: E4.40 / S020

Second Marker: Dr J A Barria

Question

Page 14 of 23

4

Part 1 The statement in itself does not constitute a restriction since we can send as many data bits/pulses as we like by increasing the size of the symbol set. But if we are power constrained then the symbols get closer together and hence if noise is present then the error probability increases

↑
(3 marks)

Part 2

$$C = \frac{I}{\max_{P(X)}} = \max_{P(X)} [H(X) - H(X|Y)] \text{ or } \max_{P(X)} [H(Y) - H(Y|X)]$$

I will except a solution in Sheet I is expressed as a function of $P(x_1), P(x_2)$ - the input probabilities - and the function I is maximised by obtaining

$$\frac{\partial I}{\partial P(x_1)} = 0; \quad \frac{\partial I}{\partial P(x_2)} = 0 \quad \text{with the}$$

solution being obtained for $P(x_1), P(x_2)$.

which satisfies the above subject to the constraint $P(x_1) + P(x_2) = 1$

Examination: Information Theory
MODEL ANSWER and MARKING SCHEME

First Examiner: Professor L F Turner

Paper Code: E4.40 / S020

Second Marker: Dr J A Barria

Question

Page 15 of 23

Will also accept the following alternative.

$$I = I(Y) - H(Y/X)$$

with channel matrix $\begin{matrix} & y_1 & y_2 \\ x_1 & \begin{bmatrix} 1-p & p \end{bmatrix} \\ x_2 & \begin{bmatrix} p & 1-p \end{bmatrix} \end{matrix}$

$$\begin{aligned} \text{Now } H(Y/X) &= -\sum_{j=1}^2 P(x_j) [p \log p + (1-p) \log (1-p)] \\ &= -(p \log p + (1-p) \log (1-p)) \end{aligned}$$

$$\text{and } H(Y) = -\{P(y_1) \log P(y_1) + P(y_2) \log P(y_2)\}$$

$$\begin{aligned} \text{But } P(y_1) &= P(x_1) P(y_1/x_1) + P(x_2) P(y_1/x_2) \\ &= P(x_1)(1-p) + P(x_2) \cdot p \end{aligned}$$

$$P(y_2) = P(x_1) \cdot p + (1-p) P(x_2)$$

$$\text{If } P(x_1) = P(x_2) \text{ then } P(y_1) = P(y_2)$$

$$\text{hence } H(Y) = 1 \text{ and}$$

$$\underline{I = 1 + p \log p + (1-p) \log (1-p)}$$

(13 marks)
13

First Examiner: Professor L F Turner

Paper Code: E4.40 / S020

Second Marker: Dr J A Barria

Question

Page 16 of 23

Part 3. In order to transmit at capacity it is necessary that the input binary digits to the channel be used with equal probabilities

Now the source is

x_1	→	0	$1/3$
x_2	→	1	$1/3$
x_3	→		$1/3$

Consider the following coding in which the x_i 's are mapped into binary digits

Case I

x_1	00	$\therefore P(0) = P(1) = 1/2$
x_2	10	\therefore looks perfect
x_3	11	

But Entropy of source $= \log_2 3 = 1.59$ bits/symbol

We are using 2 bits/symbol \therefore wasteful

Case II

Using S/F code

x_1	0
x_2	10
x_3	11

$P(0) = 0.4$

$P(1) = 0.6$

\therefore Prob match to channel
- we require $P(0) = P(1) = \frac{1}{2}$

First Examiner: Professor L F Turner

Paper Code: E4.40 / S020

Second Marker: Dr J A Barria

Question

Page 17 of 23

Case III S/F applied to extended symbol set

prob		SF code
$1/9$	$x_1 x_1$	000
	$x_1 x_2$	001
	$x_1 x_3$	010
	$x_2 x_1$	011
	$x_2 x_2$	100
	$x_2 x_3$	101
	$x_3 x_1$	110
	$x_3 x_2$	1110
$1/9$	$x_3 x_3$	1111

$$\therefore P(0) = 0.45 \quad \left. \begin{array}{l} P(1) = 0.55 \end{array} \right\} \text{getting closer to } 1/2$$

$$\begin{aligned} \text{Average code word length} \\ &= \frac{29}{9} \text{ / pair} \\ &= \frac{29}{18} = 1.61 \text{ / symbol} \end{aligned}$$

close to entropy

(9 marks)

If we repeat for coding of larger blocks of x_i 's
then $P(0)$ and $P(1)$ approach $1/2$
and $\bar{n} \rightarrow H(x)$

The point need is for proper channel codes
that is inserted between source code output
and channel to achieve reliable communication
can be needed to ensure that channel code
gives $P(0) = P(1) = 1/2$ at its output.

(Total mark = 25)

5

Part 1 The sequence $S_i = x_{i_1}, x_{i_2}, \dots, x_{i_N}$, where each x_{i_j} is chosen from the set of symbols x_1, x_2, \dots, x_n .

$$\begin{aligned} \text{Thus } \tilde{H}(S_i) &= - \sum_{i=1}^n P(S_i) \log P(S_i) \\ &= - \sum_{i_1=1}^n \dots \sum_{i_N=1}^n P(x_{i_1}, \dots, x_{i_N}) \log P(x_{i_1}, \dots, x_{i_N}) \quad (1) \end{aligned}$$

$$\text{But } P(x_{i_1}, \dots, x_{i_N}) = P(x_{i_1}, \dots, x_{i_M}) \cdot P\left(\frac{x_{i_{M+1}}}{x_{i_1}, \dots, x_{i_M}}\right)$$

Hence on substituting in (1) we obtain

$$\tilde{H}(S_i) = - \sum_{i_1=1}^n \dots \sum_{i_N=1}^n P(x_{i_1}, \dots, x_{i_N}) \left[\log P(x_{i_1}, \dots, x_{i_N}) + \log P\left(\frac{x_{i_{M+1}}}{x_{i_1}, \dots, x_{i_M}}\right) + \dots + \log P\left(\frac{x_{i_N}}{x_{i_{N-M}}, \dots, x_{i_{N-1}}}\right) \right]$$

$$\text{But } - \sum_{i_1=1}^n \dots \sum_{i_N=1}^n P(x_{i_1}, \dots, x_{i_N}) \log P\left(\frac{x_{i_j}}{x_{i_{j-M}}, \dots, x_{i_{j-1}}}\right) = H(x)$$

$j = M+1, \dots, N$

where $H(x)$ is the true entropy of the memory source.

Thus we have

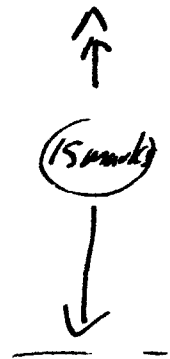
$$\begin{aligned}\tilde{H}(\sigma) &= (N-M) H(x) - \sum_{i=1}^n \dots \sum_{i_M=1}^n P(x_i, \dots, x_{i_M}) \log P(x_i, \dots, x_{i_M}) \\ &= (N-M) H(x) - \sum_{i=1}^n \dots \sum_{i_M=1}^n P(x_i, \dots, x_{i_M}) \log P(x_i, \dots, x_{i_M})\end{aligned}$$

But since $-\sum p_i \log p_i \geq 0$ we have the

$$\text{result } \tilde{H} = (N-M) H(x) + \delta \quad (2)$$

$$\delta \geq 0$$

and for $N > M$, δ is independent of N .



Part 2 Proof of Shannon's noiseless coding theorem for memory source

Take a block S_i of source symbols, which have probability $P(S_i)$

We then know that can select code word of length L_i which satisfies the Kraft/McMillan equality provided

$$\log 1/P(S_i) \leq L_i < \log 1/P(S_i) + 1 \quad (2)$$

Hence we have

$$\sum_{i=1}^N P(s_i) \log \frac{1}{P(s_i)} \leq \left[\sum_{i=1}^N P(s_i) \right] < \sum_{i=1}^N P(s_i) \log \frac{1}{P(s_i)} + \sum_{i=1}^N P(s_i)$$

Now from (2) this can be written

$$(N-M)H(x) + \delta \leq N_{ave} < (N-M)H(x) + \delta + 1$$

on dividing by N , the block length, we get

$$H(x) \leq \frac{N_{ave}}{N} = n_{ave} < H(x) + \frac{1}{N} \rightarrow 0$$

10 marks

as $N \rightarrow \infty$.

$$\therefore n_{ave} = H(x) = \text{entropy}$$

same

Implementation is generally impossible in practice since codebook size grows exponentially with N .

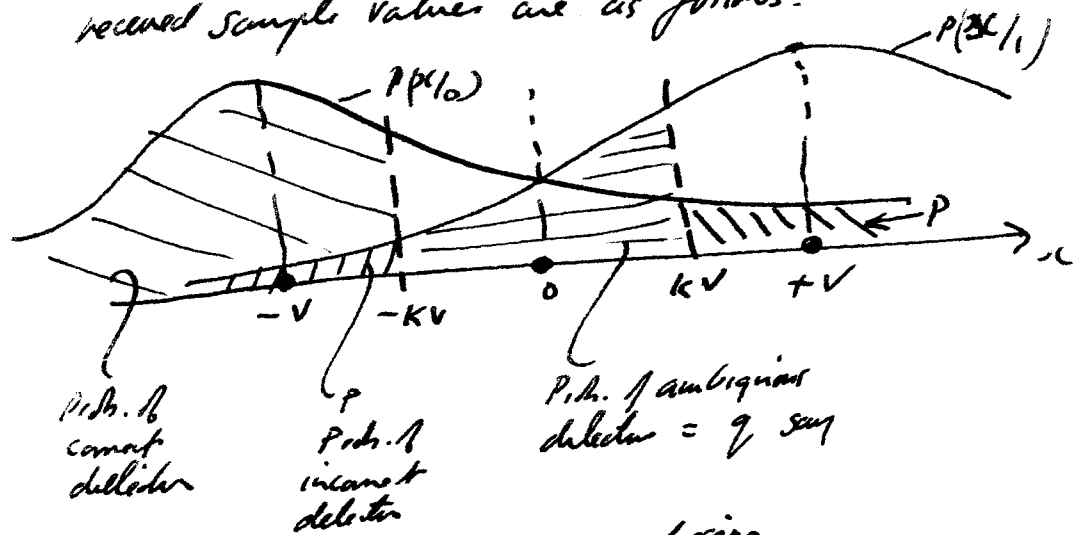
e.g. 3-bit per sample (x_1, \dots, x_{256})

So if attempt to encode 3 samples at a time

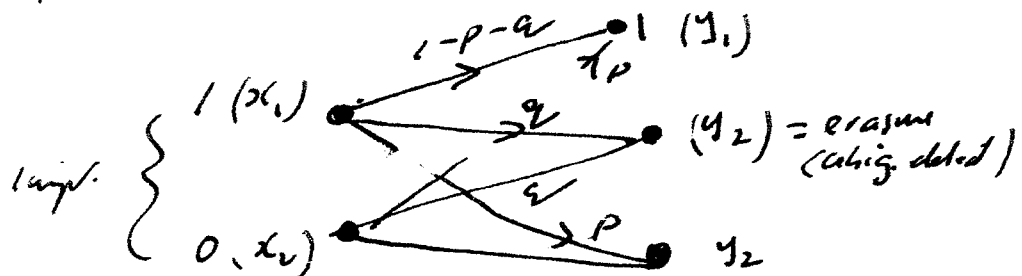
$$\text{codebook} = (256)^3 \quad !!$$

Total mark = 25

Assume that cable attenuation is neglected.
In this case the conditional pdf's relating to the received sample values are as follows:



Thus the channel transition ~~matrix~~ diagram is



and the channel transition matrix is

$$\begin{matrix} & y_1 & y_2 & y_3 \\ x_1 & \begin{bmatrix} 1-p-q & p & q \end{bmatrix} \\ x_2 & \begin{bmatrix} p & (1-p-q) & q \end{bmatrix} \end{matrix}$$

Examination: Information Theory
MODEL ANSWER and MARKING SCHEME

First Examiner: Professor L F Turner

Paper Code: E4.40 / S020

Second Marker: Dr J A Barria

Question

Page 22 of 23

Now the channel matrix is uniform from the input, but not doubly uniform

$$\therefore I = H(Y) + (1-q-p) \log(1-q-p) + q \log q + p \log p$$

↖ expect proof of this

Although channel not uniform for output it is easy to show in this case that $H(Y)$ is maximised when $P(x_1) = P(x_2) = 1/2$

↖ expect this to be shown

With these probabilities

$$P(y_1) = P(x_1)(1-p-q) + P(x_2)q = \frac{1-q}{2}$$

$$P(y_2) = P(x_1)q + P(x_2)(1-p-q) = q$$

$$P(y_3) = P(x_1)p + P(x_2)(1-p-q) = \frac{1-q}{2}$$

Hence the capacity is

$$C = (1-q) \left[1 - \log_2(1-q) \right] + (1-p-q) \log_2(1-p-q) + p \log_2 p \quad \text{bits/input}$$

—(1)

↑
15 marks
↓

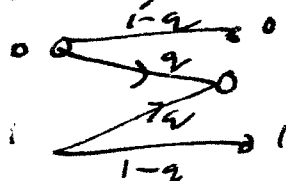
If $q=0$ then C reduces to
 $1 + p \log p + (1-p) \log (1-p)$ which
 is the capacity of the B.S.C.

The capacity in (1) is changed from that of the
 BSC since we are saying, if signals are
 somewhat uncertain we then indicate them
 - i.e. we give outputs that are more certain
 than with BSC. (but, ^{most} uncertain signals are
 indicated)



10
marks

If K becomes large, then $p \rightarrow 0$, i.e. we
 are attempting to rate out errors, at the
 expense of an increasing number of ambiguous
 outputs. The correct outputs are reduced in number,
 but only the most certain are given as
 'positive' outputs. The transition diagram becomes



and the capacity as can
 be seen from (1) becomes
 $C = (1-q)$

Total mark = 25