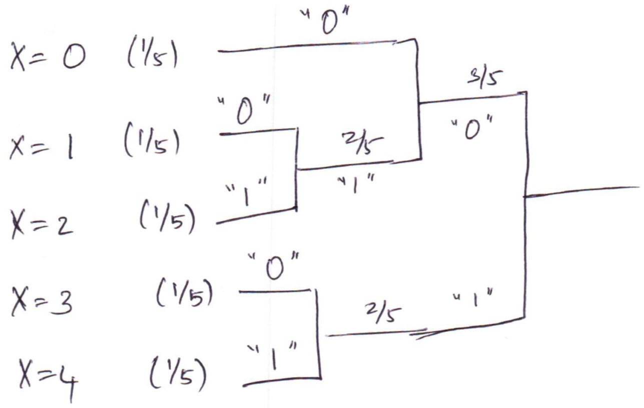
3F7 Information Theory and Coding
Engineering Tripos 2017/18 – Solutions

Question 1

(a) i) Let X be a random variable which takes on values in $\{0, 1, 2, 3, 4\}$, with $X = 0$ if all four coins have the same weight, $X = 1$ if the first coin is heavier, $X = 2$ if the second coin is heavier etc. [30%]

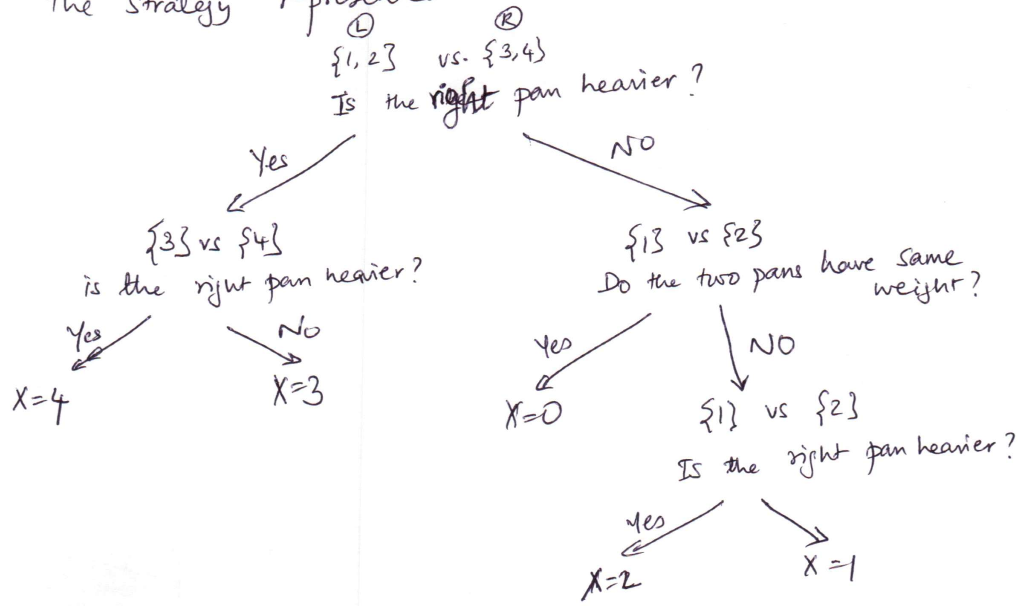
We determine a Huffman code for the distribution $P(X = 0) = P(X = 1) = \ldots = P(X = 4) = \frac{1}{5}$. Oe possible assignment from the Huffman code and the corresponding strategy is given below.



The Huffman code is:

$X=0$	:	0 0
$X=1$	:	0 1 0
$X=2$	:	0 1 1
$X=3$	:	1 0
$X=4$	:	1 1

The strategy represented as a tree is.



Ⓛ Ⓡ
{1, 2} vs. {3, 4}
Is the right pan heavier?
Yes
NO
{3} vs {4}
is the right pan heavier?
Yes
No
X=4
X=3
{1} vs {2}
Do the two pans have same weight?
Yes
NO
X=0
{1} vs {2}
Is the right pan heavier?
Yes
X=2
X=1

ii) The average number of yes/no questions is the expected code length of the Huffman code, which is

$$L = \frac{2+3+3+2+2}{5} = 2.4.$$

(Note that different strategies are possible depending on the assignments of codewords to the five possibilities, but the average number of questions in each strategy is the same. [10%]

(b) i) $H(Z|X) = H(X+Y|X) = H(Y|X)$, where the last equality holds because given X, knowing $(X+Y)$ is equivalent to knowing Y. [10%]

ii) Consider $H(X,Z)$. This can be expanded in two ways as follows [20%]

$$H(X,Z) = H(Z) + H(X|Z) = H(X) + \underbrace{H(Z|X)}_{H(Y|X) \text{ from part (i)}}$$

Furthermore, since X, Y are independent $H(Y|X) = H(Y)$.

Using this in the above equation, we have

$$H(Z) = H(X) + H(Y) - H(X|Z) = H(Y) + I(X;Z).$$

Since $I(X;Z) \geq 0$, we have shown that $H(Z) \geq H(Y)$. Swapping the roles of X and Y above, we similarly obtain $H(Z) \geq H(X)$.

iii) Let X be any random variable taking values in $\{x_1, \ldots, x_r\}$ with non-zero probabilities for each symbol. Let $Y = -X$. Then $Z = X + Y = 0$, hence $H(Z) = 0$. [10%]

But $H(X) = H(Y) > 0$.

iv) Let X, Y be independent. Consider $H(X+Y, X)$, which using the chain rule, can be expanded in two ways:

$$\begin{aligned} H(X+Y,X) &= H(X+Y) + H(X|X+Y) && (1) \\ &= H(X) + H(X+Y|X) \\ &= H(X) + H(Y|X) \\ &= H(X) + H(Y). && (2) \end{aligned}$$

Comparing the RHS of (1) and (2) we see that $H(X+Y) = H(X) + H(Y)$ if X, Y are independent and $H(X|X+Y) = 0$, i.e., the sum $(X+Y)$ uniquely identifies the pair (X,Y). In other words, [20%] the sum $(X+Y)$ takes on a distinct value for each (X,Y) pair.

Assessor's comment: *For part (a), some students interpreted the number of weighings as distinct from the number of yes/no questions. For example, with $\{1,2\}$ compared with $\{3,4\}$, the question "is the right pan heavier" may be asked; if the answer is no, "one may ask the question 'do the two pans have equal weight"? These two questions may be interpreted as corresponding to either one weighing or two weighings; full credit was given to both interpretations, as long as they were consistently applied throughout the question.*

For part (b.iv), many had the right intuition that one of the required conditions is that $(X,Y) \to (X+Y)$ should be an invertible function, but many did not specify the other requirement: (X,Y) also need to be independent.

Question 2

(a) i) The number of channel uses required is $\frac{m}{0.8\mathcal{C}}$. Let $R = 0.8\mathcal{C}$, and $n = \frac{m}{0.8\mathcal{C}}$.

Let P_X be the distribution that maximises $I(X;Y)$. (The proof of the channel coding theorem shows that one could use any input distribution that produces $I(X;Y) > R$.)

Construct a codebook by choosing 2^{nR} codewords each of length n, with each entry of each codeword chosen i.i.d. $\sim P_X$. Label these codewords $\{X^n(1), \dots, X^n(2^{nR})\}$. [25%]

To transmit message $i \in \{1, \dots, 2^{nR}\}$, the encoder transmits $X^n(i)$.

The decoder receives Y^n and uses joint typicality decoding. It decodes $\hat{W}$ as be the message if $X^n(\hat{W})$ is the unique codeword that is Y^n is P_{XY}-jointly-typical with Y. If no such codeword is found an error is declared.

(b) The capacity of the overall channel is $\mathcal{C} = \max_{P_X} I(X;YZ)$.

i) We have

$$\begin{aligned} I(X;YZ) &= I(X;Y) + I(X;Z|Y) \qquad (3)\\ &= I(X;Z) + I(X;Y|Z) \end{aligned}$$

Since $I(X;Z|Y), I(X;Y|Z)$ are both non-negative, we conclude that [15%]

$$I(X;YZ) \geq \max\{I(X;Y),\, I(X;Z)\},$$

and hence

$$\mathcal{C} = \max_{P_X} I(X;YZ) \geq \max\{\max_{P_X} I(X;Y),\, \max_{P_X} I(X;Z)\} = \max\{\mathcal{C}_1, \mathcal{C}_2\}.$$

ii) Starting with the expression in (3), we have the following chain:

$$\begin{aligned} I(X;YZ) &= I(X;Y) + I(X;Z|Y)\\ &= I(X;Y) + I(X;Z) + [I(X;Z|Y) - I(X;Z)]\\ &= I(X;Y) + I(X;Z) + [H(Z|Y) - H(Z|YX) - H(Z) + H(Z|X)]\\ &\overset{*}{=} I(X;Y) + I(X;Z) + [H(Z|Y) - H(Z)]\\ &= I(X;Y) + I(X;Z) - I(Z;Y)\\ &\leq I(X;Y) + I(X;Z). \end{aligned}$$

In the above $(*)$ holds because $H(Z|XY) = H(Z|X)$ since (Y,Z) are conditionally independent given X. Hence [20%]

$$\mathcal{C} = \max_{P_X} I(X;YZ) \;\leq\; \max_{P_X} I(X;Y) \;+\; \max_{P_X} I(X;Z) = \mathcal{C}_1 + \mathcal{C}_2$$

(c) The transition probability matrix of the channel $P_{YZ|X}$ is as follows:

		YZ			
	$P_{YZ\mid X}$	00	01	10	11
X	0	0.72	0.18	0.08	0.02
	1	0.02	0.08	0.18	0.72

For any input distribution $P_X(0) = p$, $P_X(1) = (1-p)$, the induced P_{YZ} distribution is [40%]

	YZ			
	00	01	10	11
P_{YZ}	$0.72p + 0.02(1-p)$	$0.18p + 0.08(1-p)$	$0.18(1-p) + 0.08p$	$0.02(1-p) + 0.72p$

Then the mutual information is

$$\begin{aligned}
I(X;YZ) &= H(YZ) - H(YZ|X) \\
&= H(YZ) - P_X(0) \cdot H(\{0.72, 0.18, 0.08, 0.02\}) - P_X(1) \cdot H(\{0.02, 0.08, 0.18, 0.72\}) \\
&= H(YZ) - 1.1909.
\end{aligned}$$

(Note that $H(YZ|X)$ can also be calculated as $H(Y|X) + H(Z|X) = H_2(0.9) + H_2(0.8)$

Now, since the channel acts symmetrically on inputs $X = 0$ and $X = 1$, the optimal input distribution is uniform, i.e., $p = (1-p) = \frac{1}{2}$. Using this above, the distribution $P_{YZ} = \{0.37, 0.13, 0.13, 0.37\}$, from which which $H(YZ) = 1.8267$.

Therefore, the capacity $\mathcal{C}$ is

$$\mathcal{C} = \max_{P_X} I(X;YZ) = 1.8267 - 1.1909 = 0.6358 \text{ bits.}$$

Assessor's comment: *For part (a), many students assumed that the channel was a binary input channel. The question only says that it is a discrete memoryless channel, so it may have any discrete input alphabet. Most students also did not specify how to choose the input distribution P_X to generate the codebook: this can be the P_X that maximises $I(X;Y)$, or any P_X that results in an $I(X;Y) > R = 0.8\mathcal{C}$.*

Many students struggled with the last part of the question (capacity calculation) and did not calculate the joint distribution of (Y,Z), which is the key step in computing the capacity.

Question 3

(a) i) For each symbol x in the alphabet $\mathcal{X}$, the Shannon-Fano code assigns a codeword of length $\lceil \log \frac{1}{P(x)} \rceil$. Since $\lceil a \rceil < (a+1)$ for any number a, the expected code length satisfies [15%]

$$L = \sum_x P(x) \left\lceil \log \frac{1}{P(x)} \right\rceil < \sum_x P(x) \left(1 + \log \frac{1}{P(x)} \right) = 1 + H(X).$$

ii) Consider a binary random variable X with $P(0) = (1-\delta)$, and $P(1) = \delta$ for some $\delta \in (0,1)$. Then for any given $\epsilon > 0$, by taking δ to be small enough we can ensure that $H(X) = H_2(\delta)$ is less than ϵ. The optimal code for this source is the trivial code $\{0,1\}$, which has average code length 1, which is within ϵ of $1 + H(X) = 1 + H_2(\delta)$, i.e., it is much closer to $1 + H(X)$ than to $H(X) = H_2(\delta)$. [15%]

iii) The average code length of a codeword for k-symbols is

$$\begin{aligned} L_k &= \sum_{(x_1,\dots,x_k)} P(x_1,\dots,x_k) \left\lceil \log \frac{1}{P(x_1,\dots,x_k)} \right\rceil \\ &< \sum_{x_1,\dots,x_k} P(x_1,\dots,x_k) \left(1 + \log \frac{1}{P(x_1,\dots,x_k)} \right) \\ &= 1 + H(X_1,\dots,X_k) = 1 + kH(X), \end{aligned}$$

where the last equality holds because $X_1,\dots,X_k$ are iid $\sim P$. Therefore the average code length per source symbol is bounded as [15%]

$$\frac{L_k}{k} < H(X) + \frac{1}{k} \text{ bits/source symbol.}$$

(b) i) We have

$$p_i := P(X^\Delta = x_i) = P(i\Delta \le X < (i+1)\Delta) = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(x_i)\Delta,$$

where the last equality follows from the equation defining x_i. [10%]

ii) The entropy is [10%]

$$\begin{aligned} H(X^\Delta) = -\sum_i p_i \log p_i &= -\sum_i f(x_i)\Delta \log\left(f(x_i)\Delta\right) \\ &= -\sum_i \Delta f(x_i) \log f(x_i) - \sum_i f(x_i)\Delta \log \Delta. \end{aligned}$$

iii) We have [10%]

$$\sum_i f(x_i)\Delta = \Delta \sum_i \int_{i\Delta}^{(i+1)\Delta} f(x)dx = \int_{-\infty}^{\infty} f(x)dx = 1,$$

where the last equality holds because the pdf integrates to 1. Using this in the second term of the result in part (ii), we get

$$H(X^\Delta) = -\sum_i \Delta f(x_i) \log_2 f(x_i) - \log_2 \Delta.$$

iv) For a random variable X uniform in $[0, \frac{1}{8}]$, the differential entropy is [15%]

$$h(X) = -\int_0^{1/8} 8 \log_2 8 \, dx = -3 \ \text{ bits} \ .$$

With $\Delta = 2^{-n}$, Eq. (2) in the exam paper says that

$$H(X^\Delta) \approx h(X) + n = n - 3 \ \text{ bits.}$$

Therefore, for large n, the entropy of the quantised (discrete) random variable X^Δ is $(n-3)$ bits, with the quantisation levels being $\Delta = 2^{-n}$ apart, i.e., quantised to n-bit accuracy.

v) Any number in the interval $[0, \frac{1}{8}]$ can be expressed in binary as $0.000b_4b_5b_6\ldots$. Since the first three digits of the expansion are always zero, to represent any number in $[0, \frac{1}{8}]$ with n-bit accuracy, i.e., up to bit b_n, we need only $(n-3)$ bits. [10%]

Assessor's comment: *Reasonably well answered by most of those who attempted it. For part (a.ii), many students gave examples where the expected length of the Shannon-Fano code was close to $1 + H(X)$. Note that the question asks for the **optimal** expected code length (that of the Huffman code) to be close to $(1 + H(X))$.*

Question 4

(a) Dimension $k = 3$, block length $n = 6$, rate $k/n = 0.5$. [10%]

(b) If we denote the information sequence by $\underline{x} = [x_1, x_2, x_3]$ then the codeword is generated as $\underline{x}\mathbf{G}$. Hence [15%]

$$[x_1,\, x_2,\, x_3] \cdot \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix} = [?, 1, 1, ?, ?, 0].$$

Using the unerased code bits, we get

$$x_2 + x_3 = 1, \quad x_2 = 1, \quad x_1 + x_2 = 0.$$

Therefore, $x_1 = 1, x_2 = 1, x_3 = 0$, and the transmitted codeword is $[1, 1, 1, 0, 1, 0]$.

(c) We first swap the second and third rows, and then replace the third row with the sum of the second and third rows to obtain a systematic generator matrix: [10%]

$$\begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix} \xrightarrow{\text{swap } R_2 \text{ and } R_3} \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \end{bmatrix} \xrightarrow{R_3 = R_2 + R_3} \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}.$$

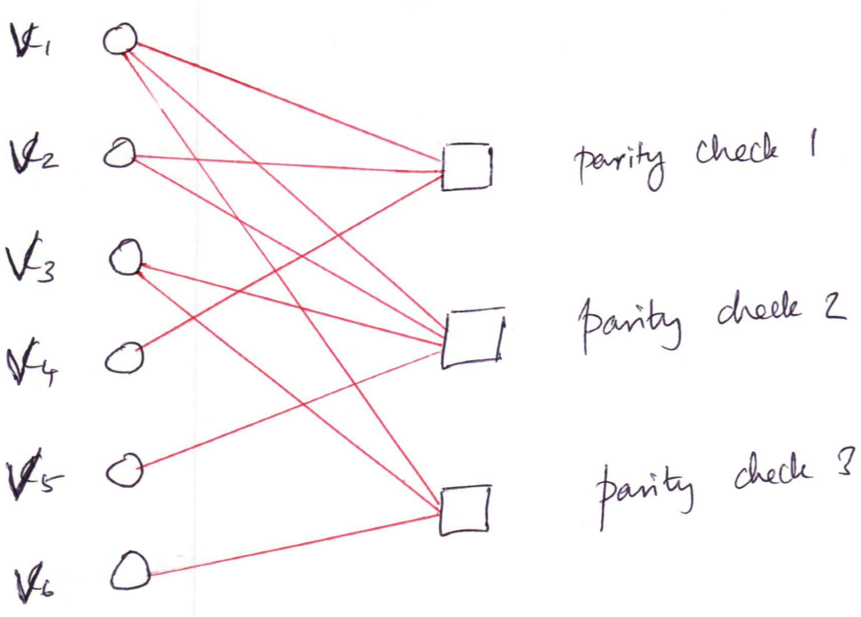
(d) WIth $\mathbf{G} = [I_k \mid P]$, the parity check matrix $\mathbf{H} = [-P^T \mid I_{n-k}]$ is [5%]

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

(e) The minimum number of columns of $\mathbf{H}$ that add up to the zero vector is 3, e.g., columns 2,4,5. So $d_{min} = 3$. [5%]

(f) The decoded codeword will be $[0, 0, 0, 0, 0, 0]$. It is unique because the received sequence has Hamming distance one from this codeword, and since the minimum distance between codeword is three, no other codeword can be at Hamming distance one from the received sequence $[0, 1, 0, 0, 0, 0]$. (This can also be verified by listing all the codewords.) [15%]

(g) The factor graph is as shown below: [10%]



(h) Since the value of the second received bit is 1, the message sent to the check nodes in the first iteration is corresponding LLR, given by

$$L(y_2) = \ln \frac{P(y_2 = 1|c_2 = 0)}{P(y_2 = 1|c_2 = 0)} = \ln \frac{0.1}{0.9} = -2.1972.$$

[10%]

(i) Since the second variable node is connected to check nodes 1 and 3, the final LLR for the second bit at the end of one complete iteration is computed as

$$L_2 = L(y_2) + L_{c_1 \to v_2} + L_{c_2 \to v_3}.$$

We already computed $L(y_2) = -2.1972$. The incoming check-to-variable messages are [20%]

$$L_{c_1 \to v_2} = 2 \tanh^{-1}([\tanh(2.1972/2)]^2) = 1.5163,$$
$$L_{c_2 \to v_2} = 2 \tanh^{-1}([\tanh(2.1972/2)]^3) = 1.1309.$$

Therefore, the final LLR for code bit 2 is

$$L_2 = -2.1972 + 1.5163 + 1.1309 = 0.45.$$

Since $L_2 > 0$, the bit is correctly decoded as 0.

Assessor's comment: *The most popular question on the exam. Well answered overall, though many had trouble computing the final LLR for the second bit in the last part of the question.*