3F7 Information Theory and Coding
Engineering Tripos 2025/26 – Solutions

Question 1

(a) Bag 1: $[R\ R\ B]$ Bag 2: $[B\ B\ R]$

(i) Since the ball is drawn is uniformly random, $P(Y = \text{Red}) = P(Y = \text{Blue}) = \frac{1}{2}$, hence $H(Y) = 1$. Similarly, $H(X) = 1$. The joint entropy is $H(Y, X) = H(X) + H(Y \mid X)$, and the conditional entropy is

$$H(Y \mid X) = \frac{1}{2}H(Y \mid X = 1) + \frac{1}{2}H(Y \mid X = 2) = \frac{1}{2}H_2(2/3) + \frac{1}{2}H_2(2/3) = H_2(2/3) = 0.918,$$

where $H_2(p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}$. Therefore $H(X, Y) = 1.918$. [20%]

(ii) We have

$$H(Z \mid Y) = P(Y = R)H(Z \mid Y = R) + P(Y = B)H(Z \mid Y = B) = \frac{1}{2}\left(H(Z \mid Y = R) + H(Z \mid Y = B)\right).$$

The conditional probability mass functions $P(Z \mid Y = R)$ and $P(Z \mid Y = B)$ can be computed using Bayes rule and the law of total probability [25%]

$$\begin{aligned} P(Z = R \mid Y = R) = \frac{P(Z = R, Y = R)}{P(Y = R)} &= \frac{P(Z = R, Y = R, X = 1) + P(Z = R, Y = R, X = 2)}{P(Y = R)} \\ &= \frac{1/2 \cdot 1/3 \ + \ 1/2 \cdot 0}{1/2} = \frac{1}{3}. \end{aligned}$$

Therefore $P(Z = B \mid Y = R) = \frac{2}{3}$. Similarly $P(Z = B \mid Y = B) = 1 - P(Z = R \mid Y = B) = \frac{1}{3}$. Hence $H(Z \mid Y = R) = H(Z \mid Y = B) = H_2(1/3)$ and $H(Z \mid Y) = H_2(1/3) = 0.918$.

(iii) We have $I(X; Z \mid Y) = H(Z \mid Y) - H(Z \mid X, Y) = H_2(1/3) - H(Z \mid X, Y)$, and

$$\begin{aligned} H(Z \mid X, Y) = {} & P(X = 1, Y = R)H(Z \mid X = 1, Y = R) + P(X = 1, Y = B)H(Z \mid X = 1, Y = B) \\ & + P(X = 2, Y = R)H(Z \mid X = 2, Y = R) + P(X = 2, Y = B)H(Z \mid X = 2, Y = B) \\ = {} & \frac{1}{2}\frac{2}{3} \cdot 1 + \frac{1}{2}\frac{1}{3} \cdot 0 + \frac{1}{2}\frac{1}{3} \cdot 0 + \frac{1}{2}\frac{2}{3} \cdot 1 = \frac{2}{3} \end{aligned}$$

Using this $I(X; Z \mid Y) = H_2(1/3) - 2/3 = 0.2516$. By the chain rule, $I(X; Y, Z) = I(X; Y) + I(X; Z \mid Y)$. From part (i), [20%]
$I(X; Y) = H(Y) - H(Y \mid X) = 1 - H_2(2/3)$. Hence $I(X; Y, Z) = \frac{1}{3}$.

(b) (i) Letting $\mathcal{Y}$ denote the range of the function $f(X)$, for each $y \in \mathcal{Y}$, we have $P_Y(y) = \sum_{x:f(x)=y} P(x)$ and $Q_Y(y) = \sum_{x:f(x)=y} Q(x)$. Then, [25%]

$$D(P_Y \| Q_Y) = \sum_{y \in \mathcal{Y}} P_y(y) \log \frac{P_y(y)}{Q_y(y)} = \sum_{y \in \mathcal{Y}} \sum_{x:f(x)=y} P(x) \log \frac{\sum_{x:f(x)=y} P(x)}{\sum_{x:f(x)=y} Q(x)}. \tag{1}$$

By the log-sum inequality, for each y, we have

$$\sum_{x:f(x)=y} P(x) \log \frac{\sum_{x:f(x)=y} P(x)}{\sum_{x:f(x)=y} Q(x)} \leq \sum_{x:f(x)=y} P(x) \log \frac{P(x)}{Q(x)}.$$

Substituting this in (1), we obtain,

$$\begin{aligned} D(P_Y \| Q_Y) &= \sum_{y \in \mathcal{Y}} \sum_{x: f(x)=y} P(x) \log \frac{\sum_{x: f(x)=y} P(x)}{\sum_{x: f(x)=y} Q(x)} \\ &\leq \sum_{y \in \mathcal{Y}} \sum_{x: f(x)=y} P(x) \log \frac{P(x)}{Q(x)} \\ &= \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} = D(P \| Q). \end{aligned}$$

(ii) The inequality $D(P_Y \| Q_Y) \leq D(P \| Q)$ is a *data processing* inequality. The relative entropy between two distributions can only decrease when you apply a function to the underlying random variables. An extreme case of this is the function $f(x) = c$, where c is a constant. In this case, $D(P_Y \| Q_Y) = 0$. Equality holds if the function f is one-to-one (invertible). [10%]

Assessor's comment: Most did part (a) correctly, but many struggled with part (b) due to incorrect heuristic applications of Bayes rule. Not many gave a correct proof for (c)(i), with some using the Jacobian (which doesn't apply to discrete random variables).

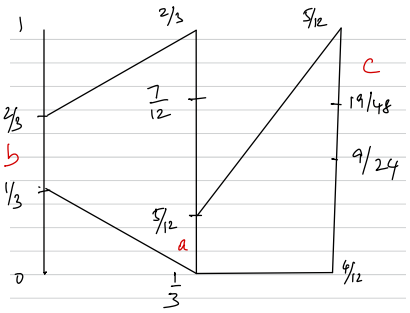
Question 2

(a) {1, 01, 001, 000}: can be Huffman code. [15%]

{00, 01, 10, 110}: cannot be a Huffman code since in any optimal code the two longest codewords should be neighbouring leaves on the tree.

{0, 10, 01, 11}: cannot be a Huffman code as it is not prefix-free.

(b) FALSE. Since $I(X;Y) = H(Y) - H(Y \mid X) = H(X) - H(X \mid Y)$, we have $I(X;Y) \le H(X) \le \log_2 4$ and also, $I(X;Y) \le H(Y) \le \log_2 3$. Therefore the capacity $\max_{P_X} I(X;Y)$ of the channel can be no larger than $\log_2 3$ bits per channel use. [15%]



(c) (i) The interval for (b, a, c) is $[\frac{19}{48}, \frac{20}{48})$, as shown above. The number of bits needed for the codeword is therefore either $\lceil \log_2 48 \rceil = 6$ or 7. The interval can be expressed as $[\frac{25.33}{64}, \frac{26.67}{64})$, or $[\frac{50.667}{128}, \frac{53.333}{128})$. The dyadic interval $[\frac{51}{128}, \frac{52}{128})$ sits within the symbol interval, so the codeword is the 7-bit binary expansion of 51, i.e, 0110011. (Can also use the binary expansion of 52; using a different assignment of intervals to a, b, c will also give a different answer.) [25%]

(ii) The best possible lower bound on the expected code length for (X_1, X_2, X_3) is [15%]

$$\begin{aligned} H(X_1, X_2, X_3) &= H(X_1) + H(X_2 \mid X_1) + \underbrace{H(X_3 \mid X_1, X_2)}_{=H(X_3|X_2)} \\ &= \log_2 3 + H(\{0.5, 0.25, 0.25\}) + H(\{0.5, 0.25, 0.25\}) = 4.585 \text{ bits.} \end{aligned}$$

(d) Consider a compression code designed for distribution Q with code lengths $\log_2 \frac{1}{Q(a)}$, $\log_2 \frac{1}{Q(b)}$, and $\log_2 \frac{1}{Q(c)}$. (These may not be integers, but we only want to use it to find the coding distribution that minimizes redundancy.) The expected code length when the source distribution is P is $L = \sum_{x \in \mathcal{X}} P(x) \log_2 \frac{1}{Q(x)} = H(P) + \underbrace{D(P\|Q)}_{\text{redundancy}}$. [30%]

Since we know $P(a) = 0.5$, to minimize redundancy, we take $Q(a) = 0.5$ and let $Q(b) = q$, $Q(c) = (0.5 - q)$. The redundancy is

$$D(P\|Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} = p \log_2 \frac{p}{q} + (0.5 - p) \log_2 \frac{0.5 - p}{0.5 - q}.$$

Since $p = 0.1$ or $p = 0.2$, the redundancy takes one of two values:

$$D_1(q) = 0.1 \log_2 \frac{0.1}{q} + 0.4 \log_2 \frac{0.4}{0.5 - q}, \quad D_2(q) = 0.2 \log_2 \frac{0.2}{q} + 0.3 \log_2 \frac{0.3}{0.5 - q}$$

Choosing $q = 0.1$ minimizes D_1, and $q = 0.2$ minimizes D_2. The optimal value of q that minimizes the worst-case redundancy $D(q) = \max\{D_1(q), D_2(q)\}$ lies between 0.1 and 0.2. Since

$D_1(q)$ increases and $D_2(q)$ decreases for $q \in [0.1, 0.2]$ the value that minimizes $D(q)$ is the one that solves $D_1(q) = D_2(q)$. This gives:

$$0.1 \log_2 \frac{0.5 - q}{q} = 0.1 \log_2 0.1 + 0.4 \log_2 0.4 - 0.2 \log_2 0.2 - 0.3 \log_2 0.3,$$

solving which we obtain $q = 0.1484$. Therefore the optimal coding distribution is $Q(a) = 0.5, Q(b) = 0.1484, Q(c) = 0.3516$.

Assessor's comment: In part (c)(i), most got the correct interval, but many did not use the correct number of bits for the binary expansion. For c(ii), many answers wrongly added 2 to the joint entropy; the extra 2 bits is required only to get an *upper* bound on the expected number of bits. Many good starts for part (d), but then went off-track with approaches such as differentiation, which don't work here.

Question 3

(a) (i) 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 $\{X^n(1), \ldots, X^n(2^{nR})\}$. [20%]
To transmit message $W \in \{1, \ldots, 2^{nR}\}$, the encoder transmits $X^n(W)$.
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.

(ii) The entropy of the source is $0.5 \log_2 \frac{1}{0.5} + 0.25 \log_2 \frac{1}{0.25} + 0.25 \log_2 \frac{1}{0.25} = 1.5$. An optimal compressor will represent m source symbols using (close to) $1.5m$ bits. If the capacity is $\mathcal{C}$ bits/channel use, then the minimum number of channel uses required is $1.5m/\mathcal{C}$. [10%]

(b) For each input symbol $x \in \{1, \ldots, M\}$, we have $P(Y = x \mid X = x) = 1 - \varepsilon$ and $P(Y = ? \mid X = x) = \varepsilon$. Therefore

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_{x \in \{0,1\}} P_X(x) H(Y|X = x) = H(Y) - H_2(\varepsilon), \qquad (2)$$

since $H(Y|X = x) = H_2(\varepsilon)$ for $x \in \{1, \ldots, M\}$. Now, let the input distribution be $P_X(1) = \alpha_1, \ldots, P_X(M) = \alpha_M$, with $(\alpha_1 + \ldots + \alpha_M) = 1$. Then the output distribution is

$$P_Y(1) = \alpha_1(1-\varepsilon),\ P_Y(2) = \alpha_2(1-\varepsilon), \ldots P_Y(M) = \alpha_M(1-\varepsilon),\ P_Y(?) = \varepsilon.$$

Hence

$$\begin{aligned} H(Y) &= \varepsilon \log \frac{1}{\varepsilon} + \sum_{i=1}^{M} \alpha_i (1-\varepsilon) \log \frac{1}{\alpha_i (1-\varepsilon)} \\ &= \varepsilon \log \frac{1}{\varepsilon} + \underbrace{\left(\sum_{i=1}^{M} \alpha_i \right)}_{1} (1-\varepsilon) \log \frac{1}{(1-\varepsilon)} + (1-\varepsilon) \sum_{i=1}^{M} \alpha_i \log \frac{1}{\alpha_i} \\ &= H_2(\varepsilon) + (1-\varepsilon) H(\{\alpha_1, \ldots, \alpha_M\}). \end{aligned}$$

Substituting this in (2), we obtain

$$I(X;Y) = (1-\varepsilon) H(\{\alpha_1, \ldots, \alpha_M\}).$$

Since the uniform distribution maximises the entropy, the maximising input distribution is $P_X(1) = \ldots = P_X(M) = 1/M$, which gives $\mathcal{C} = \max_{P_X} I(X;Y)(1-\varepsilon) \log_2 M$. [25%]

(c) (i) The capacity is $\max I(X;Y)$ where the maximum is over all input distributions on $[0, 2\pi)$. Since $Y = X + Z$ and Z is independent of X, we have

$$I(X;Y) = h(Y) - h(Y \mid X) = h(X) - h(X \mid Y).$$

Since $y = x + z \mod 2\pi$, we have $h(X \mid Y) = h(Z)$. Indeed, for any input-output pair $y \in [0, 2\pi)$, we have $x = (y - z)$ if $z \le y$, and $x = y - z + 2\pi$ if $z > y$. (Verify this for $z > y$ by plotting on a line segment from 0 to 2π and calculating x.) [25%]
The given hint implies that $h(X)$ and hence $I(X;Y)$ is maximised by the uniform input distribution.

(ii) Using part (i), taking the maximising input distribution $p_X(x) = \frac{1}{2\pi}$ for $x \in [0, 2\pi)$, we have $h(Y) = \log(2\pi)$ and hence $\mathcal{C} = \log(2\pi) - h(Z)$. For the given distribution, [20%]

$$h(Z) = -\int_0^{2\pi} p(z) \log p(z) = \frac{(\log e)\alpha}{1 - e^{-2\pi\alpha}} \int_0^{2\pi} (\alpha z) e^{-\alpha z} dz - \log \frac{\alpha}{1 - e^{-2\pi\alpha}}.$$

Integrating by parts, we get

$$\int_0^{2\pi}(\alpha z)e^{-\alpha z}dz = \int_0^{2\pi}(-z)d\left(e^{-\alpha z}\right)dz = 2\pi e^{-2\pi\alpha} + \int_0^{2\pi} e^{-\alpha z}dz = 2\pi e^{-2\pi\alpha} + \frac{1-e^{-2\pi\alpha}}{\alpha}.$$

Therefore,

$$h(Z) = \frac{(\log e)2\pi\alpha e^{-2\pi\alpha}}{(1-e^{-2\pi\alpha})} + \log\frac{e(1-e^{-2\pi\alpha})}{\alpha} \;\Rightarrow \mathcal{C} \;= \log\left(\frac{2\pi\alpha}{e(1-e^{-2\pi\alpha})}\right) - \frac{(\log e)2\pi\alpha e^{-2\pi\alpha}}{(1-e^{-2\pi\alpha})}.$$

Assessor's comment: For part (a), many answers described typical set based source coding, rather than channel coding as required by the question; some implicitly assumed that the channel was binary input. Many good attempts for part (c), but many spent a lot of time on the calculation in c(ii); it is a little messy but recognizing parts of the integrand as a probability density avoids a few integrals.

Question 4

(a) The rate of the code is $3/6 = 0.5$. [5%]

(b) Recall $[c_1, c_2, c_3, c_4, c_6] = [x_1, x_2, x_3]\mathbf{G}$. The first column of $\mathbf{G}$ corresponds to the equation defining c_1, therefore it is $[0, 1, 0]^T$. The second column corresponds to the equation for c_2, and is $[1, 0, 1]^T$ and so on. Therefore [15%]

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

(c) The eight codewords are obtained by substituting each possible value for $[x_1, x_2, x_3]$: [10%]

$$[0,0,0,0,0,0],\ [0,1,1,0,1,0],\ [1,0,1,1,1,0],\ [0,1,0,1,1,1],$$
$$[1,1,0,1,0,0],\ [1,1,1,0,0,1],\ [0,0,1,1,0,1],\ [1,0,0,0,1,1].$$

(d) We verify that $\mathbf{G}\mathbf{H}^T = \mathbf{0}$, the all-zeros 3×3 matrix: [15%]

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

From $\mathbf{H}$ we can find $\mathbf{G}_{\text{sys}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}$.

(e) The message transmitted by each variable node j in the first iteration is the LLR $L(y_j) = \ln \frac{P(y_j|c_j=0)}{y_j|c_j=1}$, for $j = 1, \ldots, 6$. For the given $\underline{y}$, these are $[\ln 9, -\ln 9, -\ln 9, \ln 9, \ln 9, -\ln 9]$. [25%]

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

$$L_3 = L(y_3) + L_{c_1 \to v_3} + L_{c_3 \to v_3}.$$

We already computed $L(y_3) = -\ln 9 = -2.1972$. The check-to-variable message $L_{c_1 \to v_3}$ is based on the messages received by c_1 in the first iteration from v_2 and v_4:

$$\begin{aligned} L_{c_1 \to v_3} &= 2 \tanh^{-1}[\tanh(L(y_2)/2) \tanh(L(y_4)/2)] \\ &= 2 \tanh^{-1}[\tanh(-2.1972/2) \tanh(L(2.1972/2))] = -1.5163. \end{aligned}$$

Similarly, $L_{c_3 \to v_3}$ is based on the messages received by c_3 in the first iteration from v_1, v_2 and v_6:

$$L_{c_3 \to v_3} = 2 \tanh^{-1}[\tanh(L(y_1)/2) \tanh(L(y_2)/2) \tanh(L(y_6)/2)] = 1.1309.$$

Therefore the final LLR for the third bit is $L_3 = -2.1972 - 1.5163 + 1.1309 = -2.5827$. Since it is negative it will be decoded as a 1.

(f) The posterior probability of the third code bit being 0 is:

$$P(c_3 = 0 \mid \underline{y}) = \sum_{\underline{c}: c_3 = 0} P(\underline{c} \mid \underline{y}) = \sum_{\underline{c}: c_3 = 0} \frac{P(\underline{c}) P(\underline{y} \mid \underline{c})}{P(\underline{y})} = \frac{1/8}{P(\underline{y})} \sum_{\underline{c}: c_3 = 0} P(\underline{y} \mid \underline{c}),$$

where the last equality holds because the codewords are all equally likely. Similarly,

$$P(c_3 = 1 \mid \underline{y}) = \frac{1/8}{P(\underline{y})} \sum_{\underline{c}: c_3 = 1} P(\underline{y} \mid \underline{c}).$$

Denoting the list of codewords found in part (c) by $\{\underline{c}_1, \ldots, \underline{c}_8\}$, and using he BSC(0.1) to compute $P(\underline{y} \mid \underline{c})$ we have: [30%]

$$\begin{aligned}\frac{P(c_3 = 0 \mid \underline{y})}{P(c_3 = 1 \mid \underline{y})} &= \frac{P(\underline{y} \mid \underline{c}_1) + P(\underline{y} \mid \underline{c}_4) + P(\underline{y} \mid \underline{c}_5) + P(\underline{y} \mid \underline{c}_8)}{P(\underline{y} \mid \underline{c}_2) + P(\underline{y} \mid \underline{c}_2) + P(\underline{y} \mid \underline{c}_6) + P(\underline{y} \mid \underline{c}_7)} \\ &= \frac{(0.1)^3(0.9)^3 + (0.1)^3(0.9)^3 + (0.1)^4(0.9)^2 + (0.1)^4(0.9)^2}{(0.1)^2(0.9)^4 + (0.1)^5(0.9) + (0.1)(0.9)^5 + (0.1)^2(0.9)^4} = 0.0224,\end{aligned}$$

which taking ln gives -3.7967 in LLR format. This has the same sign but larger magnitude than the value computed in part (e). This is because in just one iteration, belief propagation is not expected to give an accurate approximation of the posterior probability.

Assessor's comment: The most popular question in the exam, well done by most who attempted it.