3F7 Information Theory and Coding Engineering Tripos 2018/19 – Solutions

Question 1

(a) i) The set of possible values for Y is $\{0,1,2\}$. From the law of total probability, we have for $k=0,1,2$: [20%]

$$P(Y=k)=\sum_{i=1}^{6} P(X=i)\,P(Y=k \mid X=i)=\sum_{i=1}^{6}\frac{1}{6}P(Y=k \mid X=i), \qquad (1)$$

where the second equality holds because all six outcomes of the die are equally likely. For $X \in \{1,2,3,4\}$

$$P(Y=0 \mid X=i)=\frac{1}{2},\ P(Y=1 \mid X=i)=\frac{1}{2},\ P(Y=2 \mid X=i)=0,\ \text{for } i=1,2,3,4.$$

For $X \in \{5,6\}$,

$$P(Y=0 \mid X=i)=\frac{1}{4},\ P(Y=1 \mid X=i)=\frac{1}{2},\ P(Y=2 \mid X=i)=\frac{1}{4},\ \text{for } i=5,6.$$

Substituting these conditional probabilities in (1), we obtain

$$P(Y=0)=\frac{5}{12}, \qquad P(Y=1)=\frac{1}{2}, \qquad P(Y=2)=\frac{1}{12}.$$

ii) The mutual information is [15%]

$$\begin{aligned} I(X;Y)=H(Y)-H(Y \mid X) &= H\left(\left\{\frac{5}{12},\frac{1}{2},\frac{1}{12}\right\}\right)-\sum_{i=1}^{6}\frac{1}{6}H(Y \mid X=i) \\ &= H\left(\left\{\frac{5}{12},\frac{1}{2},\frac{1}{12}\right\}\right)-\frac{4}{6}\cdot 1-\frac{2}{6}H\left(\left\{\frac{1}{4},\frac{1}{2},\frac{1}{4}\right\}\right)=0.158 \text{ bits} \end{aligned}$$

(b) Let $g(x)$ be the uniform density on $[a,b]$. That is $g(x)=\frac{1}{(b-a)}$ for $a \le x \le b$, and zero otherwise. For any density f that is zero outside $[a,b]$, since the relative entropy $D(f\|g)$ is non-negative, we have

$$\begin{aligned} 0 \le D(f\|g) &= \int_a^b f(x)\log_2\frac{f(x)}{1/(b-a)}dx=\int_a^b f(x)\log_2(b-a)dx+\int_a^b f(x)\log_2 f(x)dx \qquad (2) \\ &= \log_2(b-a)-h(X). \end{aligned}$$

In (2), equality holds if and only if $f=g$, i.e., the uniform density. Therefore $h(X)\le \log_2(b-a)$, with equality if and only if X is uniformly distributed in $[a,b]$. [20%]

(c) i) Let n_{1a} denote the number of occurrences of $(X=1, Y=a)$ in (x^n,y^n), and n_{0b}, n_{1b}, n_{0c} similarly defined. From the given pmf, $H(X,Y)=2$ and [15%]

$$P_{XY}(x^n,y^n)=\left(\frac{1}{4}\right)^{n_{1a}+n_{0b}+n_{1b}+n_{0c}} \qquad \Rightarrow \qquad -\frac{1}{n}\log P_{XY}(x^n,y^n)=2\cdot\frac{n_{1a}+n_{0b}+n_{1b}+n_{0c}}{n}.$$

Therefore,

$$A_n^{XY} = \left\{ (x^n, y^n) : \quad \frac{n_{1a} + n_{0b} + n_{1b} + n_{0c}}{n} = 1 \right\}.$$

ii) The marginal pmfs are $P_X(0) = P_X(1) = \frac{1}{2}$, and $P_Y(a) = \frac{1}{4}$, $P_Y(b) = \frac{1}{2}$, $P_Y(c) = \frac{1}{4}$. [15%]
Therefore

$$H(X) = 1, \qquad H(Y) = H\left(\left\{ \frac{1}{4}, \frac{1}{2}, \frac{1}{4} \right\} \right) = \frac{3}{2}.$$

From the marginals, we have

$$P_X(x^n) = \left(\frac{1}{2}\right)^{n_1 + n_0}, \qquad P_Y(y^n) = \left(\frac{1}{2}\right)^{n_a} \left(\frac{1}{4}\right)^{n_b} \left(\frac{1}{2}\right)^{n_c}. \tag{3}$$

Thus

$$A_n^X = \left\{ x^n : \ \frac{n_0 + n_1}{n} = 1 \right\}, \qquad A_n^Y = \left\{ y^n : \frac{2n_a + n_b + 2n_c}{n} = \frac{3}{2} \right\}.$$

Since we also have $(n_a + n_b + n_c) = n$, the above implies any $y^n \in A_n^Y$ will satisfy $\frac{n_b}{n} = \frac{1}{2}$. (This last implication was not required to get full marks for this part.)

iii) Consider the sequence pair:

$$\begin{aligned} x^n &= \ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1 \\ y^n &= \ b\ b\ c\ c\ a\ a\ a\ a \end{aligned}$$

Here $n = 8$, and $n_{1a} + n_{0b} + n_{1b} + n_{0c} = n = 8$. Therefore $(x^n, y^n) \in A_n^{XY}$. However, $\frac{n_b}{n} = \frac{2}{8} \neq \frac{1}{2}$. Therefore $(x^n, y^n) \notin A_n^Y$. [15%]

(Any other example with $\frac{n_b}{n} \neq \frac{1}{2}$ or that violates $\frac{2n_a + n_b + 2n_c}{n} = \frac{3}{2}$ would also work.)

Question 2

(a) The code lengths assigned to the five source symbols by a Shannon-Fano code are $\ell_i = \lceil \log \frac{1}{p_i} \rceil$, $i = 1, \dots, 5$. In decreasing order of symbol probabilities, these are:

$$\ell_1 = 1, \ \ell_2 = 2, \ \ell_3 = 3, \ \ell_4 = 4, \ \ell_5 = 4. \tag{4}$$

The expected code length of the Shannon-Fano code is therefore [20%]

$$L = \frac{1}{2}1 + \frac{1}{4}2 + \frac{1}{8}3 + \frac{1}{16}4 + \frac{1}{16}4 = \frac{15}{8}, \tag{5}$$

which is exactly equal to the entropy of the source.

Since the Huffman code always achieves the optimal code length for a given set of probabilities, its expected code length can be no larger than than that of Shannon-Fano. Hence a Huffman code must achieve the entropy of this source. (Note that Shannon-Fano achieves the entropy whenever all the probabilities are integer powers of $\frac{1}{2}$.)

(b) i) The code lengths and expected code length are given by (4) and (5), respectively. The redundancy $R(\gamma)$ is given by [10%]

$$\begin{aligned} R(\gamma) &= \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{16} \cdot 4 + \frac{1}{16} \cdot 4 - H(S') \\ &= \frac{1}{2} + (16^{-1} - \gamma)\log(16^{-1} - \gamma) + (16^{-1} + \gamma)\log(16^{-1} + \gamma) \end{aligned} \tag{6}$$

ii) $R(\gamma)$ is the relative entropy between the two source distributions in (parts (a) and (b)). As the relative entropy is a measure of distance between the two distributions, we expect the redundancy to be maximum for $\gamma \in (0, \frac{1}{32}]$ when $\gamma = 1/32$. The two distributions are "furthest apart" from one another for this value of γ. [10%]

(c) i) For the Kraft inequality to be satisfied, we require [15%]

$$\sum_{i=1}^{5} 2^{-\ell_i} \leq 1 \quad \Rightarrow \quad \frac{1}{2^2} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^{\ell_4}} + \frac{1}{2^4} \leq 1 \quad \Rightarrow \quad \frac{1}{2^{\ell_4}} \leq \frac{5}{16} \quad \Rightarrow \quad \ell_4 \geq 1.678$$

Since the code length ℓ_4 has to be an integer, the minimum possible value is $\ell_4^* = 2$. The corresponding expected code length is

$$L = \frac{1}{4}2 + \frac{1}{4}2 + \frac{5}{24}3 + \frac{1}{6}2 + \frac{1}{8}4 = 2.458 \text{ bits.}$$

ii) The set $\mathcal{B}$ does not contain an optimal symbol code for X because an optimal code in $\mathcal{B}$ (with the expected code length computed in (i) above) assigns a longer codeword to a_3 than a_4, though a_3 has the higher symbol probability. The expe in $\mathcal{B}$ can be improved by swapping the codewords for a_3 and a_4. [10%]

iii) Arithmetic coding would be the better choice. Since the probabilities are not dyadic (integer powers of 1/2), to achieve rates close to the entropy using Huffman coding, one needs to code over long blocks of symbols. Constructing a Huffman code over blocks (super-symbols) of length k will give an expected code length within $1/k$ of the entropy, but the complexity increases exponentially with k. The complexity of arithmetic coding over a string of k source symbols increases only linearly with k and the expected code length is within $\frac{2}{k}$. [10%]

(d) We have

$$\begin{aligned} H(X) - H(X') &= p_i \log \frac{1}{p_i} + p_j \log \frac{1}{p_j} - (p_i - \epsilon) \log \frac{1}{p_i - \epsilon} - (p_j + \epsilon) \log \frac{1}{p_j + \epsilon} \\ &= p_i \log \frac{(p_i - \epsilon)}{p_i} \ + \ p_j \log \frac{(p_j + \epsilon)}{p_j} \ + \ \epsilon \log \frac{p_j + \epsilon}{p_i - \epsilon}. \qquad (7) \end{aligned}$$

We now use the inequality $\ln x \leq (x-1)$, or equivalently, $\log x \leq (x-1)/\ln 2$. Using this in (7), we obtain [25%]

$$\begin{aligned} H(X) - H(X') &\leq \left[p_i \left(-\epsilon/p_i\right) + p_j(\epsilon/p_j) + \epsilon \left(\frac{p_j + \epsilon}{p_i - \epsilon} - 1 \right) \right] \frac{1}{\ln 2} \\ &= \frac{\epsilon}{\ln 2} \left(\frac{p_j + \epsilon}{p_i - \epsilon} - 1 \right) < 0 \quad \text{since we are given } (p_j + \epsilon) < (p_i - \epsilon). \end{aligned}$$

Hence $H(X') > H(X)$, as required.

(Note: One could also prove the result in other ways. E.g., by differentiating (7) with respect to ϵ, and showing that the derivative is negative. One could also recognise that $p_i \log \frac{(p_i - \epsilon)}{p_i} + p_j \log \frac{(p_j + \epsilon)}{p_j} = -D(P_X \| P_{X'}) \leq 0$, where $P_X, P_{X'}$ denote the two pmfs.)

Question 3

(a) i) Consider the mutual information $I(X;Y\,g(Y))$, and expand in two different ways using the chain rule. [15%]

$$I(X;Y\,g(Y)) = I(X;Y) + I(X;g(Y)|Y) = I(X;g(Y)) + I(X;Y|g(Y)) \tag{8}$$

The term $I(X;g(Y)|Y) = H(g(Y)|Y) - H(g(Y)|X,Y) = 0$, because given Y, there is no uncertainty in $g(Y)$. Using this in (8) gives

$$I(X;Y) = I(X;g(Y)) + I(X;Y|g(Y)).$$

Since $I(X;Y|g(Y)) \geq 0$, this implies $I(X;Y) \geq I(X;g(Y))$.

ii) We have [15%]

$$\begin{aligned}
0 \leq D(P\|Q) &= \sum_x P(x)\log_2 \frac{P(x)}{Q(x)} \\
&= \frac{1}{\ln 2}\sum_x P(x)\ln\frac{P(x)}{Q(x)} \\
&\overset{(a)}{\leq} \frac{1}{\ln 2}\sum_x P(x)\left[\frac{P(x)}{Q(x)} - 1\right] \\
&= \frac{1}{\ln 2}\left\{\left[\sum_x \frac{P(x)^2}{Q(x)}\right] - 1\right\} \qquad \Rightarrow \qquad \sum_x \frac{P(x)^2}{Q(x)} \geq 1, \quad \text{as required.}
\end{aligned}$$

In the above (a) is obtained using $\ln x \leq (x-1)$.

(b) i) We first show that $I(X;Z|Y) = 0$, and hence [10%]

$$I(X;Y,Z) = I(X;Y) + I(X;Z|Y) = I(X;Y). \tag{9}$$

We have

$$I(X;Z|Y) = H(Z|Y) - H(Z|X,Y) = 0,$$

since Z is uniquely determined from Y, and hence $H(Z|Y) = H(Z|X,Y) = 0$.

ii) We compute $I(X;Y,Z)$ as follows:

$$I(X;Y,Z) = H(Y,Z) - H(Y,Z|X) = H(Z) + H(Y|Z) - H(Z|X) - H(Y|X,Z). \tag{10}$$

We compute each of the terms as follows: [15%]

$$\begin{aligned}
H(Z) &= H_2(\alpha) && (11)\\
H(Y|Z) &= \alpha H(Y|Z=1) + (1-\alpha)H(Y|Z=2) = \alpha H(Y_1) + (1-\alpha)H(Y_2) && (12)\\
H(Z|X) &= 0 \quad \text{since } Z \text{ is uniquely determined by } X && (13)\\
H(Y|X,Z) &= \alpha H(Y|X,Z=1) + (1-\alpha)H(Y|X,Z=2) = \alpha H(Y_1|X_1) + (1-\alpha)H(Y_2|X_2). && (14)
\end{aligned}$$

Substituting (11) – (14) in (9), we obtain

$$\begin{aligned}
I(X;Y) = I(X;YZ) &= H_2(\alpha) + \alpha H(Y_1) + (1-\alpha)H(Y_2) - \alpha H(Y_1|X_1) - (1-\alpha)H(Y_2|X_2) \\
&= \alpha I(X_1;Y_1) \;+\; (1-\alpha)I(X_2;Y_2) \;+\; H_2(\alpha). \qquad (15)
\end{aligned}$$

Note: One could also compute $I(X;Y,Z)$ via $I(X;Y \mid Z) + I(X;Z)$. Then $I(X;Z) = H(Z) = H_z(\alpha)$, and $I(X;Y \mid Z) = \alpha I(X_1;Y_1) + (1-\alpha)I(X_2;Y_2)$.

(c) Note that $I(X_1;Y_1) \leq \mathcal{C}_1$ and $I(X_2;Y_2) \leq \mathcal{C}_2$, with equality if we choose the capacity achieving input distribution for each of the individual channels. When choosing the channel 1 with prob. α and channel 2 with probability $(1-\alpha)$, we can maximise mutual information for the chosen channel by using its capacity achieving input distribution. We then optimize over α to find the capacity of the union channel:

$$\mathcal{C} = \max_{P_X} I(X;Y) = \max_{\alpha\in[0,1]} \; \alpha\mathcal{C}_1 \,+\, (1-\alpha)\mathcal{C}_2 \,+\, H_2(\alpha), \tag{16}$$

We differentiate $f(\alpha) = \alpha\mathcal{C}_1 \,+\, (1-\alpha)\mathcal{C}_2 \,+\, H_2(\alpha)$ to find the maximum: [25%]

$$\frac{df(\alpha)}{d\alpha} = \mathcal{C}_1 - \mathcal{C}_2 + \log_2\frac{(1-\alpha)}{\alpha}, \tag{17}$$

where we have used the given hint $\frac{dH_2(\alpha)}{d\alpha} = \log_2\frac{(1-\alpha)}{\alpha}$. Equating (17) to 0, we find that the maximising value of α is

$$\alpha^* = \frac{2^{\mathcal{C}_1}}{2^{\mathcal{C}_1}+2^{\mathcal{C}_2}}, \qquad 1-\alpha^* = \frac{2^{\mathcal{C}_2}}{2^{\mathcal{C}_1}+2^{\mathcal{C}_2}}. \tag{18}$$

Using this value of α^* in (16), we obtain

$$\begin{aligned}\mathcal{C} &= \frac{2^{\mathcal{C}_2}\mathcal{C}_1}{2^{\mathcal{C}_1}+2^{\mathcal{C}_2}} + \frac{2^{\mathcal{C}_2}\mathcal{C}_2}{2^{\mathcal{C}_1}+2^{\mathcal{C}_2}} + \frac{2^{\mathcal{C}_1}}{2^{\mathcal{C}_1}+2^{\mathcal{C}_2}}\log\frac{2^{\mathcal{C}_1}+2^{\mathcal{C}_2}}{2^{\mathcal{C}_1}} + \frac{2^{\mathcal{C}_2}}{2^{\mathcal{C}_1}+2^{\mathcal{C}_2}}\log\frac{2^{\mathcal{C}_1}+2^{\mathcal{C}_2}}{2^{\mathcal{C}_2}} \\ &= \log(2^{\mathcal{C}_1}+2^{\mathcal{C}_2}).\end{aligned}$$

Hence $2^{\mathcal{C}} = 2^{\mathcal{C}_1}+2^{\mathcal{C}_2}$.

iii) This is a union channel with channel one being a binary symmetric channel with crossover probability 0.1, and channel 2 being the trivial channel whose only input and output is the symbol 2. We therefore have [20%]

$$\mathcal{C}_1 = 1 - H_2(0.9) = 0.531, \qquad \mathcal{C}_2 = 0,$$

from which the capacity is $\mathcal{C} = \log(2^{\mathcal{C}_1}+2^{\mathcal{C}_2}) = 1.29$. The capacity achieving input distribution for channel 1 is $P(0) = P(1) = \frac{1}{2}$. From (18), the optimal value $\alpha^* = 0.591$, and hence the maximising input distribution is

$$P(X=0) = \frac{0.591}{2} = 0.2955, \qquad P(X=1) = 0.2955, \qquad P(X=2) = 0.409.$$

Question 4

(a) Block length $n = 6$, $(n - k) = 4$, hence dimension $k = 2$, , rate $k/n = 2/6 = 1/3$. [10%]

(Some candidates astutely observed that $\mathbf{H}$ had only three independent rows, so the true dimension and rate were 3 and 1/2, respectively. This answer was also given full marks.

(b) The codeword is $\underline{c} = [1, c_2, c_3, c_4, 1, 1]$. Using $\underline{c}\,\mathbf{H}^T = \underline{0}$, we get the following equations: [15%]

$$1 + c_2 + c_4 = 0, \qquad 1 + c_3 + 1 = 0, \qquad c_2 + c_3 + 1 = 0, \qquad c_4 + 1 + 1 = 0,$$

from which $c_3 = 0,\ c_2 = 1,\ c_4 = 0$, and hence $\underline{c} = [1, 1, 0, 0, 1, 1]$.

(c) Since each column has two ones, and row has three ones, the edge perspective polynomials are [5%]

$$\lambda(x) = x, \quad \rho(x) = x^2.$$

(d) The variable-to-check message along an edge depends on the channel output and the incoming message along the other edge. The outgoing message is erased if and only if the channel input is erased *and* the incoming message is an erasure. Hence

$$p_t = \epsilon q_t. \tag{19}$$

The check-to-variable message along an edge in iteration t is erased if *any* of the incoming messages (sent in iteration $t - 1$) along the two other edges is erased. Therefore,

$$q_t = 1 - P(\text{no incoming messages erased}) = 1 - (1 - p_{t-1})^2. \tag{20}$$

Combining the two equations, we get [20%]

$$p_t = \epsilon \left(1 - (1 - p_{t-1})^2\right).$$

The initial variable to check messages are erased if and only if the channel output corresponding to that variable is erased. Therefore $p_0 = \epsilon$.

(e) The key assumption when computing each outgoing message is the independence of the incoming messages arriving along each edge, i.e., we treat the probability of erasure along each edge as independent. This is not strictly true in a graph with cycles, however it is a reasonable assumption when the number of variable nodes (and check nodes) is very large and the factor graph has been generated by picking one at random from the ensemble with a given degree distribution. [10%]

(f) We note that outputs in the range $[-1, 0)$ can be produced only from the input $x = 0$, and outputs in the range $(1, 2]$ can be produced only from the input $x = 1$. Therefore, the likelihood ratio is [10%]

$$L(y) = \frac{f(y \,|\, x = 0)}{f(y \,|\, x = 1)} = \begin{cases} \frac{1/2}{0} = \infty, & -1 \leq y < 0, \\ \frac{1/2}{1/2} = 1, & 0 \leq y \leq 1, \\ \frac{0}{1/2} = 0, & 1 < y \leq 2. \end{cases} \tag{21}$$

(g) From the likelihood ratio, we observe that the channel is equivalent to a binary erasure channel: the output is an erasure if it lies in the interval $[0, 1]$, otherwise the input symbol is uniquely determined from the output. Therefore the codeword corresponding to the output sequence $\underline{y} = [0.1,\ 1.2,\ -0.6,\ 1.7,\ 0.5,\ 0.9]$ has the form $\underline{c} = [c_1,\ 1,\ 0,\ 1,\ c_5,\ c_6]$. Using $\underline{c}\,\mathbf{H}^T = \underline{0}$, we find:

$$c_1 + 1 + 1 = 0, \quad c_1 + c_6 = 0, \qquad 1 + c_5 = 0,$$

or $c_1 = 0, c_6 = 0, c_5 = 1$. The codeword is $\underline{c} = [0,\ 1,\ 0,\ 1,\ 1,\ 0]$. [15%]

(h) From the equivalent binary erasure channel explained in part (g), the input symbols are (0/1), and the output is an erasure with probability 1/2, or equal to the input symbol with probability 1/2. Since the capacity of the binary erasure channel with erasure probability ϵ is $(1-\epsilon)$, the capacity of the channel is $1-0.5=0.5$ bits/channel use. [15%]