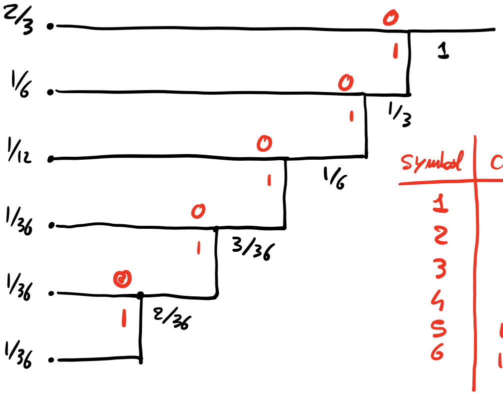
Information Theory and Applications (Module ID: 40981)

Written Test 1: Solutions

1. [15 %] Consider the discrete i.i.d. source X defined on the alphabet $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$, with probability distribution $P_X(1) = 2/3, P_X(2) = 1/6, P_X(3) = 1/12, P_X(4) = P_X(5) = P_X(6) = 1/36$. Find a Huffman code for this source that encodes single symbols (i.e., with input block length $n = 1$), and compare the achieved average coding length with the source entropy $H(X)$.

Solution:



Symbol	codeword
1	0
2	10
3	110
4	1110
5	11110
6	11111

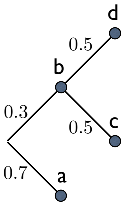
$$\ell_1 = \frac{2}{3}\cdot 1 + \frac{1}{6}\cdot 2 + \frac{1}{12}\cdot 3 + \frac{1}{36}\cdot 4 + \frac{1}{36}\cdot 5 + \frac{1}{36}\cdot 5$$

$$= 1.6389$$

$$H(X) = \frac{2}{3}\log\frac{3}{2} + \frac{1}{6}\log 6 + \frac{1}{12}\log 12 + \frac{3}{36}\log 36$$

$$= 1.5504 \quad \text{bits}$$

2. [30 %] Consider the tree source $\{X_i : i = 1, 2, 3, \ldots\}$ in the figure:



a) Compute its entropy rate (in bits per symbol).

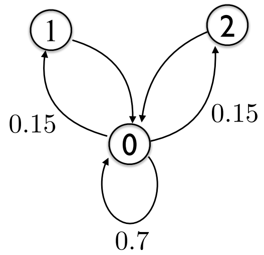
b) Consider the following sequence generated by the source:

$$aaabcaabdabcaabdaaabcaabd$$

parse the sequence using Lempel-Ziv parsing, and encode the resulting sequence using the Lempel-Ziv algorithm. Then, calculate the resulting normalized coding length in bit/symbol (notice: in LZ coding the output length in bits is given by $c(\lceil \log c \rceil + \lceil \log |\mathcal{X}| \rceil)$ where c is the number of phrases in the LZ parsing).

c) We neglect the memory in the source and treat the source as if it was i.i.d., using a Huffman code with probabilities $p_a = \mathbb{P}(X_i = a)$, $p_b = \mathbb{P}(X_i = b)$, $p_c = \mathbb{P}(X_i = c)$, $p_d = \mathbb{P}(X_i = d)$. Provide such Huffman code, and compute its rate (average normalized coding length in bits per source symbol).

Hint: here the key difficulty is to compute the marginal distribution of the source, i.e., the probabilities p_a, p_b, p_c, p_d given above. A way to do this is to represent the source as generated by an underlying Markov Chain that produces a source symbol in correspondence of each state transition, and compute the stationary distribution of this Markov Chain. Then, use the law of total probability conditioning on the Markov Chain state (distributed according to its stationary distribution). For example, a possible chain is represented in the figure, together with the symbol-transition correspondence:



transition	symbol
0-0	a
0-1	b
0-2	b
1-0	c
2-0	d

Solution:

a) Here we use the result seen in class (one of the problem sets) about tree sources and the renewal-reward theorem. The entropy rate is given by

$$H(\mathcal{X}) = \frac{\text{entropy of the ``terminal leaves pmf''}}{\text{mean renewal time}}$$

where the mean renewal time is the average time (number of steps, or symbols) for which the source resets, i.e., it goes back to the root of the tree.

The source emits symbols “a” with probability 0.7, symbols “bc” with probability $0.3 * 0.5 = 0.15$, and symbols ”bd” with probability $0.3 * 0.5 = 0.15$. Hence, the entropy of the terminal leaves a, bc, and bd, is

$$H = -0.7 * \log 0.7 - 0.15 * \log 0.15 - 0.15 * \log 0.15 = 1.1813 \text{ bits}$$

The mean renewal time is

$$T = 0.7 * 1 + 0.15 * 2 + 0.15 * 2 = 1.3 \text{ symbols}$$

Finally, the entropy rate is given by

$$H(\mathcal{X}) = \frac{1.1813}{1.3} = 0.9087 \text{ bits/symbol}$$

b) The LZ parsing is given by

$$a, aa, b, c, aab, d, ab, ca, abd, aaa, bc, aabd$$

We have 12 blocks. The encoding is give by

$$(0,a),(1,a),(0,b),(0,c),(2,b),(0,d),(1,b),(4,a),(7,d),(2,a),(3,c),(5,d)$$

Using $\lceil \log(12) \rceil = 4$ bits for the index, and $\lceil \log |\mathcal{X}| \rceil = 2$ bits for the symbols, the total length in binary digits is $12 * (4 + 2) = 72$, yielding a normalized length of $72/25 = 2.88$ bit/symbol. Noice that this is much larger than the entropy rate since the sequence is too short for the asymptotic optimality to manifest.

c) The stationary distribution of the MC in the figure is obtained in the derivation below

$$\underline{\underline{P}} = \begin{bmatrix} 0.7 & 0.15 & 0.15 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

$$\underline{\pi} = \underline{\pi}\,\underline{\underline{P}} \implies \underline{\pi}\left(\underline{\underline{P}} - \underline{\underline{I}}\right) = 0$$

$$(\pi_1 \ \ \pi_2 \ \ \pi_3) \begin{bmatrix} -0.3 & 0.15 & 0.15 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} = 0$$

$$\begin{cases} -0.3\,\pi_1 + \pi_2 + \pi_3 = 0 \\ 0.15\,\pi_1 - \pi_2 = 0 \\ \pi_1 + \pi_2 + \pi_3 = 1 \end{cases}$$

$$\implies \quad \pi_2 = 0.15\,\pi_1$$

$$\pi_3 = 0.3\,\pi_1 - 0.15\,\pi_1 = 0.15\,\pi_1$$

$$\pi_1 + 0.15\,\pi_1 + 0.15\,\pi_1 = 1$$

$$\boxed{\begin{aligned} \pi_1 &= \frac{1}{1.3} = 0.7692 \\ \pi_2 &= 0.1154 \\ \pi_3 &= 0.1154 \end{aligned}}$$

It follows that the marginal probability of symbol "a" is

$$p_a = \pi_1 * P_{0,0} = 0.7692 * 0.7 = 0.5384$$

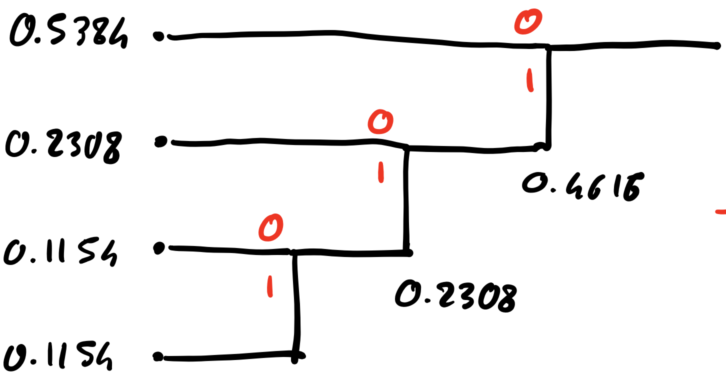
The marginal probability of symbol "b" is

$$p_b = \pi_1 * P_{0,1} + \pi_1 * P_{0,2} = 0.2308$$

The marginal probability of symbols "c" and "d" is

$$p_c = p_d = \pi_2 = \pi_3 = 0.1154$$

The corresponding Huffman code is given below



a	0
b	10
c	110
d	111

$$\ell_1 = 1\cdot 0.5384 + 2\cdot 0.2308 + 3\cdot 0.1154 + 3\cdot 0.1154$$

$$= 1.2308$$

3. [15 %] Let X be a discrete memoryless source on an alphabet $\mathcal{X}$ with pmf P_X and let $g : \mathcal{X} \to \mathbb{R}$ be a function (not generally non-negative).

   We want to generalize the typical average lemma to generic functions g for which the mean $\mathbb{E}[g(X)] = \sum_{x\in\mathcal{X}} g(x)P_X(x)$ exists, i.e., such that $\sum_{x\in\mathcal{X}} |g(x)|P_X(x) < \infty$.

   In particular, prove that for any $\mathbf{x} \in \mathcal{T}_\epsilon^{(n)}(X)$ we have

$$\mathbb{E}[g(X)] - \delta(\epsilon) \leq \frac{1}{n}\sum_{i=1}^{n} g(x_i) \leq \mathbb{E}[g(X)] + \delta(\epsilon)$$

   where $\delta(\epsilon)$ vanishes as $\epsilon \downarrow 0$.

   Hint: for any function you can always write $g(x) = g_+(x) - g_-(x)$ where $g_+(x) = \max\{g(x), 0\}$ and $g_-(x) = -\min\{g(x), 0\}$ are the positive and negative parts of g.

   Solution:

From the definition, we can write

$$g(x) = g_+(x) - g_-(x)$$

where g_+ and g_- are non-negative functions. Then, the typical average lemma yields that, for any $\mathbf{x} \in \mathcal{T}_\epsilon^{(n)}(X)$,

$$(1-\epsilon)\mathbb{E}[g_+(X)] \leq \frac{1}{n}\sum_{i=1}^n g_+(x_i) \leq (1+\epsilon)\mathbb{E}[g_+(X)]$$

and

$$(1-\epsilon)\mathbb{E}[g_-(X)] \leq \frac{1}{n}\sum_{i=1}^n g_-(x_i) \leq (1+\epsilon)\mathbb{E}[g_-(X)]$$

Multiplying by -1 all terms of the second inequality, we obtain

$$-(1-\epsilon)\mathbb{E}[g_-(X)] \geq -\frac{1}{n}\sum_{i=1}^n g_-(x_i) \geq -(1+\epsilon)\mathbb{E}[g_-(X)]$$

or, equivalently,

$$-(1+\epsilon)\mathbb{E}[g_-(X)] \leq -\frac{1}{n}\sum_{i=1}^n g_-(x_i) \leq -(1-\epsilon)\mathbb{E}[g_-(X)]$$

Finally, adding up the first and the last inequalities we find

$$(1-\epsilon)\mathbb{E}[g_+(X)]-(1+\epsilon)\mathbb{E}[g_-(X)] \leq \frac{1}{n}\sum_{i=1}^n g_+(x_i) - \frac{1}{n}\sum_{i=1}^n g_-(x_i) \leq (1+\epsilon)\mathbb{E}[g_+(X)]-(1-\epsilon)\mathbb{E}[g_-(X)]$$

which can be rewritten as

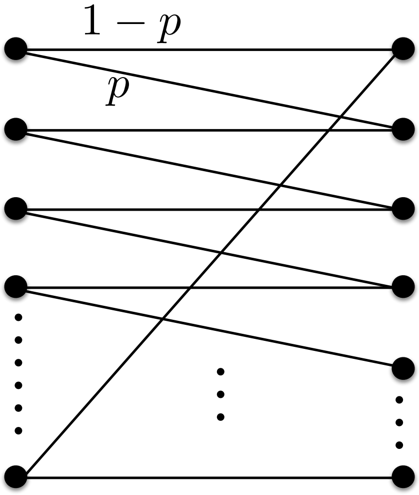
$$\mathbb{E}[g(X)] - \epsilon(\mathbb{E}[g_+(X)] + \mathbb{E}[g_-(X)]) \leq \frac{1}{n}\sum_{i=1}^n g(x_i) \leq \mathbb{E}[g(X)] + \epsilon(\mathbb{E}[g_+(X)] + \mathbb{E}[g_-(X)])$$

from which the final result follows by letting

$$\delta(\epsilon) = \epsilon(\mathbb{E}[g_+(X)] + \mathbb{E}[g_-(X)])$$

Notice that the absolute summability of the function $g(x)$ guarantees that both $\mathbb{E}[g_+(X)]$ and $\mathbb{E}[g_-(X)]$ exist (finite), and therefore we have that $\delta(\epsilon)$ vanishes as $\epsilon \downarrow 0$.

4. [10 %] Calculate the capacity of the DMC represented in the figure below, where $|\mathcal{X}| = |\mathcal{Y}| = K \geq 2$ (some generic integer), and $p \in (0, 1)$. The result is an expression in terms of K and p.



Solution:

This is a strongly symmetric channel for which

$$C = \log|\mathcal{X}| - \mathcal{H}((1-p,p,0,\dots,0) = \log K - \mathcal{H}_2(p)$$

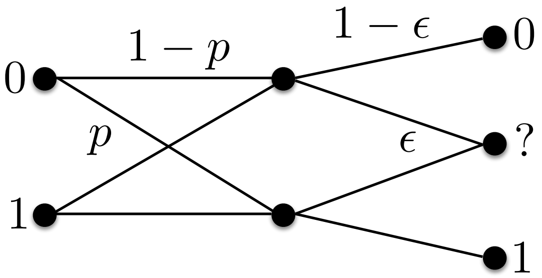
5. [30 %] Consider the concatenation of the BSC and the BEC given in the figure below.

a) Calculate the capacity of the concatenated channel in terms of $p \in (0,1/2)$ and $\epsilon \in (0,1)$.

b) Suppose now that in between the BSC and the BEC we introduce a relay, i.e., a device that can process arbitrarily long blocks of the output of the BSC and translate them into arbitrarily long blocks of input for the BEC. Show that the capacity of the concatenation with the relay in this case is given by $C_{\rm relay} = \min\{C_{\rm bsc}, C_{\rm bec}\}$ where $C_{\rm bsc}$ and $C_{\rm bec}$ are the individual capacities of the BSC and of the BEC, respectively.

c) Show that the capacity of the concatenated channel (without relay) is upperbounded by $C_{\rm relay}$.

Hint: in order to answer question b) you have to prove an achievability and a converse. For the achievability, you may consider a specific realization of the relay that decoded the (coded) input block of the BSC, retrieves the information message, and re-encode the message into a codeword for the BEC (this type of relaying is called "decode and forward", and in this case it turns out to be optimal). For the converse, consider the data processing inequality.



Solution:

a) Let's call the input of the BSC X, its output (and input of the BEC) Y, and the output of the BEC Z. The concatenated channel from X to Z has transition probability matrix

$$\mathbf{P} = \left[\begin{array}{ccc} (1-p)(1-\epsilon) & \epsilon & p(1-\epsilon) \\ p(1-\epsilon) & \epsilon & (1-p)(1-\epsilon) \end{array} \right]$$

This is a weakly symmetric channel, whose capacity is given by

$$C = \max_{P_X} I(X;Z) = \max_{P_X} H(Z) - \mathcal{H}((1-p)(1-\epsilon), \epsilon, p(1-\epsilon))$$

Our goal is to maximize the entropy of Z with respect to the input distribution P_X. Let $P_X(1) = \alpha$ and $P_X(0) = 1-\alpha$. Then,

$$\begin{aligned} P_Z(0) &= (1-\alpha)(1-p)(1-\epsilon) + \alpha p(1-\epsilon) \\ P_Z(1) &= \alpha(1-p)(1-\epsilon) + (1-\alpha)p(1-\epsilon) \\ P_Z(?) &= (1-\alpha)\epsilon + \alpha\epsilon = \epsilon \end{aligned}$$

such that

$$H(Z) = \mathcal{H}\left(P_Z(0), \epsilon, P_Z(1)\right)$$

Remember that entropy is increased as the pmf gets closer to a uniform distribution. Since $P_Z(?) = \epsilon$, independently of α, we maximize the entropy by making $P_Z(0) = P_Z(1)$, i.e., by choosing $\alpha = 1/2$. The result is:

$$C_{\text{conc}} = \mathcal{H}\left(\frac{1-\epsilon}{2}, \epsilon, \frac{1-\epsilon}{2}\right) - \mathcal{H}\left((1-p)(1-\epsilon), \epsilon, p(1-\epsilon)\right).$$

b) Define

$$C_{\text{bsc}} = \max_{P_X} \; I(X;Y)$$

and

$$C_{\text{bec}} = \max_{P_Y} \; I(Y;Z)$$

It is clear that we can encode information bits at rate up to C_{bsc} and these can be reliably decoded by the intermediate relay. Then, the relay can re-encode the bits at rate up to C_{bec} and send them reliably to the destination. Hence, this decode and forward strategy achieves the rate $\min\{C_{\text{bsc}}, C_{\text{bec}}\}$.

In order to show that this is indeed the capacity, we need to show that it is impossible to do better (i.e., we need a converse statement). Notice that the relay channel can be improved by replacing either the BSC or the BEC with a perfect channel that does not cause any error or erasure, with transition matrix

$$\mathbf{P} = \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right]$$

The capacity of the original relay channel cannot be larger than that of either enhanced channels. Hence, by contradiction, if there exist codes of rate $R > C_{\text{bsc}}$ and vanishing probability of error for the relay channel, these would violate the capacity of of the enhanced channel obtained by replacing the BEC with the perfect channel. Following a symmetric argument, if there exist codes of rate $R > C_{\rm bec}$ and vanishing probability of error for the relay channel, these would violate the capacity of of the enhanced channel obtained by replacing the BSC with the perfect channel. It follows that $C_{\rm relay} = \min\{C_{\rm bsc}, C_{\rm bec}\}$.

c) Given the Markov chain $X \to Y \to Z$, the data processing inequality implies:

$$\begin{aligned} I(X;Y) &\geq I(X;Z) \\ I((Y;Z) &\geq I(X;Z). \end{aligned}$$

The above two inequalities hold for any joint distribution of X, Y, Z satisfying the above said Markov chain. Let $X^{(1)}$ denote the X maximizing $I(X;Y)$, and $X^{(c)}$ denote the X maximizing $I(X;Z)$. Then, we have

$$C_{\rm bsc} = I(X^{(1)};Y) \geq I(X^{(c)};Y) \geq I(X^{(c)};Z) = C_{\rm conc}$$

Consider now the second inequality. With a similar argument, let $Y^{(2)}$ be the distribution of Y that maximizes $I(Y;Z)$, and let $Y^{(c)}$ induced by $X^{(c)}$. Then, we have

$$C_{\rm conc} = I(X^{(c)};Z) \leq I(Y^{(c)};Z) \leq I(Y^{(2)};Z) = C_{\rm bec}$$

Hence, we conclude that $C_{\rm conc} \leq C_{\rm relay}$.