Information Theory I

Fall Semester 2017

Prof. Dr. A. Lapidoth

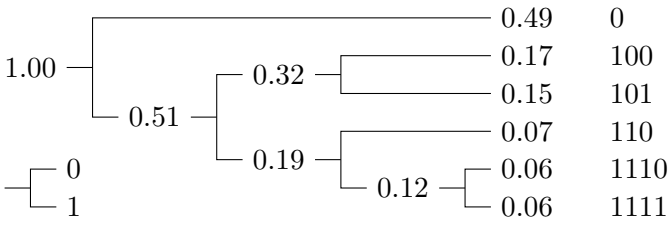
Signal and Information Processing Laboratory
Institut für Signal- und Informationsverarbeitung

Model Answers to Exercise 5 of October 18, 2017

http://www.isi.ee.ethz.ch/teaching/courses/it1.html

Problem 1 *Huffman Coding*

a) The Huffman tree for this distribution is

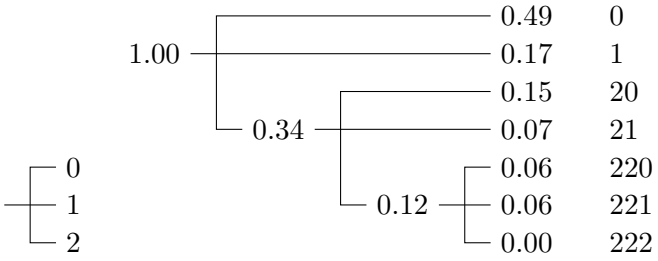


The expected length of the codewords for the binary Huffman code is

$$\begin{aligned}\mathrm{L}(C) &= \mathsf{E}[l(X)] \\ &= 0.49 \cdot 1 + 0.17 \cdot 3 + 0.15 \cdot 3 + 0.07 \cdot 3 + 0.06 \cdot 4 + 0.06 \cdot 4 \\ &= 2.14 \text{ bits.}\end{aligned}$$

Note that $H(X) \approx 2.105$ bits.

b) The Huffman procedure can be adapted to ternary codes (by combining the three least likely symbols in each step). But note that every full ternary tree contains $3k+2$ leaves (for some integer k), so to make sure that combining the three least likely symbols does not occur any loss in optimality, we might need to add dummy symbols until the number of symbols is equal to $3k+2$. The ternary Huffman tree then is



This code has an expected length

$$\begin{aligned}\mathrm{L}(C) &= \mathsf{E}[l(X)] \\ &= 0.49 \cdot 1 + 0.17 \cdot 1 + 0.15 \cdot 2 + 0.07 \cdot 2 + 0.06 \cdot 3 + 0.06 \cdot 3 \\ &= 1.46 \text{ ternary symbols.}\end{aligned}$$

Note that $H_3(X) \approx 1.328$ ternary symbols.

Problem 2 *Bad Codes*

a) $\{0, 10, 11\}$ is a Huffman code, e.g., for the distribution $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$.

b) The code $\{00, 01, 10, 110\}$ can be shortened to $\{00, 01, 10, 11\}$ without losing its instantaneous property, and is therefore not optimal. Thus, it cannot be a Huffman code. Alternatively, it is not a Huffman code because there is a unique longest codeword.

c) The code $\{01, 10\}$ can be shortened to $\{0, 1\}$ without losing its instantaneous property, and is therefore not optimal and, consequently, not a Huffman code.

Problem 3 *Optimal Codeword Lengths*

We give two proofs for both parts: short proofs and more verbose proofs.

a) We perform induction on m. For $m = 2$, the claim is trivially true. For $m = 3$, the claim is also true: the two least probable messages (i.e., 2 and 3) will be assigned a codeword of length 2 and message 1 will be assigned a codeword of length 1. For $m \geq 4$, we will show that $p_{m-1} + p_m < \frac{2}{5}$. Because $p_1 > \frac{2}{5}$, message 1 will still be the unique most probable message after one step of Huffman's algorithm, i.e., $p_1 > p'_2 \geq \ldots \geq p'_{m-1}$. Therefore, by induction, message 1 will not be combined with any other message except in the last step of Huffman's algorithm, which proves the claim for all m.

   We show $p_{m-1} + p_m < \frac{2}{5}$ for $m \geq 4$ by contradiction. Assume that $p_{m-1} + p_m \geq \frac{2}{5}$. Then, we must have $p_2 \geq \frac{1}{5}$ because otherwise $p_{m-1} + p_m \leq p_2 + p_2 < \frac{2}{5}$, which would contradict the assumption. But this leads to the desired contradiction, because $p_1 > \frac{2}{5}$, $p_2 \geq \frac{1}{5}$, and $p_{m-1} + p_m \geq \frac{2}{5}$ is impossible as the probabilities sum up to more than one. The contradiction can be visualized as follows:

$$\underbrace{p_1}_{>\frac{2}{5}} > \underbrace{p_2}_{\geq\frac{1}{5}} \geq \ldots \geq \underbrace{p_{m-1} \geq p_m}_{p_{m-1}+p_m \geq \frac{2}{5}}$$

b) Note that $m \geq 4$ must hold: if the most probable message has probability $p_1 < \frac{1}{3}$, it is not possible to have three or fewer messages, because otherwise the probabilities would not sum up to one. Now run Huffman's algorithm until three messages are left. If, in any step, message 1 was combined with another message, its codeword length will be at least 2, and the claim follows. Otherwise, it will be combined with another message in the next step: As $p_1 < \frac{1}{3}$, at least one of the other two remaining messages must have probability larger than $\frac{1}{3}$ (in order for the probabilities to sum up to one). Therefore, message 1 cannot be the most probable message, it will be combined with another message in the next step, and its codeword length will be at least 2.

For the verbose proofs, we can without loss of generality assume

$$l_1 \leq l_2 \leq \ldots \leq l_m$$

because of the assumption

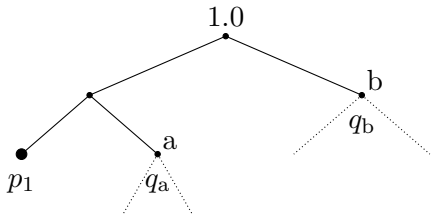
$$p_1 > p_2 \geq p_3 \geq \ldots \geq p_m. \tag{1}$$

a) We prove by contradiction that for

$$p_1 > \frac{2}{5} \tag{2}$$

the most probable codeword x_1 must have length $l_1 = 1$.

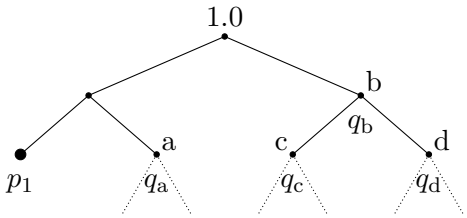
Suppose that there exists a Huffman code with $l_1 = 2$. Then, the corresponding tree looks as follows:



where node a (which may but need not be a codeword) with probability q_a must exist because otherwise the corresponding code would be strictly suboptimal and hence not a Huffman code. Thus, we know

$$p_1 + q_a + q_b = 1, \tag{3}$$
$$q_a \leq q_b, \tag{4}$$
$$p_1 \leq q_b, \tag{5}$$

where (3) follows from the properties of probability mass functions, and (4) and (5) hold because otherwise the code would not be optimal. (Remember that the code under consideration has to be optimal since it is by assumption a Huffman code.) Equations (5) and (2) imply $q_b > \frac{2}{5}$. Moreover, because of (1), node b with probability q_b cannot be a codeword. Thus, the tree has to look as follows:



Now, we can consider the construction algorithm of the Huffman code. If node a and p_1 are combined before nodes c and d, then we know that

$$q_c \geq \max\{p_1, q_a\} > \frac{2}{5},$$
$$q_d \geq \max\{p_1, q_a\} > \frac{2}{5}.$$

These two conditions and (2) imply $p_1 + q_c + q_d > \frac{6}{5}$, which contradicts (3). Thus, nodes c and d are combined first. Equations (3), (2), and (5) imply $q_a < \frac{1}{5}$. We therefore have

$$q_c + q_d = q_b > \frac{2}{5},$$
$$q_c \leq \min\{p_1, q_a\} < \frac{1}{5},$$
$$q_d \leq \min\{p_1, q_a\} < \frac{1}{5},$$

which contradicts $q_{\mathrm{b}} > \frac{2}{5}$. We conclude that $\ell(x_1) = 2$ must be wrong.

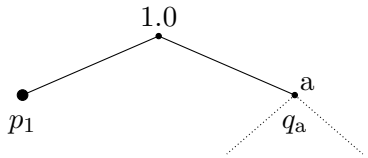
As we argue next, the assumption $l_1 > 2$, leads to a contradiction because of the above analysis for the case $l_1 = 2$. To see this, consider the node on the shortest path between the node corresponding to the probability p_1 and the root that is separated from the root by two edges. Observe that the probability q, which corresponds to this node, has to be greater than p_1. Hence, we obtain a contradiction from the analysis for the case $l_1 = 2$ and we therefore conclude that the length of the codeword for x_1 must be 1.

b) We prove by contradiction that for

$$p_1 < \frac{1}{3} \tag{6}$$

the most probable codeword x_1 must have length $l_1 \geq 2$.

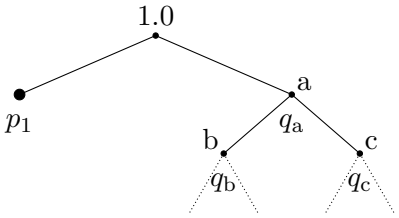
For contradiction, suppose there exists a Huffman code with $l_1 = 1$. Then the corresponding tree looks as follows:



Thus, we know

$$p_1 + q_{\mathrm{a}} = 1.$$

The last relation and (6) imply that $q_{\mathrm{a}} > \frac{2}{3}$. Because of (1) p_1 is the largest probability and node a therefore cannot be a codeword. Thus the tree has to look as follows:



Moreover,

$$q_{\mathrm{b}} + q_{\mathrm{c}} = q_{\mathrm{a}} > \frac{2}{3}, \tag{7}$$

$$q_{\mathrm{b}} \leq p_1, \tag{8}$$

$$q_{\mathrm{c}} \leq p_1, \tag{9}$$

where (8) and (9) hold since otherwise in the construction of the Huffman code one would have combined either node b or node c with the node that corresponds to x_1 instead of combining the nodes b and c. From (8) and (9) we obtain

$$q_{\mathrm{b}} + q_{\mathrm{c}} \leq 2 \cdot p_1 < \frac{2}{3} \tag{10}$$

and hence a contradiction to (7). Thus, the assumption that $l_1 = 1$ must be wrong.

Problem 4 *AEP*

a) From the weak law of large numbers, we know that for IID random variables Y_i, we have

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Y_i = \mathsf{E}[Y_1]$$

in probability. We use the fact that the random variables $X_1, X_2, \ldots, X_n$ are IID (and hence $-\log q(X_1), -\log q(X_2), \ldots, -\log q(X_n)$ are also IID) to obtain

$$\begin{aligned}
\lim_{n\to\infty}\left(-\frac{1}{n}\log q(X_1,\ldots,X_n)\right) &= \lim_{n\to\infty}\left(-\frac{1}{n}\log\prod_{i=1}^{n} q(X_i)\right)\\
&= \lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}\big(-\log q(X_i)\big)\right)\\
&= \mathsf{E}[-\log q(X_1)] \qquad \text{in probability}\\
&= \sum_{x=1}^{m} p(x)\log\frac{1}{q(x)}\\
&= \sum_{x=1}^{m} p(x)\log\frac{p(x)}{q(x)} - \sum_{x=1}^{m} p(x)\log p(x)\\
&= D(p\|q) + H(p).
\end{aligned}$$

b) Similarly, we get

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\log\frac{p(X_1,\ldots,X_n)}{q(X_1,\ldots,X_n)} &= \lim_{n\to\infty}\left(\frac{1}{n}\log\prod_{i=1}^{n}\frac{p(X_i)}{q(X_i)}\right)\\
&= \lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}\log\frac{p(X_i)}{q(X_i)}\right)\\
&= \mathsf{E}\left[\log\frac{p(X_1)}{q(X_1)}\right] \qquad \text{in probability}\\
&= \sum_{x=1}^{m} p(x)\log\frac{p(x)}{q(x)}\\
&= D(p\|q).
\end{aligned}$$

Problem 5 *Proof of Theorem 3.3.1 in Cover & Thomas: High-Probability Sets and the Typical Set*

Without loss of generality we will concentrate on the logarithm to the base 2 and the entropy given in bits.

a) Using $\Pr(\mathcal{A}\cup\mathcal{B}) = \Pr(\mathcal{A}) + \Pr(\mathcal{B}) - \Pr(\mathcal{A}\cap\mathcal{B})$, we get

$$\begin{aligned}
\Pr(\mathcal{A}\cap\mathcal{B}) &= \underbrace{\Pr(\mathcal{A})}_{\geq 1-\epsilon_1} + \underbrace{\Pr(\mathcal{B})}_{\geq 1-\epsilon_2} - \underbrace{\Pr(\mathcal{A}\cup\mathcal{B})}_{\leq 1}\\
&\geq 1-\epsilon_1 + 1 - \epsilon_2 - 1\\
&= 1 - \epsilon_1 - \epsilon_2.
\end{aligned}$$

For n sufficiently large, we have $\Pr\big(\mathcal{A}_\epsilon^{(n)}\big) \geq 1-\epsilon$, and choosing $\mathcal{A} \triangleq \mathcal{A}_\epsilon^{(n)}$, $\mathcal{B} \triangleq \mathcal{B}_\delta^{(n)}$, $\epsilon_1 = \epsilon$, and $\epsilon_2 = \delta$ immediately leads to

$$\Pr\big(\mathcal{A}_\epsilon^{(n)} \cap \mathcal{B}_\delta^{(n)}\big) \geq 1-\epsilon-\delta.$$

b) i) This follows from Part a).

ii) This holds because $\mathcal{A}_\epsilon^{(n)}$ and $\mathcal{B}_\delta^{(n)}$ are sets of sequences $\mathbf{x} \in \mathcal{X}^n$.

iii) This holds because every $\mathbf{x} \in \mathcal{A}_\epsilon^{(n)}$, and thus every $\mathbf{x} \in \mathcal{A}_\epsilon^{(n)} \cap \mathcal{B}_\delta^{(n)}$, must satisfy

$$P_{\mathbf{X}}(\mathbf{x}) \leq 2^{-n(H(X)-\epsilon)}.$$

iv) This holds because the summands do not depend on $\mathbf{x}$.

v) This holds because the number of elements in $\mathcal{A} \cap \mathcal{B}$ cannot be larger than the number of elements in $\mathcal{B}$.

c) We set $\epsilon = \min\big\{\frac{\delta'}{2}, \frac{1-\delta}{2}\big\} > 0$ (we will quickly see why). From Part b) we know that there exists a n_0 such that for all $n \geq n_0$,

$$\big|\mathcal{B}_\delta^{(n)}\big| \geq (1-\epsilon-\delta)2^{n(H(X)-\epsilon)}.$$

Because $1-\epsilon-\delta \geq \frac{1-\delta}{2} > 0$ by our choice of ϵ, we can take the logarithm on both sides and divide both sides by n, which leads to

$$\frac{1}{n}\log_2\big|\mathcal{B}_\delta^{(n)}\big| \geq H(X) - \epsilon - \frac{1}{n}\log_2\frac{1}{1-\epsilon-\delta}.$$

For $n \geq \max\big\{n_0, \frac{1}{\epsilon}\log_2\frac{1}{1-\epsilon-\delta}\big\}$, we obtain

$$\frac{1}{n}\log_2\big|\mathcal{B}_\delta^{(n)}\big| \geq H(X) - \epsilon - \epsilon \geq H(X) - \delta',$$

where the last inequality again follows from our choice of ϵ.