ESE 532
HW#4 Solutions

Problem 4.1

$$\begin{aligned} H(X) &= -\sum_{i=1}^{6} p_i \log_2 p_i = -(0.1\log_2 0.1 + 0.2\log_2 0.2 \\ &\quad +0.3\log_2 0.3 + 0.05\log_2 0.05 + 0.15\log_2 0.15 + 0.2\log_2 0.2) \\ &= 2.4087 \quad \text{bits/symbol} \end{aligned}$$

If the source symbols are equiprobable, then $p_i = \frac{1}{6}$ and

$$H_u(X) = -\sum_{i=1}^{6} p_i \log_2 p_i = -\log_2 \frac{1}{6} = \log_2 6 = 2.5850 \quad \text{bits/symbol}$$

As it is observed the entropy of the source is less than that of a uniformly distributed source.

Problem 4.2
If the source is uniformly distributed with size N, then $p_i = \frac{1}{N}$ for $i = 1, \ldots, N$. Hence,

$$\begin{aligned} H(X) &= -\sum_{i=1}^{N} p_i \log_2 p_i = -\sum_{i=1}^{N} \frac{1}{N} \log_2 \frac{1}{N} \\ &= -\frac{1}{N} N \log_2 \frac{1}{N} = \log_2 N \end{aligned}$$

Problem 4.3

$$H(X) = -\sum_i p_i \log p_i = \sum_i p_i \log \frac{1}{p_i}$$

By definition the probabilities p_i satisfy $0 < p_i \leq 1$ so that $\frac{1}{p_i} \geq 1$ and $\log \frac{1}{p_i} \geq 0$. It turns out that each term under summation is positive and thus $H(X) \geq 0$. If X is deterministic, then $p_k = 1$ for some k and $p_i = 0$ for all $i \neq k$. Hence,

$$H(X) = -\sum_i p_i \log p_i = -p_k \log 1 = -p_k 0 = 0$$

Note that $\lim_{x\to 0} x \log x = 0$ so if we allow source symbols with probability zero, they contribute nothing in the entropy.

Problem 4.4
1)

$$\begin{aligned} H(X) &= -\sum_{k=1}^{\infty} p(1-p)^{k-1} \log_2 (p(1-p)^{k-1}) \\ &= -p\log_2(p) \sum_{k=1}^{\infty} (1-p)^{k-1} - p\log_2(1-p) \sum_{k=1}^{\infty} (k-1)(1-p)^{k-1} \\ &= -p\log_2(p) \frac{1}{1-(1-p)} - p\log_2(1-p) \frac{1-p}{(1-(1-p))^2} \\ &= -\log_2(p) - \frac{1-p}{p} \log_2(1-p) \end{aligned}$$

2) Clearly $p(X = k|X > K) = 0$ for $k \leq K$. If $k > K$, then

$$p(X = k|X > K) = \frac{p(X = k, X > K)}{p(X > K)} = \frac{p(1-p)^{k-1}}{p(X > K)}$$

But,

$$\begin{aligned} p(X>K) &= \sum_{k=K+1}^{\infty} p(1-p)^{k-1} = p\left(\sum_{k=1}^{\infty}(1-p)^{k-1} - \sum_{k=1}^{K}(1-p)^{k-1}\right) \\ &= p\left(\frac{1}{1-(1-p)} - \frac{1-(1-p)^K}{1-(1-p)}\right) = (1-p)^K \end{aligned}$$

so that

$$p(X=k|X>K) = \frac{p(1-p)^{k-1}}{(1-p)^K}$$

If we let $k = K + l$ with $l = 1, 2, \ldots$, then

$$p(X=k|X>K) = \frac{p(1-p)^K(1-p)^{l-1}}{(1-p)^K} = p(1-p)^{l-1}$$

that is $p(X = k|X > K)$ is the geometrically distributed. Hence, using the results of the first part we obtain

$$\begin{aligned} H(X|X>K) &= -\sum_{l=1}^{\infty} p(1-p)^{l-1}\log_2(p(1-p)^{l-1}) \\ &= -\log_2(p) - \frac{1-p}{p}\log_2(1-p) \end{aligned}$$

Problem 4.11
1)

$$\begin{aligned} H(X) &= -(.05\log_2 .05 + .1\log_2 .1 + .1\log_2 .1 + .15\log_2 .15 \\ &\quad +.05\log_2 .05 + .25\log_2 .25 + .3\log_2 .3) = 2.5282 \end{aligned}$$

2) After quantization, the new alphabet is $B = \{-4, 0, 4\}$ and the corresponding symbol probabilities are given by

$$\begin{aligned} p(-4) &= p(-5) + p(-3) = .05 + .1 = .15 \\ p(0) &= p(-1) + p(0) + p(1) = .1 + .15 + .05 = .3 \\ p(4) &= p(3) + p(5) = .25 + .3 = .55 \end{aligned}$$

Hence, $H(Q(X)) = 1.4060$. As it is observed quantization decreases the entropy of the source.

Problem 4.20
From the discussion in the beginning of Section 4.2 it follows that the total number of sequences of length n of a binary DMS source producing the symbols 0 and 1 with probability p and $1-p$ respectively is $2^{nH(p)}$. Thus if $p = 0.3$, we will observe sequences having $np = 3000$ zeros and $n(1-p) = 7000$ ones. Therefore,

$$\#\text{ sequences with 3000 zeros } \approx 2^{8813}$$

Another approach to the problem is via the Stirling's approximation. In general the number of binary sequences of length n with k zeros and $n-k$ ones is the binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

To get an estimate when n and k are large numbers we can use Stirling's approximation

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

Hence,

$$\#\text{ sequences with 3000 zeros } = \frac{10000!}{3000!7000!} \approx \frac{1}{21\sqrt{2\pi 30\cdot 70}}10^{10000}$$

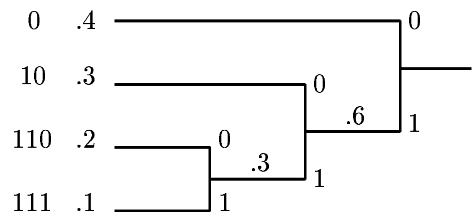
Problem 4.22
1) The entropy of the source is

$$H(X) = -\sum_{i=1}^{4} p(a_i)\log_2 p(a_i) = 1.8464 \quad \text{bits/output}$$

2) The average codeword length is lower bounded by the entropy of the source for error free reconstruction. Hence, the minimum possible average codeword length is $H(X) = 1.8464$.

3) The following figure depicts the Huffman coding scheme of the source. The average codeword length is

$$\bar{R}(X) = 3 \times (.2 + .1) + 2 \times .3 + .4 = 1.9$$

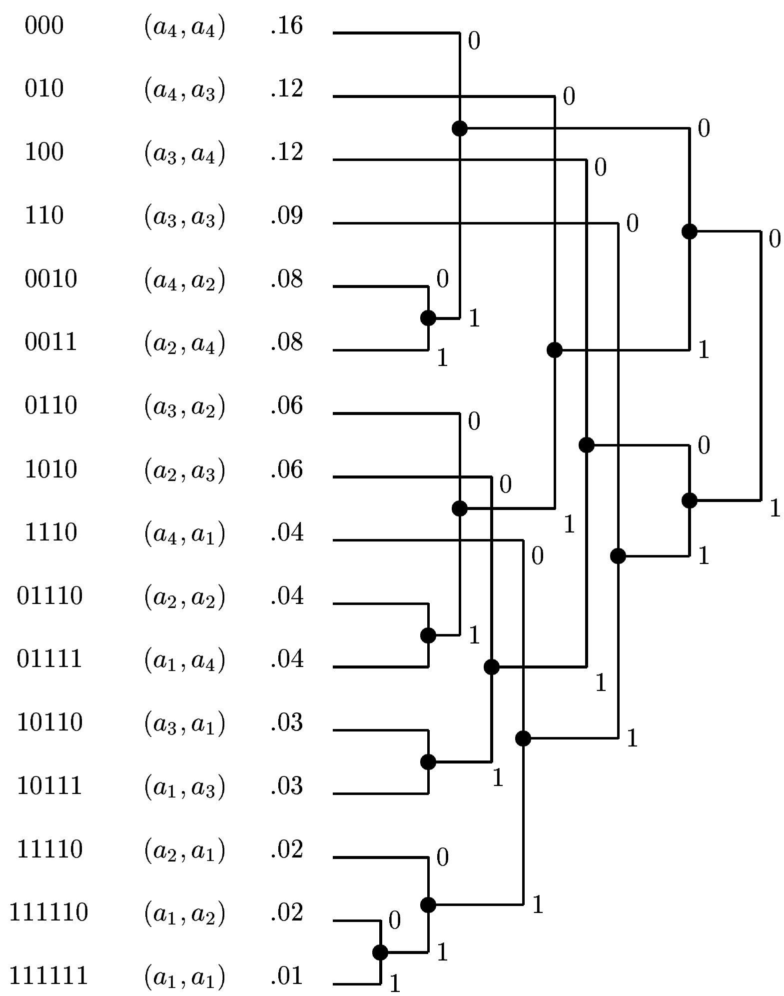


4) For the second extension of the source the alphabet of the source becomes $\mathcal{A}^2 = \{(a_1, a_1), (a_1, a_2), \ldots (a_4, a_4)\}$ and the probability of each pair is the product of the probabilities of each component, i.e. $p((a_1, a_2)) = .2$. A Huffman code for this source is depicted in the next figure. The average codeword length in bits per pair of source output is

$$\bar{R}_2(X) = 3 \times .49 + 4 \times .32 + 5 \times .16 + 6 \times .03 = 3.7300$$

The aberage codeword length in bits per each source output is $\bar{R}_1(X) = \bar{R}_2(X)/2 = 1.865$.

5) Huffman coding of the original source requires 1.9 bits per source output letter whereas Huffman coding of the second extension of the source requires 1.865 bits per source output letter and thus it is more efficient.



Please let me know if you find any mistake in this solution
