Solutions to Set #2
Data Compression, Huffman code and AEP

1. **Huffman coding.**
Consider the random variable

$$X = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ 0.50 & 0.26 & 0.11 & 0.04 & 0.04 & 0.03 & 0.02 \end{pmatrix}$$

(a) Find a binary Huffman code for X.

(b) Find the expected codelength for this encoding.

(c) Extend the Binary Huffman method to Ternarry (Alphabet of 3) and apply it for X.

Solution: Huffman coding.

(a) The Huffman tree for this distribution is

Codeword								
1	x_1	0.50	0.50	0.50	0.50	0.50	0.50	1
01	x_2	0.26	0.26	0.26	0.26	0.26	0.50	
001	x_3	0.11	0.11	0.11	0.11	0.24		
00011	x_4	0.04	0.04	0.08	0.13			
00010	x_5	0.04	0.04	0.05				
00001	x_6	0.03	0.05					
00000	x_7	0.02						

(b) The expected length of the codewords for the binary Huffman code is 2 bits. ($H(X) = 1.99$ bits)

(c) The ternary Huffman tree is

Codeword					
0	x_1	0.50	0.50	0.50	1.0
1	x_2	0.26	0.26	0.26	
20	x_3	0.11	0.11	0.24	
21	x_4	0.04	0.04		
222	x_5	0.04	0.09		
221	x_6	0.03			
220	x_7	0.02			

This code has an expected length 1.33 ternary symbols. ($H_3(X) =$ 1.25 ternary symbols).

2. **Codes.**
Let $X_1, X_2, \ldots,$ i.i.d. with

$$X = \begin{cases} 1, & \text{with probability } 1/2 \\ 2, & \text{with probability } 1/4 \\ 3, & \text{with probability } 1/4. \end{cases}$$

Consider the code assignment

$$C(x) = \begin{cases} 0, & \text{if } x = 1 \\ 01, & \text{if } x = 2 \\ 11, & \text{if } x = 3. \end{cases}$$

(a) Is this code nonsingular?

(b) Uniquely decodable?

(c) Instantaneous?

(d) Entropy Rate is defined as

$$H(\mathcal{X}) \triangleq \lim_{n\to\infty} \frac{H(X^n)}{n}. \tag{1}$$

What is the entropy rate of the process

$$Z_1 Z_2 Z_3 \ldots = C(X_1)C(X_2)C(X_3)\ldots?$$

Solution: Codes.

(a) **Yes**, this code is nonsingular because $C(x)$ is different for every x.

(b) **Yes**, this code is uniquely decodable. Reversing the codewords

$$C'(x) = \begin{cases} 0, & \text{if } x = 1 \\ 10, & \text{if } x = 2 \\ 11, & \text{if } x = 3 \end{cases}$$

gives an instantaneous code, and thus a uniquely decodable code. Therefore the reversed extension is uniquely decodable, and so the extension itself is also uniquely decodable.

(c) **No**, this code is not instantaneous because $C(1)$ is a prefix of $C(2)$.

(d) The expected codeword length is

$$L(C(x)) = 0.5 \times 1 + 0.25 \times 2 + 0.25 \times 2 = \frac{3}{2}.$$

Further, the entropy rate of the i.i.d. X^n is

$$H(\mathcal{X}) = H(X) = H(.5, .25, .25) = \frac{3}{2}.$$

So the code is a uniquely decodable code with $L = H(\mathcal{X})$, and therefore the sequence is maximally compressed with $H(\mathcal{Z}) = 1$ bit. If $H(\mathcal{Z})$ were less than its maximum of 1 bit then the Z^n sequence could be further compressed to its entropy rate, and X^m could also be compressed further by blockcoding. However, this would result in $L_m < H(\mathcal{X})$ which contradicts theorem 5.4.2 of the text. So $H(\mathcal{Z}) = 1$ bit.

Note that the Z^n sequence is not i.i.d. $\sim \text{Br}(\frac{1}{2})$, even though $H(\mathcal{Z}) = 1$ bit. For example, $P\{Z_1 = 1\} = \frac{1}{4}$, and a sequence starting $10\ldots$ is not allowed. However, once $Z_i = 0$ for some i then Z_k is Bernoulli($\frac{1}{2}$) for $k > i$, so Z^n is asymptotically Bernoulli($\frac{1}{2}$) and gives the entropy rate of 1 bit.

3. **Huffman via MATLAB**

(a) Give a Huffman encoding into an alphabet of size D = 2 of the following probability mass function:

$$\left(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}\right)$$

(b) Assume you have a file of size 1,000 symbols where the symbols are distributed i.i.d. according to the pmf above. After applying the Huffman code, what would be the pmf of the compressed binary file and what would be the expected length?

(c) Generate a sequence (using MATLAB or any other software) of length $10,000$ symbols of X with i.i.d probability P_X. Assume the alphabet of X is $\mathcal{X} = (0, 1, \ldots, 6)$.

(d) What is the percentage of each symbol $(0, 1, \ldots, 6)$ in the sequence that was generated by MATLAB. Explain this result using the law of large numbers.

(e) Represent each symbol in $\mathcal{X}$ using the simple binary representation. Namely, $X = 0$ represent as $'000'$, $X = 1$ represent as $'001'$, $X = 2$ represent as $'010'$, $\ldots$, $X = 6$ represent as $'110'$.

(f) What is the length of the simple representation. What percentage of $'0'$ and $'1'$ do you have in this representation?

(g) Now, compress the $10,000$ symbols of X, into bits using Huffman code.

(h) What is the length of the compressed file? What is the percentage of $'0'$ and $'1'$ do you have in this representation?

(i) Explain the results.

Solution:

(a) The code is presented in Fig 1.

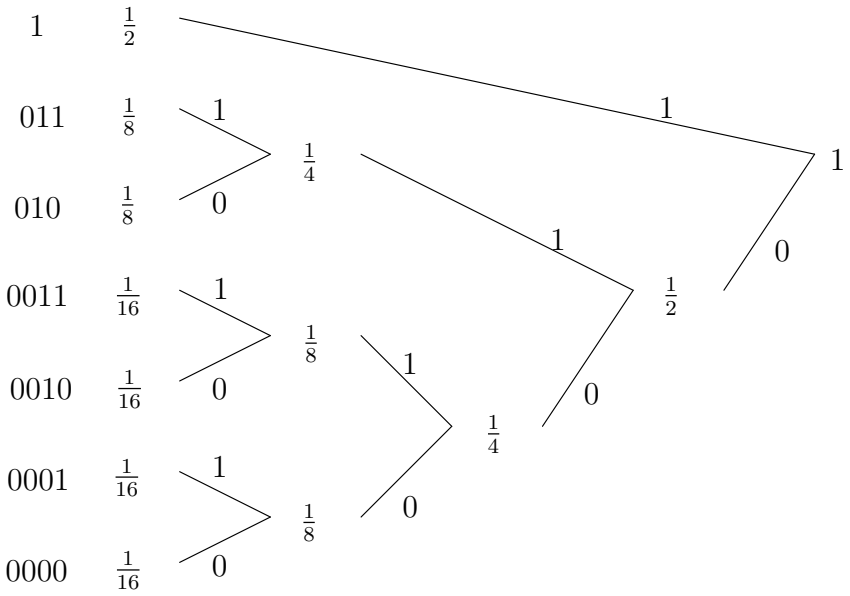


Figure 1: Huffman

(b) Huffman code is optimal code and achieves the entropy for dyadic distribution. If the distribution of the digits is not $Bernoulli(\frac{1}{2})$ you can compress it further. The binary digits of the data would be equally distributed after applying the Huffman code and therefore $p_0 = p_1 = \frac{1}{2}$.

The expected length would be:

$$E[l] = \frac{1}{2}\cdot 1 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3 + \frac{1}{16}\cdot 4 + \frac{1}{16}\cdot 4 + \frac{1}{16}\cdot 4 + \frac{1}{16}\cdot 4 = 2.25$$

Therefore, the expected length of 1000 symbols would be 2250 bits.

(c-i) The results from simulation should be as follows:

```
Drawn symbols
-------------
Size:                 10000
Percentage of x=0:    0.4984
Percentage of x=1:    0.125
Percentage of x=2:    0.1262
Percentage of x=3:    0.0636
Percentage of x=4:    0.0575
Percentage of x=5:    0.0647
Percentage of x=6:    0.0646

Before compression
------------------
Size:                 30000
Percentage of b=0:    0.77213
Percentage of b=1:    0.22787

After compression
-----------------
Size:                 22463
Percentage of b=0:    0.49958
Percentage of b=1:    0.50042

Compressions rate (symbol -> binary): 2.2463
```

Explanation: The symbols $X(i),\ i \in \{1, 2, \ldots, 10,000\}$ are drawn i.i.d. according to P_X. Therefore, by the L.L.N, the percentage of appearances of each symbol $x \in \mathcal{X}$, should be approximately $P_X(x)$. Next, we represent each symbol with 3 binary symbols. Since the file in uncompressed, the percentage of 0's is not half. The expected percentage can be calculated from the PMF. In our case,

$$\begin{aligned}\mathbb{E}\left[\#\{0\text{ś in file}\}\right] =& 10,000\big(P_x(0)\cdot 3 + P_x(1)\cdot 2 + P_x(2)\cdot 2 \\ &+ P_x(3)\cdot 1 + \cdots + P_x(6)\cdot 1\big) \\ =& 23,125\end{aligned}$$

Since there are $30,000$ bits for representation, the expected percentage is 77.08%. After Huffman's code is applied, the number of appearances of 0's and 1's are almost equal. This is since lossless compression maximize entropy (now 0 and 1 in the file are uniform).

4. **Entropy and source coding of a source with infinite alphabet** (15 points)

   Let X be an i.i.d. random variable with an infinite alphabet, $\mathcal{X} = \{1, 2, 3, ...\}$. In addition let $P(X = i) = 2^{-i}$.

   (a) What is the entropy of the random variable?

   (b) Find an optimal variable length code, and show that it is indeed optimal.

   Solution

   (a)

$$\begin{aligned}H(X) &= -\sum_{x\in\mathcal{X}} p(x)\log p(x) \\ &= -\sum_{i=1}^{\infty} 2^{-i}\log_2(2^{-i}) \\ &= -\sum_{i=1}^{\infty} \frac{-i}{2^i} = 2\end{aligned}$$

(b) Coding Scheme:

1	0
2	10
3	110
4	1110
5	11110
.	.
.	.
.	.

Average Length:

$$L^* = \sum_{i=1}^{\infty} p(x=i)L(i) = \sum_{i=1}^{\infty} \frac{i}{2^i} = 2 = H(X)$$

Hence it is the Optimal Code.

5. **Bad wine.**
One is given 6 bottles of wine. It is known that precisely one bottle has gone bad (tastes terrible). From inspection of the bottles it is determined that the probability p_i that the i^{th} bottle is bad is given by $(p_1, p_2, \ldots, p_6) = (\frac{7}{26}, \frac{5}{26}, \frac{4}{26}, \frac{4}{26}, \frac{3}{26}, \frac{3}{26})$. Tasting will determine the bad wine.

Suppose you taste the wines one at a time. Choose the order of tasting to minimize the expected number of tastings required to determine the bad bottle. Remember, if the first 5 wines pass the test you don't have to taste the last.

(a) What is the expected number of tastings required?

(b) Which bottle should be tasted first?

Now you get smart. For the first sample, you mix some of the wines in a fresh glass and sample the mixture. You proceed, mixing and tasting, stopping when the bad bottle has been determined.

(c) What is the minimum expected number of tastings required to determine the bad wine?

(d) What mixture should be tasted first?

Solution: Bad wine.

(a) If we taste one bottle at a time, the corresponding number of tastings are $\{1, 2, 3, 4, 5, 5\}$ with some order. By the same argument as in Lemma 5.8.1, to minimize the expected length $\sum p_i l_k$ we should have $l_j \leq l_k$ if $p_j > p_k$. Hence, the best order of tasting should be from the most likely wine to be bad to the least.

The expected number of tastings required is

$$\begin{aligned}\sum_{i=1}^{6} p_i l_i &= 1 \times \frac{7}{26} + 2 \times \frac{5}{26} + 3 \times \frac{4}{26} + 4 \times \frac{4}{26} + 5 \times \frac{3}{26} + 5 \times \frac{3}{26} \\ &= \frac{75}{26} \\ &= 2.88\end{aligned}$$

(b) The first bottle to be tasted should be the one with probability $\frac{7}{26}$.

(c) The idea is to use Huffman coding.

(01)	7	7	8	11	15	26
(11)	5	6	7	8	11	
(000)	4	5	6	7		
(001)	4	4	5			
(100)	3	4				
(101)	3					

The expected number of tastings required is

$$\begin{aligned}\sum_{i=1}^{6} p_i l_i &= 2 \times \frac{7}{26} + 2 \times \frac{5}{26} + 3 \times \frac{4}{26} + 3 \times \frac{4}{26} + 3 \times \frac{3}{26} + 3 \times \frac{3}{26} \\ &= \frac{66}{26} \\ &= 2.54\end{aligned}$$

Note that $H(p) = 2.52$ bits.

(d) The mixture of the first, third, and forth bottles should be tasted first, (or equivalently the mixture of the second, fifth and sixth).

6. **Relative entropy is cost of miscoding.**
Let the random variable X have five possible outcomes $\{1, 2, 3, 4, 5\}$. Consider two distributions on this random variable

Symbol	$p(x)$	$q(x)$	$C_1(x)$	$C_2(x)$
1	1/2	1/2	0	0
2	1/4	1/8	10	100
3	1/8	1/8	110	101
4	1/16	1/8	1110	110
5	1/16	1/8	1111	111

(a) Calculate $H(p)$, $H(q)$, $D(p||q)$ and $D(q||p)$.

(b) The last two columns above represent codes for the random variable. Verify that the average length of C_1 under p is equal to the entropy $H(p)$. Thus C_1 is optimal for p. Verify that C_2 is optimal for q.

(c) Now assume that we use code C_2 when the distribution is p. What is the average length of the codewords. By how much does it exceed the entropy $H(p)$?

(d) What is the loss if we use code C_1 when the distribution is q?

Solution: Relative entropy is cost of miscoding.

(a)

$$\begin{aligned} H(p) &= \sum_i -p_i \log p_i \\ &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - 2\cdot\frac{1}{16}\log\frac{1}{16} \\ &= \frac{15}{8}. \end{aligned}$$

Similarly, $H(q) = 2$.

$$\begin{aligned} D(p||q) &= \sum_i p_i \log\frac{p_i}{q_i} \\ &= \frac{1}{2}\log\frac{1/2}{1/2} + \frac{1}{4}\log\frac{1/4}{1/8} + \frac{1}{8}\log\frac{1/8}{1/8} + 2\cdot\frac{1}{16}\log\frac{1/16}{1/8} \\ &= \frac{1}{8}. \end{aligned}$$

Similarly, $D(q||p) = \frac{1}{8}$.

(b) The average codeword length for C_1 is

$$El_1 = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + 2 \cdot \frac{1}{16} \cdot 4 = \frac{15}{8}.$$

Similarly, the average codeword length for C_2 is 2.

(c)

$$E_p l_2 = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 3 + \frac{1}{8} \cdot 3 + 2 \cdot \frac{1}{16} \cdot 3 = 2,$$

which exceeds $H(p)$ by $D(p||q) = \frac{1}{8}$.

(d) Similarly, $E_q l_1 = \frac{17}{8}$, which exceeds $H(q)$ by $D(q||p) = \frac{1}{8}$.

7. **Shannon code.** Consider the following method for generating a code for a random variable X which takes on m values $\{1, 2, \ldots, m\}$ with probabilities $p_1, p_2, \ldots, p_m$. Assume that the probabilities are ordered so that $p_1 \geq p_2 \geq \cdots \geq p_m$. Define

$$F_i = \sum_{k=1}^{i-1} p_i, \tag{2}$$

the sum of the probabilities of all symbols less than i. Then the codeword for i is the number $F_i \in [0, 1]$ rounded off to l_i bits, where $l_i = \lceil \log \frac{1}{p_i} \rceil$.

(a) Show that the code constructed by this process is prefix-free and the average length satisfies

$$H(X) \leq L < H(X) + 1. \tag{3}$$

(b) Construct the code for the probability distribution $(0.5, 0.25, 0.125, 0.125)$.

Solution to Shannon code.

(a) Since $l_i = \lceil \log \frac{1}{p_i} \rceil$, we have

$$\log \frac{1}{p_i} \leq l_i < \log \frac{1}{p_i} + 1 \tag{4}$$

which implies that

$$H(X) \leq L = \sum p_i l_i < H(X) + 1. \tag{5}$$

The difficult part is to prove that the code is a prefix code. By the choice of l_i, we have

$$2^{-l_i} \leq p_i < 2^{-(l_i - 1)}. \tag{6}$$

Thus F_j, $j > i$ differs from F_i by at least 2^{-l_i}, and will therefore differ from F_i is at least one place in the first l_i bits of the binary expansion of F_i. Thus the codeword for F_j, $j > i$, which has length $l_j \geq l_i$, differs from the codeword for F_i at least once in the first l_i places. Thus no codeword is a prefix of any other codeword.

(b) We build the following table

Symbol	Probability	F_i in decimal	F_i in binary	l_i	Codeword
1	0.5	0.0	0.0	1	0
2	0.25	0.5	0.10	2	10
3	0.125	0.75	0.110	3	110
4	0.125	0.875	0.111	3	111

The Shannon code in this case achieves the entropy bound (1.75 bits) and is optimal.

8. **An AEP-like limit.** Let $X_1, X_2, \ldots$ be i.i.d. drawn according to probability mass function $p(x)$. Find

$$\lim_{n \to \infty} [p(X_1, X_2, \ldots, X_n)]^{\frac{1}{n}}.$$

Solution: An AEP-like limit.
$X_1, X_2, \ldots$, i.i.d. $\sim p(x)$. Hence $\log(X_i)$ are also i.i.d. and

$$\begin{aligned} \lim(p(X_1, X_2, \ldots, X_n))^{\frac{1}{n}} &= \lim 2^{\log(p(X_1, X_2, \ldots, X_n))^{\frac{1}{n}}} \\ &= 2^{\lim \frac{1}{n} \sum \log p(X_i)} \\ &= 2^{E(\log(p(X)))} \\ &= 2^{-H(X)} \end{aligned}$$

by the strong law of large numbers.

9. **AEP**. Let $X_1, X_2, \ldots$ be independent identically distributed random variables drawn according to the probability mass function $p(x), x \in \{1, 2, \ldots, m\}$. Thus $p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$. We know that $-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \to H(X)$ in probability. Let $q(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} q(x_i)$, where q is another probability mass function on $\{1, 2, \ldots, m\}$.

   (a) Evaluate $\lim -\frac{1}{n} \log q(X_1, X_2, \ldots, X_n)$, where $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$.

   (b) Now evaluate the limit of the log likelihood ratio $\frac{1}{n} \log \frac{q(X_1,\ldots,X_n)}{p(X_1,\ldots,X_n)}$ when $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$. Thus the odds favouring q are exponentially small when p is true.

Solution: AEP.

(a) Since the $X_1, X_2, \ldots, X_n$ are i.i.d., so are $q(X_1), q(X_2), \ldots, q(X_n)$, and hence we can apply the strong law of large numbers to obtain

$$\begin{aligned}
\lim -\frac{1}{n} \log q(X_1, X_2, \ldots, X_n) &= \lim -\frac{1}{n} \sum \log q(X_i) \\
&= -E(\log q(X)) \text{ w.p. } 1 \\
&= -\sum p(x) \log q(x) \\
&= \sum p(x) \log \frac{p(x)}{q(x)} - \sum p(x) \log p(x) \\
&= D(\mathbf{p}||\mathbf{q}) + H(\mathbf{p}).
\end{aligned}$$

(b) Again, by the strong law of large numbers,

$$\begin{aligned}
\lim -\frac{1}{n} \log \frac{q(X_1, X_2, \ldots, X_n)}{p(X_1, X_2, \ldots, X_n)} &= \lim -\frac{1}{n} \sum \log \frac{q(X_i)}{p(X_i)} \\
&= -E(\log \frac{q(X)}{p(X)}) \text{ w.p. } 1 \\
&= -\sum p(x) \log \frac{q(x)}{p(x)} \\
&= \sum p(x) \log \frac{p(x)}{q(x)} \\
&= D(\mathbf{p}||\mathbf{q}).
\end{aligned}$$

10. **Empirical distribution of a sequence** Before starting the question, below are two facts that you may consider to use:

    - Stirling approximation: $n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$.
    - Consider a sequence of length n that consist of two different numbers. The first number appears n_1 times and the second number appears n_2 times such that $n_1 + n_2 = n$. The number of different combinations of such sequences is given by $\binom{n}{n_1\ n_2} = \frac{n!}{n1!n2!}$.

    A fair dice with 6 faces was thrown n times, where n is a very large number.

    (a) Find how many different sequences there exists with an empirical pmf $(p_1, p_2, ..., p_6)$, where p_i is the portion of the sequence that is equal to $i \in \{1, 2, ..., 6\}$.
    In this section you can assume that $n! \approx (\frac{n}{e})^n$ since only the power of $\frac{n}{e}$ will matter.

    (b) Now, we were told that the portion of odd numbers in the sequence is 2/3 (i.e., $p_1 + p_3 + p_5 = 2/3$). For n very large, what is the most likely empirical pmf of the sequence. **Hint:** Define:

    $$X = \begin{cases} 1 & p_1 \\ 3 & p_3 \\ 5 & p_5 \end{cases}, \ Y = \begin{cases} 2 & p_2 \\ 4 & p_4 \\ 6 & p_6 \end{cases}, \ Z = \begin{cases} X & \frac{2}{3} \\ Y & \frac{1}{3} \end{cases}.$$

    Think why maximizing $H(Z)$ means maximizing $H(X)$, $H(Y)$.

    (c) What is the cardinality of the weak typical set with respect to the pmfs that you found/given in the previous subquestions, i.e., (a) and (b)?

    Remark 1 *The weak typical set is the typical set we learned in the AEP lecture.*

    Solution

(a) The number of combinations N is given by:

$$\begin{aligned} N &= \binom{n}{n_1 n_2 n_3 n_4 n_5 n_6} \\ &= \frac{n!}{n_1! n_2! n_3! n_4! n_5! n_6!} \\ &\overset{(a)}{=} \frac{\left(\frac{n}{e}\right)^n}{\left(\frac{n_1}{e}\right)^{n_1}\left(\frac{n_2}{e}\right)^{n_2}\left(\frac{n_3}{e}\right)^{n_3}\left(\frac{n_4}{e}\right)^{n_4}\left(\frac{n_5}{e}\right)^{n_5}\left(\frac{n_6}{e}\right)^{n_6}}, \end{aligned}$$

where (a) follows by Stirling's approximation. Thus,

$$\begin{aligned} \log(N) &= \log n - n_1 \log n_1 - n_2 \log n_2 - \ldots - n_6 \log n_6 \\ &= -\sum_{i=1}^{6} n_i \log\left(\frac{n_i}{n}\right), \end{aligned}$$

Finally, we get:

$$\begin{aligned} N &\approx 2^{-\sum_{i=1}^{6} n_i \log \frac{n_i}{n}} \\ &= 2^{-n\sum_{i=1}^{6} \frac{n_i}{n} \log \frac{n_i}{n}} \\ &= 2^{nH(\frac{n_1}{n}, \frac{n_2}{n}, \frac{n_3}{n}, \frac{n_4}{n}, \frac{n_5}{n}, \frac{n_6}{n})} \\ &\overset{(a)}{\approx} 2^{nH(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})} \\ &\overset{(b)}{=} 2^{6nH(\frac{1}{6})}, \end{aligned}$$

where:
(a) For sufficient large n we have $\frac{n_i}{n} \approx \frac{1}{6}$.
(b) i.i.d.
The result is probable since the typical set and the set defined by the law of large numbers converge as the number of samples goes to infinty.

(b) First, observe that:

$$\frac{n_1 + n_3 + n_5}{n} = \frac{2}{3} \Leftrightarrow p_1 + p_2 + p_3 = \frac{2}{3}$$

We would like to find $\{p_1, p_2, p_3, p_4, p_5, p_6\}$ which, under the given constraints maximizes entropy. This results in the biggest typical

set which means the event will be most likely. Equivalently:

$$X = \begin{cases} 1 & p_1 \\ 3 & p_3 \\ 5 & p_5 \end{cases}, \; Y = \begin{cases} 2 & p_2 \\ 4 & p_4 \\ 6 & p_6 \end{cases}, \; Z = \begin{cases} X & \frac{2}{3} \\ Y & \frac{1}{3} \end{cases}$$

Define the following indicator auxiliary variable:

$$I = \begin{cases} 1 & Z = X \\ 0 & Z = Y \end{cases}$$

Now, let us maximize the entropy of Z.

$$\begin{aligned} H(Z) &\overset{(a)}{=} H(Z, I) \\ &= H(I) + H(Z|I) \\ &= H(\frac{2}{3}) + \frac{1}{3}H(Y) + \frac{2}{3}H(X), \end{aligned}$$

where (a) follows since

$$\begin{aligned} H(Z, I) &= H(Z, I) \\ &= H(Z) + \underbrace{H(I|Z)}_{Deterministic\ Given\ Z} \\ &= H(Z). \end{aligned}$$

Accordingly, maximizing $H(Z)$ means maximizing $H(X),\ H(Y)$. As we've shown at class uniform distribution maximizes entropy. Thus,

$$H(X),\ H(Y) \sim Uniform \Rightarrow p_1,\ p_3,\ p_5 = \frac{2}{9},\ p_2,\ p_4,\ p_6 = \frac{1}{9}.$$

11. **drawing a codebook** Let X_i be a r.v. i.i.d distributed according to $P(x)$. We draw codebook of 2^{nR} codewords of X^n independently using $P(x)$ and i.i.d.. We would like to answer the question: what is the probability that the first codeword would be identical to another codeword in the codebook as n goes to infinity.

    - Let x^n be a sequence in the typical set $A_\epsilon^n(X)$. What is the asymptotic probability (you may provide an upper and lower bound) as $n \to \infty$ that we draw a sequence X^n i.i.d distributed according to $P(x)$ and we get x^n.

- Using your answer from the previous sub-question find an $\overline{\alpha}$ such that if $R < \overline{\alpha}$ the probability that the first codewaord in the codebook appears twice or more in the codebook goes to zero as $n \to \infty$.
- Find an $\underline{\alpha}$ such that if $R > \underline{\alpha}$ the probability that the first codewaord in the codebook appears twice or more in the codebook goes to 1 as $n \to \infty$.

Solution: drawing a codebook.

(a) By the definition of the typical set, the probability of every $x^n \in A_\epsilon^{(n)}$ to be drawn is bounded by

$$2^{-n(H(X)+\epsilon)} \leq p(x^n) \leq 2^{-n(H(X)-\epsilon)}$$

notice that the question regards to the probability of a specific sequence $x^n \in A_\epsilon^{(n)}$ and not the probability of the whole set (which is almost 1).

(b) We want to find $\bar{\alpha}$ such that if $R < \bar{\alpha}$ then $\Pr(\exists i \neq 1 : x^n(i) = x^n(1))$ goes to zero. Consider the following derivations:

$$\begin{aligned}
\Pr(\exists i \neq 1 : x^n(i) = x^n(1)) &= \Pr(\bigcup_{i=2}^{2^{nR}} \{x^n(i) = x^n(1)\}) \\
&\overset{(a)}{\leq} \sum_{i=2}^{2^{nR}} \Pr(x^n(i) = x^n(1)) \\
&\overset{(b)}{\leq} \sum_{i=2}^{2^{nR}} 2^{-n(H(X)-\epsilon)} \\
&\leq 2^{nR} 2^{-n(H(X)-\epsilon)} \\
&= 2^{nR-(H(X)-\epsilon)}
\end{aligned}$$

where (a) follows from the union bound and (b) follows from section a. If we set $\bar{\alpha} = H(X) - \epsilon$ this probability goes to zero.

(c) We want to find $\underline{\alpha}$ such that if $R > \underline{\alpha}$ then $P(A) = 1 - \Pr(\forall i \neq$

$1 : x^n(i) \neq x^n(1))$ goes to 1. Consider the following derivations:

$$\begin{aligned}
\Pr(\forall i \neq 1 : x^n(i) \neq x^n(1)) &= \prod_{i=2}^{2^{nR}} \Pr(x^n(i) \neq x^n(1)) \\
&\leq \prod_{i=2}^{2^{nR}} (1 - \Pr(x^n(i) = x^n(1))) \\
&\overset{(a)}{\leq} \prod_{i=2}^{2^{nR}} (1 - 2^{-n(H(X)+\epsilon)}) \\
&\overset{(b)}{\leq} (1 - 2^{-n(H(X)+\epsilon)})^{2^{nR}} \\
&\overset{(c)}{\leq} e^{-2^{nR}2^{-n(H(X)+\epsilon)}} \\
&\leq e^{-2^{nR-(H(X)+\epsilon)}}
\end{aligned}$$

where (a) follows from section a, (b) follows by adding $i = 1$ to the product and since the drawing is i.i.d, and (c) follows since $(1 - x)^n \leq e^{-nx}$. If we set $R > H(X) + \epsilon$ this probability goes to zero and thus $P(A)$ goes to 1. Hence, $\underline{\alpha} = H(X) + \epsilon$.

12. **Saving the princess.** A princess was abducted and was put in one of K rooms. Each room is labeled by a number $1, 2, \ldots, K$. Each room is of size s_i where $i = 1, 2, \ldots, K$. The probability of the princess to be in room i is p_i, and proportional to the size of the room s_i, namely, $p_i = \alpha s_i$ where α is a constant.

    (a) Find α.

    (b) In order to save the princess you need to find in which room she is. You may ask the demon a yes/no question. Like is she in room number 1 or is she in room 2 or 5 or is she in a room of odd number, and so on. You will save the princess if and only if the expected number of questions is the minimum possible. What would be the questions you should ask the demon in order to save the princess?

    Solution: Saving the princess.

(a) Since p is a probability mass function, it must satisfy

$$1 = \sum_{i=1}^{K} p_i = \sum_{i=1}^{K} \alpha s_i = \alpha \sum_{i=1}^{K} s_i$$

Therefore, $\alpha = \frac{1}{\sum_{i=1}^{K} s_i}$.

(b) The idea is to use Huffman code with the probability that the princess should be in a specific room. Follow the solution of *Bad wine* question. Here, we will build questions like *is the princess in room 1 or 2* instead of mixing wine, and we want to find the princess instead of bad wine bottle.

13. **Lossless source coding with side information.**
Consider the lossless source coding with side information that is available at the encoder and decoder, where the source X and the side information Y are i.i.d. $\sim P_{X,Y}(x, y)$.

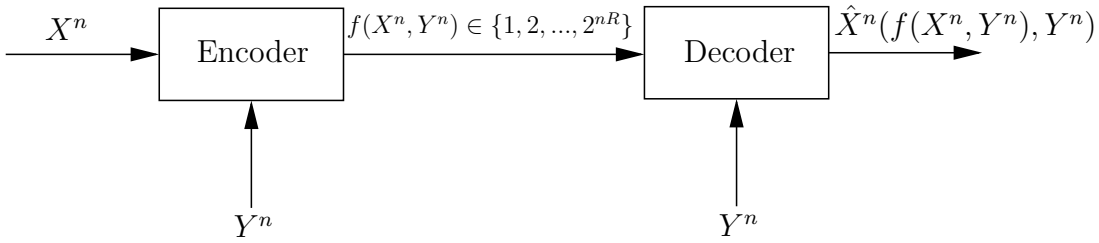


Figure 2: Lossless source coding with side information at the encoder and decoder.

Show that a code with rate $R < H(X|Y)$ can not be achievable, and interpret the result.

Hint: Let $T \triangleq f(X^n, Y^n)$. Consider

$$\begin{aligned} nR &\geq H(T) \\ &\geq H(T|Y^n), \end{aligned} \tag{7}$$

and use similar steps, including Fano's inequality, as we used in the class to prove the converse where side information was not available.

Solution Sketch of the solution (please fill in the explanation for each step):

$$\begin{aligned} nR &\geq H(T) \\ &\geq H(T|Y^n), \\ &\geq I(X^n;T|Y^n) \\ &= H(X^n|Y^n) - H(X^n|T,Y^n) \\ &= nH(X|Y) - \epsilon_n, \end{aligned}$$

where $\epsilon_n \to 0$.

14. **Challenge: Optimal code for an infinite alphabet** This question is a challenge.

15. **Conditional Information Divergence**

    (a) Let X, Z be random variables jointly distributed according to $P_{X,Z}$. We define the conditional informational divergence as follows:

    $$D\big(P_{X|Z}\big|\big|Q_{X|Z}\big|P_Z\big) = \sum_{(x,z)\in\mathcal{X}\times\mathcal{Z}} P_{X,Z}(x,z)\log\left(\frac{P_{X|Z}(x|z)}{Q_{X|Z}(x|z)}\right).$$

    With respect to this definition, prove for each relation if it is **true** or **false**:

    For any pair of random variables A, B that are jointly distributed according to $P_{A,B}$,

    i.
    $$D\big(P_{A,B}\big|\big|Q_{A,B}\big) = D\big(P_A\big|\big|Q_A\big) + D\big(P_{B|A}\big|\big|Q_{B|A}\big|P_A\big).$$

    ii.
    $$D\big(P_{A,B}\big|\big|P_AP_B\big) = D\big(P_{B|A}\big|\big|P_B\big|P_A\big).$$

    iii.
    $$I(A;B) = D\big(P_{B|A}\big|\big|P_B\big|P_A\big).$$

    iv.
    $$D\big(P_{A|B}\big|\big|Q_{A|B}\big|P_B\big) = \sum_{b\in\mathcal{B}} P_B(b)D\big(P_{A|B=b}\big|\big|Q_{A|B=b}\big).$$

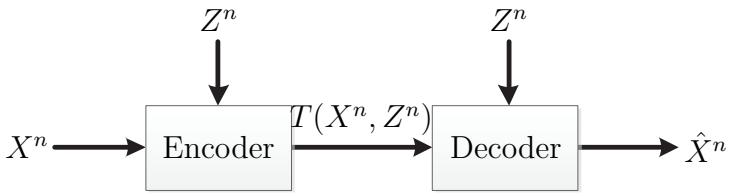


Figure 3: Source coding with side information.

(b) Consider the setting in Fig. 3.

We would like to compress the source sequence X^n losslessly using a prefix code with side information Z^n which is available to the encoder and the decoder. The sources (X^n, Z^n) are distributed i.i.d. according to $P_{X,Z}$ and that all the distribution and conditional distributions are dyadic (i.e., P_X is dyadic if $P_X(x) = 2^{-i}$, for some i, for all $x \in \mathcal{X}$). We denote the average number of bits per symbol needed to compress the source X^n as L.

i. What is the minimal L?

ii. Although the distribution of (X^n, Z^n) is $P_{X,Z}$, the distribution that is used design the optimal prefix code is $Q_{X|Z}P_Z$. What is the actual L (average bits per symbol) of this code?

iii. Now, the distribution that is used to design the prefix code is $Q_{X,Z}$. What is the actual L now?

Solutions See question 4 in Final Exam 2014 Moed A (with solutions).

16. **True or False of a constrained inequality**:

Given are three discrete random variables X, Y, Z that satisfy $H(Y|X, Z) = 0$.

(a) Copy the next relation and write **true** or **false** (If true, prove the statement, and if not provide a counterexample).

$$I(X;Y) \geq H(Y) - H(Z)$$

(b) What are the conditions for which the equality $I(X;Y) = H(Y) - H(Z)$ holds.

(c) Assume that the conditions for $I(X;Y) = H(Y) - H(Z)$ are satisfied. Is it true that there exists a function such that $Z = g(Y)$?

Solution

(a) First, note that $I(X;Y) = H(Y) - H(Y|X)$. Consider

$$\begin{aligned} 0 =& H(Y|X,Z) \\ =& H(X,Y,Z) - H(X,Z) \\ =& H(X) + H(Y|X) + H(Z|X,Y) - H(X) - H(Z|X) \\ =& H(Y|X) - I(Z;Y|X) \end{aligned}$$

Therefore, we can conclude that

$$H(Y|X) = I(Z;Y|X) \leq H(Z|X) \leq H(Z). \tag{8}$$

It follows that

$$H(Y) - H(Y|X) \geq H(Y) - H(Z).$$

(b) To satisfy equality, we must satisfy equalities in (8). First equality is obtained iff $H(Z|X,Y) = 0$, implies that Z is a function of X and Y. The second if $H(Z|X) = H(Z)$, which implies that Z is a function of X.

(c) False. Consider the following counter example. Let X and Z be independent, each distributed according to Bernoulli($\frac{1}{2}$) and $Y = X \bigoplus Z$. Then, $Y = f(X,Z)$, $H(Z|X) = H(Z)$ and $Z = X \bigoplus Y$ so $H(Z|X,Y) = 0$. However, Z is not a function of Y.

17. **True or False**: Copy each relation and write **true** or **false**.

(a) Let $X - Y - Z - W$ be a Markov chain, then the following holds:

$$I(X;W) \leq I(Y;Z).$$

(b) For two probability distributions, p_{XY} and q_{XY}, that are defined on $\mathcal{X} \times \mathcal{Y}$, the following holds:

$$D(p_{XY}||q_{XY}) \geq D(p_X||q_X).$$

(c) If X and Y are dependent and also Y and Z are dependent, then X and Z are dependent.

Solution

(a) True. By data processing inequality, if $A - B - C$ form a Markov chain, then $I(A;C) \leq I(A;B)$. Here, we have $X - Y - Z$ and $X - Z - W$ as Markov chains.

$$\begin{aligned} I(X;W) \leq & I(X;Z) \\ \leq & I(Y;Z) \end{aligned}$$

(b) True. Consider the definition of a conditional divergence,

$$D\big(P_{X|Z}\big|\big|Q_{X|Z}\big|P_Z\big) = \sum_{(x,z)\in\mathcal{X}\times\mathcal{Z}} P_{X,Z}(x,z)\log\left(\frac{P_{X|Z}(x|z)}{Q_{X|Z}(x|z)}\right).$$

From previous question on the conditional divergence, we learned that

$$D\big(P_{X,Y}\big|\big|Q_{X,Y}\big) = D\big(P_X\big|\big|Q_Y\big) + D\big(P_{Y|X}\big|\big|Q_{Y|X}\big|P_X\big),$$

where

$$D\big(P_{Y|X}\big|\big|Q_{Y|X}\big|P_X\big) = \sum_{x\in\mathcal{X}} P_X(x) D\big(P_{Y|X=x}\big|\big|Q_{Y|X=x}\big),$$

which is non-negative. We conclude that

$$D(p_{XY}||q_{XY}) \geq D(p_X||q_X).$$

(c) False. Here is a counterexample. Let X and W be two *independent* random variables. Let $Y = X + W$ and $Z = W$. Then, Y and X are dependent, Z and Y are dependent, and Z is independent of X.

18. **Huffman Code** : Let X^n be a an i.i.d. source that is distributed according to p_X:

x	0	1	2	3
$p_X(x)$	0.5	0.25	0.125	0.125

(a) Find $H(X)$.

(b) Build a binary Huffman code for the source X.

(c) What is the expected length of the resulting compressed sequence.

(d) What is the expected number of zeros in the resulting compressed sequence.

(e) Let $\tilde{X}^n$ be an another source distributed i.i.d. according to $p_{\tilde{X}}$.

$\tilde{x}$	0	1	2	3
$p_{\tilde{X}}(\tilde{x})$	0.3	0.4	0.1	0.2

What is the expected length of compressing the source $\tilde{X}$ using the code constructed in (b).

(f) Answer (d) for the code constructed in (b) and the source $\tilde{X}^n$.

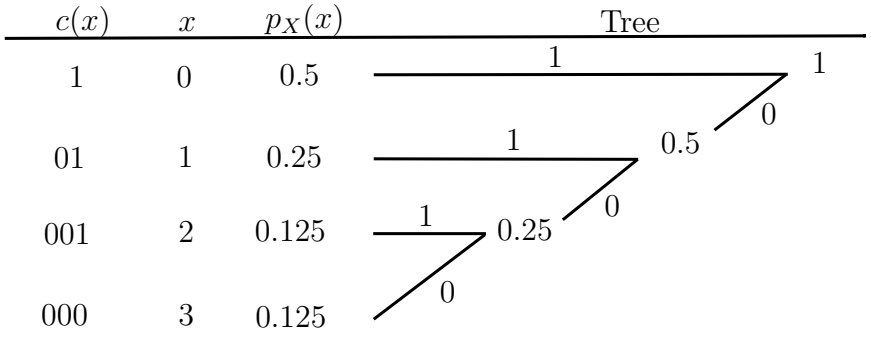
(g) Is the relative portion of zeros (the quantity in (d) divided by the quantity in (c)) after compressing the source X^n and the source $\tilde{X}^n$ different? For both sources, explain why there is or there is not a difference.

Solution

(a) The entropy of X is

$$\begin{aligned} H(X) =& -0.5 \cdot \log 0.5 - 0.25 \cdot \log 0.25 - 2 \cdot 0.125 \cdot \log 0.125 \\ =& 1.75 \end{aligned}$$

(b)



(c) Denote the length of a codeword by $L(c(x_i))$. Then

$$\begin{aligned}
L(c^n(x^n)) =& \sum_{i=1}^{n} L(c(x_i)) \\
=& \sum_{i=1}^{n} \left[p_X(0) \cdot L(c(0)) + p_X(1) \cdot L(c(1)) + p_X(2) \cdot L(c(2)) + p_X(3) \cdot L(c(3))\right] \\
=& n\left(0.5 \cdot 1 + 0.25 \cdot 2 + 0.125 \cdot 3 + 0.125 \cdot 3\right) \\
=& 1.75n
\end{aligned}$$

The expected length of the sequence is $nR^* = 1.75n$. Note that the distribution on X is dyadic, and therefore the Huffman code is optimal. Therefore, $nR = nH(X)$.

(d) Let $N(0|c)$ denote the number of zeros in a codeword c, and $c^n(x^n) = [c(x_1), \ldots, c(x_n)]$.

$$\begin{aligned}
\mathbb{E}\left[N(0|c^n(X^n))\right] =& \mathbb{E}\left[\sum_{i=1}^{n} N(0|c(X_i))\right] \\
=& \sum_{i=1}^{n} \mathbb{E}\left[N(0|c(X_i))\right] \\
=& \sum_{i=1}^{n} [p_X(0) \cdot N(0|c(0)) + p_X(1) \cdot N(0|c(1)) + p_X(2) \cdot N(0|c(2)) \\
& + p_X(3) \cdot N(0|c(3))] \\
=& \sum_{i=1}^{n} \left[0.5 \cdot 0 + 0.25 \cdot 1 + 0.125 \cdot 2 + 0.125 \cdot 3\right] \\
=& 0.875n
\end{aligned}$$

Since the code is optimal, the number of zeros is half of the expected length (see the following sub-question).

(e) Denote the length of a codeword by $L(c(x_i))$. Then

$$\begin{aligned}
L(c^n(x^n)) =& \sum_{i=1}^{n} L(c(x_i)) \\
=& \sum_{i=1}^{n} \left[p_X(0) \cdot L(c(0)) + p_X(1) \cdot L(c(1)) + p_X(2) \cdot L(c(2)) + p_X(3) \cdot L(c(3))\right] \\
=& n\left(0.3 \cdot 1 + 0.4 \cdot 2 + 0.1 \cdot 3 + 0.2 \cdot 3\right) \\
=& 2n
\end{aligned}$$

(f) The expected number of zeros is

$$\begin{aligned}
\mathbb{E}\left[N(0|c^n(\tilde{X}^n))\right] =& \mathbb{E}\left[\sum_{i=1}^{n} N(0|c(\tilde{X}_i))\right] \\
=& \sum_{i=1}^{n} \mathbb{E}\left[N(0|c(\tilde{X}_i))\right] \\
=& \sum_{i=1}^{n} [p_{\tilde{X}}(0) \cdot N(0|c(0)) + p_{\tilde{X}}(1) \cdot N(0|c(1)) + p_{\tilde{X}}(2) \cdot N(0|c(2)) \\
& + p_{\tilde{X}}(3) \cdot N(0|c(3))] \\
=& \sum_{i=1}^{n} \left[0.3 \cdot 0 + 0.4 \cdot 1 + 0.1 \cdot 2 + 0.2 \cdot 3\right] \\
=& 1.2n
\end{aligned}$$

Note that the expected number of zeros is not half of the expected length. It implies that the code is not optimal.

$$R^* = H(\tilde{X}) = 1.846$$

(g) For X^n we used optimal code with varying length. Therefore, the expected number of zeros is half of the compressed sequence. However, ee used a code that is not optimal for $\tilde{X}^n$. Henceforth, the compression rate is not optimal, and the expected number of zeros is not necessarily half of the expected length. Note that the expected length is not optimal too, since $H(\tilde{X}) \approxeq 1.8464$, which is not equal to $\frac{\mathbb{E}\left[L(c(\tilde{X}^n))\right]}{n}$.