Homework 3 Solutions

Chapter 4

3) *Shuffles increase entropy.*

$$\begin{aligned} H(TX) &\geq H(TX|T) && (218)\\ &= H(T^{-1}TX|T) && (219)\\ &= H(X|T) && (220)\\ &= H(X). && (221) \end{aligned}$$

The inequality follows from the fact that conditioning reduces entropy and the first equality follows from the fact that given T, we can reverse the shuffle.

6) *Monotonicity of entropy per element.*

a) By the chain rule for entropy,

$$\begin{aligned} \frac{H(X_1, X_2, \ldots, X_n)}{n} &= \frac{\sum_{i=1}^n H(X_i|X^{i-1})}{n} && (246)\\ &= \frac{H(X_n|X^{n-1}) + \sum_{i=1}^{n-1} H(X_i|X^{i-1})}{n} && (247)\\ &= \frac{H(X_n|X^{n-1}) + H(X_1, X_2, \ldots, X_{n-1})}{n}. && (248) \end{aligned}$$

From stationarity it follows that for all $1 \leq i \leq n$,

$$H(X_n|X^{n-1}) \leq H(X_i|X^{i-1}),$$

which further implies, by averaging both sides, that,

$$\begin{aligned} H(X_n|X^{n-1}) &\leq \frac{\sum_{i=1}^{n-1} H(X_i|X^{i-1})}{n-1} && (249)\\ &= \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}. && (250) \end{aligned}$$

Combining (248) and (250) yields,

$$\begin{aligned} \frac{H(X_1, X_2, \ldots, X_n)}{n} &\leq \frac{1}{n}\left[\frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1} + H(X_1, X_2, \ldots, X_{n-1})\right]\\ &= \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}. && (251) \end{aligned}$$

b) By stationarity we have for all $1 \leq i \leq n$,

$$H(X_n|X^{n-1}) \leq H(X_i|X^{i-1}),$$

which implies that

$$\begin{aligned} H(X_n|X^{n-1}) &= \frac{\sum_{i=1}^n H(X_n|X^{n-1})}{n} && (252)\\ &\leq \frac{\sum_{i=1}^n H(X_i|X^{i-1})}{n} && (253)\\ &= \frac{H(X_1, X_2, \ldots, X_n)}{n}. && (254) \end{aligned}$$

11) *Stationary processes.*

a) $H(X_n|X_0) = H(X_{-n}|X_0)$.
This statement is true, since

$$\begin{aligned} H(X_n|X_0) &= H(X_n, X_0) - H(X_0) \qquad (269)\\ H(X_{-n}|X_0) &= H(X_{-n}, X_0) - H(X_0) \qquad (270) \end{aligned}$$

and $H(X_n, X_0) = H(X_{-n}, X_0)$ by stationarity.

b) $H(X_n|X_0) \geq H(X_{n-1}|X_0)$.
This statement is not true in general, though it is true for first order Markov chains. A simple counterexample is a periodic process with period n. Let $X_0, X_1, X_2, \ldots, X_{n-1}$ be i.i.d. uniformly distributed binary random variables and let $X_k = X_{k-n}$ for $k \geq n$. In this case, $H(X_n|X_0) = 0$ and $H(X_{n-1}|X_0) = 1$, contradicting the statement $H(X_n|X_0) \geq H(X_{n-1}|X_0)$.

c) $H(X_n|X_1^{n-1}, X_{n+1})$ is non-increasing in n.
This statement is true, since by stationarity $H(X_n|X_1^{n-1}, X_{n+1}) = H(X_{n+1}|X_2^n, X_{n+2}) \geq H(X_{n+1}|X_1^n, X_{n+2})$ where the inequality follows from the fact that conditioning reduces entropy.

Chapter 5

27) *Test for unique decodability.*

The proof of the Sardinas-Patterson test has two parts. In the first part, we will show that if there is a code string that has two different interpretations, then the code will fail the test. The simplest case is when the concatenation of two codewords yields another codeword. In this case, S_2 will contain a codeword, and hence the test will fail.

In general, the code is not uniquely decodeable, iff there exists a string that admits two different parsings into codewords, e.g.

$$x_1x_2x_3x_4x_5x_6x_7x_8 = x_1x_2, x_3x_4x_5, x_6x_7x_8 = x_1x_2x_3x_4, x_5x_6x_7x_8. \tag{414}$$

In this case, S_2 will contain the string x_3x_4, S_3 will contain x_5, S_4 will contain $x_6x_7x_8$, which is a codeword. It is easy to see that this procedure will work for any string that has two different parsings into codewords; a formal proof is slightly more difficult and using induction.

In the second part, we will show that if there is a codeword in one of the sets $S_i, i \geq 2$, then there exists a string with two different possible interpretations, thus showing that the code is not uniquely decodeable. To do this, we essentially reverse the construction of the sets. We will not go into the details - the reader is referred to the original paper.

a) Let S_1 be the original set of codewords. We construct S_{i+1} from S_i as follows: A string y is in S_{i+1} iff there is a codeword x in S_1, such that xy is in S_i or if there exists a $z \in S_i$ such that zy is in S_1 (i.e., is a codeword). Then the code is uniquely decodable iff none of the S_i, $i \geq 2$ contains a codeword. Thus the set $S = \cup_{i \geq 2} S_i$.

b) A simple upper bound can be obtained from the fact that all strings in the sets S_i have length less than l_{max}, and therefore the maximum number of elements in S is less than $2^{l_{max}}$.

c) i) $\{0, 10, 11\}$. This code is instantaneous and hence uniquely decodable.

ii) $\{0, 01, 11\}$. This code is a suffix code (see problem 11). It is therefore uniquely decodable. The sets in the Sardinas-Patterson test are $S_1 = \{0, 01, 11\}$, $S_2 = \{1\} = S_3 = S_4 = \ldots$.

iii) $\{0, 01, 10\}$. This code is not uniquely decodable. The sets in the test are $S_1 = \{0, 01, 10\}$, $S_2 = \{1\}$, $S_3 = \{0\}, \ldots$. Since 0 is codeword, this code fails the test. It is easy to see otherwise that the code is not UD - the string 010 has two valid parsings.

iv) $\{0, 01\}$. This code is a suffix code and is therefore UD. THe test produces sets $S_1 = \{0, 01\}$, $S_2 = \{1\}$, $S_3 = \phi$.

v) $\{00, 01, 10, 11\}$. This code is instantaneous and therefore UD.

vi) $\{110, 11, 10\}$. This code is uniquely decodable, by the Sardinas-Patterson test, since $S_1 = \{110, 11, 10\}$, $S_2 = \{0\}$, $S_3 = \phi$.

vii) $\{110, 11, 100, 00, 10\}$. This code is UD, because by the Sardinas Patterson test, $S_1 = \{110, 11, 100, 00, 10\}$, $S_2 = \{0\}$, $S_3 = \{0\}$, etc.

d) We can produce infinite strings which can be decoded in two ways only for examples where the Sardinas Patterson test produces a repeating set. For example, in part (ii), the string 011111... could be parsed either as 0,11,11,... or as 01,11,11,.... Similarly for (viii), the string 10000... could be parsed as 100,00,00,... or as 10,00,00,.... For the instantaneous codes, it is not possible to construct such a string, since we can decode as soon as we see a codeword string, and there is no way that we would need to wait to decode.

30) *Cost of miscoding*

a) $H(p) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{8}\log 8 + \frac{1}{16}\log 16 + \frac{1}{16}\log 16 = 1.875$ bits.
$H(q) = \frac{1}{2}\log 2 + \frac{1}{8}\log 8 + \frac{1}{8}\log 8 + \frac{1}{8}\log 8 + \frac{1}{8}\log 8 = 2$ bits.
$D(p||q) = \frac{1}{2}\log\frac{1/2}{1/2} + \frac{1}{4}\log\frac{1/4}{1/8} + \frac{1}{8}\log\frac{1/8}{1/8} + \frac{1}{16}\log\frac{1/16}{1/8} + \frac{1}{16}\log\frac{1/16}{1/8} = 0.125$ bits.
$D(p||q) = \frac{1}{2}\log\frac{1/2}{1/2} + \frac{1}{8}\log\frac{1/8}{1/4} + \frac{1}{8}\log\frac{1/8}{1/8} + \frac{1}{8}\log\frac{1/8}{1/16} + \frac{1}{8}\log\frac{1/8}{1/16} = 0.125$ bits.

b) The average length of C_1 for $p(x)$ is 1.875 bits, which is the entropy of p. Thus C_1 is an efficient code for $p(x)$. Similarly, the average length of code C_2 under $q(x)$ is 2 bits, which is the entropy of q. Thus C_2 is an efficient code for q.

c) If we use code C_2 for $p(x)$, then the average length is $\frac{1}{2} * 1 + \frac{1}{4} * 3 + \frac{1}{8} * 3 + \frac{1}{16} * 3 + \frac{1}{16} * 3 = 2$ bits. It exceeds the entropy by 0.125 bits, which is the same as $D(p||q)$.

d) Similary, using code C_1 for q has an average length of 2.125 bits, which exceeds the entropy of q by 0.125 bits, which is $D(q||p)$.

32) *Bad Wine*

a) If we taste one bottle at a time, to minimize the expected number of tastings the order of tasting should be from the most likely wine to be bad to the least. The expected number of tastings required is

$$\begin{aligned}\sum_{i=1}^{6} p_i l_i &= 1\times\frac{8}{23} + 2\times\frac{6}{23} + 3\times\frac{4}{23} + 4\times\frac{2}{23} + 5\times\frac{2}{23} + 5\times\frac{1}{23}\\ &= \frac{55}{23}\\ &= 2.39\end{aligned}$$

b) The first bottle to be tasted should be the one with probability $\frac{8}{23}$.

c) The idea is to use Huffman coding. With Huffman coding, we get codeword lengths as $(2, 2, 2, 3, 4, 4)$.

The expected number of tastings required is

$$\begin{aligned}\sum_{i=1}^{6} p_i l_i &= 2\times\frac{8}{23} + 2\times\frac{6}{23} + 2\times\frac{4}{23} + 3\times\frac{2}{23} + 4\times\frac{2}{23} + 4\times\frac{1}{23}\\ &= \frac{54}{23}\\ &= 2.35\end{aligned}$$

d) The mixture of the first and second bottles should be tasted first.

45) *Random "20" questions.*

a) Obviously, Huffman codewords for X are all of length n. Hence, with n deterministic questions, we can identify an object out of 2^n candidates.

b) Observe that the total number of subsets which include both object 1 and object 2 or neither of them is 2^{m-1}. Hence, the probability that object 2 yields the same answers for k questions as object 1 is $(2^{m-1}/2^m)^k = 2^{-k}$.

More information theoretically, we can view this problem as a channel coding problem through a noiseless channel. Since all subsets are equally likely, the probability the object 1 is in a specific random subset is $1/2$. Hence, the question whether object 1 belongs to the kth subset or not corresponds to the kth bit of the random codeword for object 1, where codewords X^k are Bern$(1/2)$ random k-sequences.

Object	Codeword
1	$0110\ldots1$
2	$0010\ldots0$
$\vdots$	

Now we observe a noiseless output Y^k of X^k and figure out which object was sent. From the same line of reasoning as in the achievability proof of the channel coding theorem, i.e. joint typicality, it is obvious the probability that object 2 has the same codeword as object 1 is 2^{-k}.

c) Let

$$1_j = \begin{cases} 1, & \text{object } j \text{ yields the same answers for } k \text{ questions as object } 1 \\ 0, & \text{otherwise} \end{cases},$$

$$\text{for } j = 2, \ldots, m.$$

Then,

$$\begin{aligned}
E(\#\text{ of objects in } \{2,3,\ldots,m\} \text{ with the same answers}) &= E(\sum_{j=2}^{m} 1_j) \\
&= \sum_{j=2}^{m} E(1_j) \\
&= \sum_{j=2}^{m} 2^{-k} \\
&= (m-1)2^{-k} \\
&= (2^n-1)2^{-k}.
\end{aligned}$$

d) Plugging $k = n + \sqrt{n}$ into (c) we have the expected number of $(2^n - 1)2^{-n-\sqrt{n}}$.

e) Let N by the number of wrong objects remaining. Then, by Markov's inequality

$$\begin{aligned}
P(N \geq 1) &\leq EN \\
&= (2^n-1)2^{-n-\sqrt{n}} \\
&\leq 2^{-\sqrt{n}} \\
&\to 0,
\end{aligned}$$

where the first equality follows from part (d).