LUND UNIVERSITY
Electrical and Information Technology

Information Theory

Problems and Solutions

2015

Contents

Problems

Chapter 1

Chapter 2

2.1. (a) Let X be a binary stochastic variable with $P(X = 0) = P(X = 1) = \frac{1}{2}$, and let Y be another independent binary stochastic variable with $P(Y = 0) = p$ and $P(Y = 1) = 1 - p$. Consider the modulo two sum $Z = X \oplus Y$. Show that Z is independent of Y for all values of p.

(b) Let X be a stochastic variable uniformly distributed over $\{1, 2, \ldots, M\}$. Let Y be independent of X, with an arbitrary probability function over $\{1, 2, \ldots, M\}$. Consider the sum $Z = X + Y$, mod M. Show that Z is independent of Y.

2.2. Two cards are drawn from an ordinary deck of cards. What is the probability that neither of them is a heart?

2.3. Two persons flip a fair coin n times each. What is the probability that they have the same number of heads?

2.4. Use that the logarithm is a concave function for positive arguments to show that[1]

$$(x_1 x_2)^{\frac{1}{2}} \leq \frac{x_1 + x_2}{2}, \quad x_1 x_2 \in \mathbb{Z}^+$$

Hint: A function $f(\cdot)$ is concave in the interval (a, b) if

$$\lambda f(x_1) + (1 - \lambda) f(x_2) \leq f\left(\lambda x_1 + (1 - \lambda) x_2\right),$$

for all choices of x_1, x_2 and λ such that $a \leq x_1 \leq x_2 \leq b$ and $0 \leq \lambda \leq 1$.

2.5. Consider a binary vector of length $N = 10$ where the bits are i.i.d. with $P(X = 0) = p = 0.2$. Construct a table where you list, for each possible number of zeros in the vector, the number of vectors with that number of zeros, the probability for each vector and the probability for the distribution of zeros.

Chapter 3

3.1. The so called IT-inequality is in the text described as a consequence of the fact that the functions $\ln x$ lower than $x - 1$, with equality if and only if $x = 1$. Show that this relation

$$x - 1 \geq \log_b x, \quad b > 1$$

only holds for the case when $b = e$.

[1] The proof can be extended to show that for positive numbers the geometric mean is upper bounded by the arithmetic mean,

$$\left(\prod_{k=1}^{N} x_k\right)^{\frac{1}{N}} \leq \frac{1}{N} \sum_{k=1}^{N} x_k$$

3.2. Show that, for all positive x, $\ln x \geq 1 - \frac{1}{x}$ with equality if and only if $x = 1$.

3.3. Let X be the outcome of a throw with a fair dice, and let Y be Even if X is even and Odd otherwise. Determine

(a) $I(X = 2; Y = \text{Even})$, $I(X = 3; Y = \text{Even})$, $I(X = 2 \text{ or } X = 3; Y = \text{Even})$.
(b) $I(X = 4)$, $I(Y = \text{Odd})$.
(c) $H(X)$, $H(Y)$.
(d) $H(X, Y)$, $H(X|Y)$, $H(Y|X)$.
(e) $I(X; Y)$.

3.4. Let X_1 and X_2 be two variables describing the outcome of a throw with two dice and let $Y = X_1 + X_2$ be the total number of pips.

(a) What is the probability function for the stochastic variable Y?
(b) Determine $H(X_1)$ and $H(Y)$.
(c) Determine $I(Y; X_1)$.

3.5. The joint probability of X and Y is given by

	$P(X,Y)$	$Y=0$	$Y=1$
X	0	$\frac{1}{3}$	$\frac{1}{3}$
	1	0	$\frac{1}{3}$

Calculate

(a) $P(X)$, $P(Y)$, $P(X|Y)$, and $P(Y|X)$
(b) $H(X)$ and $H(Y)$
(c) $H(X|Y)$ and $H(Y|X)$
(d) $H(X, Y)$
(e) $I(X, Y)$

3.6. The joint probability of X and Y is given by

	$P(X,Y)$	$Y=a$	$Y=b$	$Y=c$
X	A	$\frac{1}{12}$	$\frac{1}{6}$	0
	B	0	$\frac{1}{9}$	$\frac{1}{5}$
	C	$\frac{1}{18}$	$\frac{1}{4}$	$\frac{2}{15}$

Calculate

(a) $P(X)$, $P(Y)$, $P(X|Y)$, and $P(Y|X)$
(b) $H(X)$ and $H(Y)$
(c) $H(X|Y)$ and $H(Y|X)$

(d) $H(X,Y)$

(e) $I(X,Y)$

3.7. The joint probability of X and Y is given by

		Y		
$P(X,Y)$		a	b	c
X	0	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{3}$
	1	$\frac{1}{4}$	0	$\frac{1}{6}$

Calculate

(a) $P(X)$, $P(Y)$, $P(X|Y)$, and $P(Y|X)$

(b) $H(X)$ and $H(Y)$

(c) $H(X|Y)$ and $H(Y|X)$

(d) $H(X,Y)$

(e) $I(X,Y)$

3.8. Consider an experiment where we are given two coins. The first is a fair coin, while the second has heads on both sides. Choose with equal probability one of the coins, and flip it twice. How much information do we get about the identity of the coin by studying the number of heads from the flips?

3.9. Consider two dice where the first has equal probability for all six numbers. The second has a small weight close to the surface of number 1. Let X be the outcome of a throw with one of the dice, then the corresponding probability distributions for the dice are given below.

x	1	2	3	4	5	6
$p(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
$q(x)$	$\frac{1}{14}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{5}{14}$

(a) What is the entropy of a throw with the fair dice and the manipulated dice, respectively?

(b) What is the relative entropy from the fair to the manipulated dice?

(c) What is the relative entropy from the manipulated to the fair dice?

3.10. Let X be the the number of flips by a fair coin until the head comes up.

(a) What is the probability distribution of the number of coin flips, X?

(b) What is the expected value of the number of coin flips, $E[X]$?

(c) What is the entropy of the number of coin flips, $H(X)$?

(d) Repeat problem (a)-(c) for an un-fair coin with $P(\text{head}) = p$ and $P(\text{tail}) = q = 1 - p$.

Hint: The following standard sums might be helpful (for $|a| < 1$)

$$\sum_{n=0}^{\infty} a^n = \frac{1}{1-a} \qquad \sum_{n=0}^{\infty} n a^{n-1} = \frac{1}{(1-a)^2}$$

3.11. Consider the joint distribution of X and Y given by

$$p(x,y) = k^2 2^{-(x+y)}, \qquad x,y = 0,1,2,\ldots$$

(a) Derive $P(X < 4, Y < 4)$

(b) Derive the joint entropy.

(c) Derive the conditional probability $H(X|Y)$.

3.12. On several occasions in the text book we make use of Stirling's approximation to relate the binomial function with the binary entropy. There are actually different versions of the approximation in the literature, with different accuracy (and difficulty). Here we will consider one of the basic versions.

Consider the logarithm of the faculty function

$$y(n) = \log n!$$

(a) View $y(n)$ as a sum and interprete as a trapezoid approximation of an integral. Use this to show that

$$\log n! \approx n \ln n - n + 1 + \frac{1}{2} \ln n$$

or, equivalently,

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

(b) Consider the error made by the approximation and show that a better approximation can be given by

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

3.13. Consider the two distributions $p(x,y)$ and $q(x,y)$ over the 2-dimensional space $\mathcal{X} \times \mathcal{Y}$. Verify that

$$\begin{aligned} D\big(p(x,y)||q(x,y)\big) &= D\big(p(x)||q(x)\big) + \sum_x D\big(p(y|x)||q(y|x)\big)p(x) \\ &= D\big(p(y)||q(y)\big) + \sum_y D\big(p(x|y)||q(x|y)\big)p(y) \end{aligned}$$

and that, if X and Y are independent,

$$D\big(p(x,y)||q(x,y)\big) = D\big(p(x)||q(x)\big) + D\big(p(y)||q(y)\big)$$

3.14. Sometimes a function called *Cross Entropy*, closely related to the relative entropy, is used. It is defined as

$$H(p,q) = -\sum_x p(x) \log q(x)$$

Show that

$$H(p,q) = D\big(p||q\big) - H_p(X)$$

3.15. (a) Show that if α, β and γ form a probability distribution, then

$$H(\alpha, \beta, \gamma) = h(\alpha) + (1-\alpha)h\Big(\tfrac{\beta}{1-\alpha}\Big)$$

(b) Show that if $p_1, p_2, p_3, \ldots, p_n$ form a probability distribution, then

$$H(p_1, p_2, \ldots, p_n) = h(p_1) + (1-p_1)H\Big(\tfrac{p_2}{1-p_1}, \tfrac{p_3}{1-p_1}, \ldots, \tfrac{p_n}{1-p_1}, \Big)$$

3.16. Consider two urns, numbered 1 and 2. Urn 1 has four white balls and three black balls, while Urn 2 has three white balls and seven black. Choose one of the urns with equal probability, and draw one ball from it. Let X be the colour of that ball and Y the number of the chosen urn.

(a) Derive the uncertainty of X.

(b) How much information is obtained about Y when observing X?

(c) Introduce a third urn, Urn 3, with only one white ball (and no black). Redo problems a and b for this case.

3.17. Show that

$$I(X; Y, Z) = I(X; Y) + I(X; Z|Y)$$

3.18. In statistics, sometimes it is desirable to compare distributions and have a measure of how different they are. One way is, of course, to use the relative entropy $D(p||q)$ as a measure. However, it is not difficult to find distributions such that $D(p||q) \neq D(q||p)$. That is, the relative entropy is not a symmetric measure. Since, symmetry is one of the basic criterion for a metric this property is desirable. Below are some of them that are based on the relative entropy.

(a) One direct way to get a symmetric measurement of the difference between two distributions is the Jeffrey's divergence

$$D_J(p||q) = D(p||q) + D(q||p)$$

named after the statistician Harold Jeffreys. Show that it can be written as (for discrete distributions)

$$D_J(p||q) = \sum_x \Big(p(x) - q(x)\Big) \log \frac{p(x)}{q(x)}$$

(b) To get around the problem that there can occur infinite values in the Jeffrey's divergence, Lin introduced in 1991 the so called Jensen-Shannon divergence,

$$D_{JS}(p||q) = \frac{1}{2} D\Big(p||\frac{p+q}{2}\Big) + \frac{1}{2} D\Big(q||\frac{p+q}{2}\Big)$$

Show that an alternative way to write this is

$$D_{JS}(p||q) = H\Big(\frac{p+q}{2}\Big) - \frac{H(p) + H(q)}{2} \tag{1}$$

3.19. Let $p(x)$ and $q(x)$ be two probability functions for the random variable X. Use the relative entropy to show that

$$\sum_x \frac{p^2(x)}{q(x)} \geq 1$$

with equality if and only if $p(x) = q(x)$ for all x.

3.20. A Markov source with output symbols $\{A, B, C\}$, is characterised by the graph in Figure 1.

(a) What is the stationary distribution for the source?

(b) Determine the entropy of the source, H_∞.

(c) Consider a memory-less source with the same probability distribution as the stationary distribution calculated in (a). What is the entropy for the memory-less source?

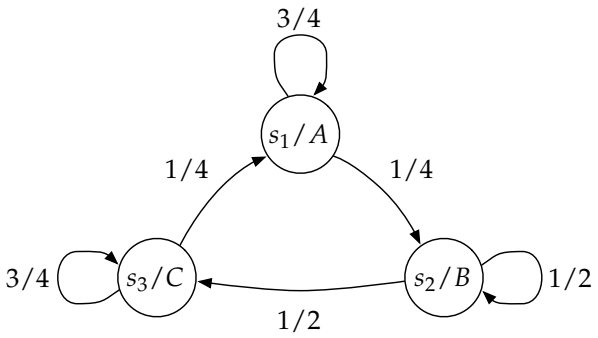


Figure 1: A Markov graph for the source in Problem 3.20

3.21. The engineer Inga is going to spend her vacation in the archipelago. She decides to go to a small archipelago with only four islands with boat connections. The four islands are connected with four different boat lines, and one sightseeing tour around the largest island, see the map in Figure 2.

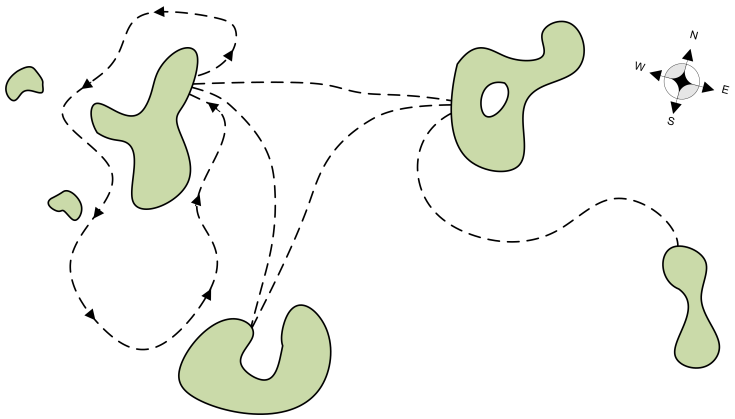


Figure 2: A map over the islands and the boat connections.

To avoid planning the vacation rout too much, she decides to take a boat every day. She will choose a boat line going out from the island with equal probability. All the boat lines are routed both ways every day, except the sightseeing tour that is only one-way.

(a) When Inga has travelled around in the archipelago for a long time, what is the probabilities for being on the islands?

(b) Inga has promised to write home and tell her friends about her travel. How many bits, in average, does she need to write per day to describe her rout? Assume that she will choose a starting island for her vacation according to the distribution in (a).

3.22. Consider a miniature chessboard with 3×3 squares as in Figure 3. For sake of clarity we have numbered the squares. On this chessboard we put a bishop. the bishop in a chess set can move

diagonally any number of squares. In this example we will consider a sort of random walk of

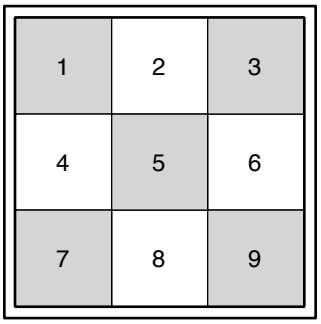


Figure 3: A 3×3 chessboard.

the bishop on this miniature chessboard. The rules are simple; When the bishop can move to n different squares it chooses one with equal probability, $1/n$. It is not allowed for the bishop to not move. Assume that the game starts at time $T = -\infty$ and let X_i, $i = -\infty, \ldots, 0, \ldots \infty$ be the position at time $T = i$. Consider the sequence of positions from time $T = 0$ as a sequence generated by a Markov source, $\boldsymbol{X} = X_0, X_1, \ldots$.

(a) If the bishop starts at a white square, what is $H_\infty(X)$?

(b) If the bishop starts at a black square, what is $H_\infty(X)$?

3.23. Use the strong form of Stirling's approximation

$$\sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n \le n! \le \sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n e^{\frac{1}{12n}}$$

and justify the steps in the following calculations, where n is an integer and $0 < p < 1$, $q = 1 - p$.

$$\begin{aligned}
\binom{n}{np} &\overset{(a)}{\le} \frac{1}{\sqrt{2\pi npq}} \frac{1}{p^{np}q^{nq}} e^{\frac{1}{12n}} \\
&\overset{(b)}{=} \frac{1}{\sqrt{2\pi npq}} 2^{nh(p)} e^{\frac{1}{12n}} \\
&\overset{(c)}{\le} \frac{1}{\sqrt{\pi npq}} 2^{nh(p)}
\end{aligned}$$

$$\begin{aligned}
\binom{n}{np} &\overset{(d)}{\ge} \frac{1}{\sqrt{2\pi npq}} \frac{1}{p^{np}q^{nq}} e^{-\frac{1}{12npq}} \\
&= \frac{1}{\sqrt{2\pi npq}} 2^{nh(p)} e^{-\frac{1}{12npq}} \\
&\overset{(e)}{\ge} \frac{1}{\sqrt{8npq}} 2^{nh(p)}
\end{aligned}$$

where, in the last inequality, it can be assumed that $12npq \ge 9$.

The above calculations gives a useful bound on the binomial coefficient,

$$\frac{1}{\sqrt{8npq}} 2^{nh(p)} \le \binom{n}{np} \le \frac{1}{\sqrt{\pi npq}} 2^{nh(p)}$$

Chapter 4

4.1. Consider the code $\{0, 01\}$.

(a) Is it a prefix-free?

(b) Is it uniquely decodable?

(c) Is it non-singular?

4.2. Consider the random variable X and the binary (prefix-free) code

x	$p(x)$	y
x_1	0.2	0
x_2	0.2	100
x_3	0.2	101

x	$p(x)$	y
x_4	0.2	110
x_4	0.1	1110
x_6	0.1	1111

(a) Draw a binary tree that represents the code.

(b) Derive the entropy $H(X)$ and the average codeword length $E[L]$.

(c) Is it possible that the code is an optimal code for X?

4.3. For each of the following sets of codeword lengths decide if there exists a binary prefix-free code. If it exists, construct a code.

(a) $\{1, 2, 3, 4, 5\}$

(b) $\{2, 2, 3, 3, 4, 4, 5, 5\}$

(c) $\{2, 2, 2, 3, 4, 4, 4, 5\}$

(d) $\{2, 3, 3, 3, 4, 5, 5, 5\}$

4.4. Show, by induction, that a binary Huffman tree with n leaves has $n - 1$ inner nodes.

4.5. For a given k-ary source the estimated probability function is $q(x)$, $x \in \{0, 1, \ldots, k-1\}$. Consider the case when an optimal source code for this source is chosen (assume that this is possible neglecting the truncation error in the logarithm). Let the true distribution for the symbols be $p(x)$, $x \in \{0, 1, \ldots, N-1\}$. Show that the relative entropy is the penalty in bits per source symbol due to the estimation error.

Use the result above to give an interpretation of the mutual information $I(X, Y)$, viewed from a source coding perspective.

4.6. Is the code given in Problem 4.2 optimal?

4.7. Consider a random variable with distribution according to

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7
$p(x)$	0.3	0.2	0.2	0.1	0.1	0.06	0.04

Construct a binary optimal source code for the source.

4.8. A binary information source with probabilities $P(X = 0) = \frac{3}{5}$ and $P(X = 1) = \frac{2}{5}$ produces a sequence. The sequence is split in blocks of three bits each, that should be encoded separately. In the table below such a code is given. Is the suggested code optimal? If not, construct one and derive the average gain in code bits per source bit.

x	y
000	0
001	100
010	110
011	11100

x	y
100	101
101	11101
110	11110
111	11111

4.9. Consider a binary Huffman code with codewords $\{y_1, y_2, \ldots, y_n\}$, $n > 2$. According to the Huffman code construction, we can assume the codewords are ordered such that the corresponding probabilities are in decreasing order, i.e. $p_1 \geq p_2 \geq \cdots \geq p_n$, and the corresponding lengths has increasing order, $l_1 \leq l_2 \leq \cdots \leq l_n$.

(a) Show that if $p_1 > \frac{2}{5}$ we must have $l_1 = 1$.

(b) Show that if $p_1 < \frac{1}{3}$ we must have $l_1 \geq 2$.

4.10. Consider the first order statistics of the letters in the English alphabet given below. It lists, for each letter, the total average number of occurrences out of a 1000 letter text.

A	73
B	9
C	30
D	44
E	130

F	28
G	16
H	35
I	74
J	2

K	3
L	35
M	25
N	78
O	74

P	27
Q	3
R	77
S	63
T	93

U	27
V	13
W	16
X	5
Y	19
Z	1

Construct a binary Huffman code for this set of letters. What is the average codewords length? How does this compare to the entropy of the letters? How does it compare to a coding where the statistics is not taken into consideration?

4.11. Assume a binary random variable X with $P(X = 0) = 0.1$ and $P(X = 1) = 0.9$.

(a) Find the average codeword length of an optimal source code for X.

(b) Consider vectors of n i.i.d. X, $\boldsymbol{X} = X_1 X_2 \ldots X_n$. Construct optimal source codes for the cases $n = 2, 3, 4$. What is the average codeword lengths per binary source symbol?

(c) Compare the above results with the entropy of X.

4.12. We are given six bottles of wine. It is known that exactly one of them is bad. From a visual inspection the following probabilities p_i for bottle i is bad are estimated,

$$(p_1, \ldots, p_6) = \left(\frac{8}{23}, \frac{6}{23}, \frac{4}{23}, \frac{2}{23}, \frac{2}{23}, \frac{1}{23},\right)$$

To determine which bottle is bad the wine should be tasted.

(a) Suppose the wines are tasted one by one. Choose the order of tasting so the expected number of tastings to find the bad wine is minimised. Which bottle should be tasted first? What is the expected number of tastings? (It is not necessary to taste the last bottle).

(b) Then a smarter solution. Instead of tasting the wines one by one you are allowed to mix some bottles in a glass to taste. What mixture should be tasted first? What is the expected number of tastings to find the bad wine?

Chapter 5

5.1. Encode the text

```
IF IF = THEN THEN THEN = ELSE ELSE ELSE = IF;
```

using the LZ77 algoritm with $S = 7$ and $B = 7$. How many code symbols were generated? If each letter in the text is translated to binary form with eight bits, what is the compression ratio?

5.2. A text has been encoded with the LZ78 algorithm and the following sequence of codewords was obtained,

Index	Codeword
1 :	$(0, t)$
2 :	$(0, i)$
3 :	$(0, m)$
4 :	$(0, \sqcup)$
5 :	$(1, h)$
6 :	$(0, e)$
7 :	$(4, t)$
8 :	$(0, h)$
9 :	$(2, n)$
10 :	$(7, w)$
11 :	$(9, \sqcup)$
12 :	$(1, i)$
13 :	$(0, n)$
14 :	$(0, s)$
15 :	$(3, i)$
16 :	$(5, .)$

Decode to get the text back.

5.3. Encode the text

```
IF IF = THEN THEN THEN = ELSE ELSE ELSE = IF;
```

using the LZ78 algoritm. How many code symbols were generated? If each letter in the text is translated to binary form with eight bits, what is the compression ratio?

5.4. Consider the sequence

```
Nat the bat swat at Matt the gnat
```

Encode and calculate the compression rate using

(a) LZ77 with $S = 10$ and $B = 3$.

(b) LZSS with $S = 10$ and $B = 3$.

(c) LZ78.

(d) LZW with predefined alphabet of size 256.

5.5. Consider the sequence

```
six sick hicks nick six slick bricks with picks and sticks
```

(a) What source alphabet should be used?

(b) Use the LZ77 with a window size $N = 8$ to encode and decode the sequence with a binary code alphabet.

(c) Use the LZ78 to encode and decode the sequence with a binary code alphabet.

(d) How many code symbols were generated?

5.6. Use the LZ78 algorithm to encode and decode the string

```
the friend in need is the friend indeed
```

with a binary code alphabet. What is the minimal source alphabet? How many code symbols were generated?

5.7. Consider a binary memory-less source where $P(0) = p$ and $P(1) = q = 1 - p$. In a sequence of n symbols, the share of 1s tends to q as n becomes large.

(a) How many sequences of length n has the share of ones equal to q?

(b) How many bits per source symbol is required to represent the sequences in a).

(c) Show that as $n \to \infty$ the number bits per source symbol required to represent the sequences in a) is the entropy, $h(q) = h(p)$.

Hint: Use Stirling's formula to approximate $n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$.

Chapter 6

6.1. Show that for all jointly ε-typical sequences, $(\boldsymbol{x}, \boldsymbol{y}) \in A_\varepsilon(X, Y)$, we get

$$2^{-n(H(X|Y)+2\varepsilon)} \leq p(\boldsymbol{x}|\boldsymbol{y}) \leq 2^{-n(H(X|Y)-2\varepsilon)}$$

6.2. A binary memory-less source with $P(X = 0) = \frac{49}{50}$ and $P(X = 1) = \frac{1}{50}$ generates sequences of length $n = 100$. Choose $\epsilon = \frac{1}{50}\log 7$.

(a) Which sequence is the most probable?

(b) Is the most probable sequence ϵ-typical?

(c) How many ϵ-typical sequences of length $n = 100$ are there?

(d) Give numerical upper and lower bounds on the number of ϵ-typical sequences of length $n = 100$.

6.3. A string is 1 meter long. It is split in two pieces where one is twice as long as the other. With probability $3/4$ the longest part is saved and with probability $1/4$ the short part is saved. Then, the same split is done with the saved part, and this continues the same way with a large number of splits. How large share of the string is, in average, saved at each split?

Hint: Consider the distribution of saved parts for the most common type of sequence.

6.4. One is given a communication channel with transition probabilities $p(y|x)$ and channel capacity $C = \max_{p(x)} I(X;Y)$. A helpful statistician preprocesses the output by forming $\tilde{Y} = g(Y)$. He claims that this will strictly improve the capacity.

(a) Show that he is wrong.

(b) Under what conditions does he not strictly decrease the capacity?

6.5. Let $X \in \mathbb{Z}_{11} = \{0, 1, \ldots, 10\}$, be a random variable used as input to an additive channel,

$$Y = X + Z, \qquad \text{mod } 11$$

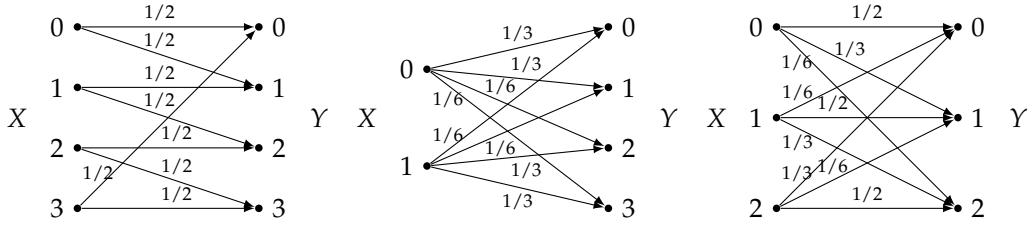
where $p(Z = 1) = p(Z = 2) = p(Z = 3) = \frac{1}{3}$. Assume that X and Z are statistically independent.

(a) What is the capacity of the channel?

(b) What distribution on $p(x)$ gives the capacity?

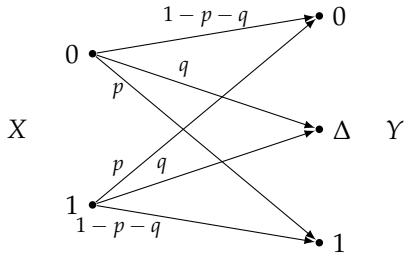
6.6. Consider the discrete memoryless channel $Y = X \cdot Z$ where X and Z are independent binary random variables. Let $P(Z = 1) = \alpha$. Find the capacity of this channel and the maximising distribution on X.

6.7. In Shannon's original paper from 1948, the following discrete memoryless channels are given. Calculate their channel capacities.

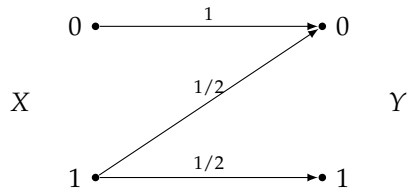


(a) Noisy typewriter (b) Soft decoding (c) 3-ary channel

6.8. Consider the binary erasure channel below. Calculate the channel capacity.

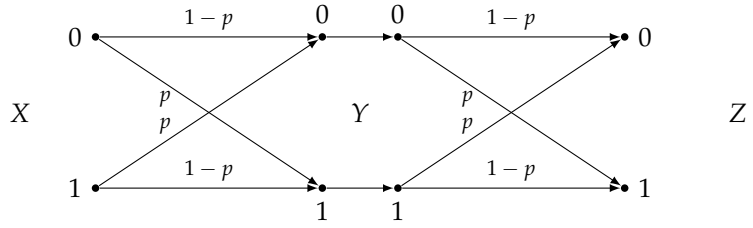


6.9. Determine the channel capacity for the following Z-channel.



Hint: $D(h(p)) = D(p)\log(\frac{1-p}{p})$.

6.10. Cascade two binary symmetric channels as in the following picture. Determine the channel capacity.



6.11. Consider the discrete memoryless channel shown in Figure 4.

(a) What is the channel capacity and for what distribution on X is it reached?

(b) Assume that for the sequence $\boldsymbol{X}$ we have $P(X = 0) = 1/6$ and $P(X = 0) = 5/6$, and that the source is memoryless. Find an optimal code to compress the sequence $\boldsymbol{Y}$. What is the average codeword length?

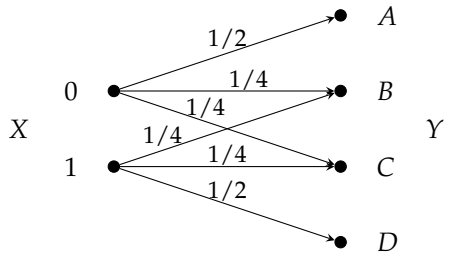


Figure 4: A discrete memoryless channel

6.12. Consider two channels in cascade, one BSC and one BEC, according to Figure 5. Derive the channel capacity.

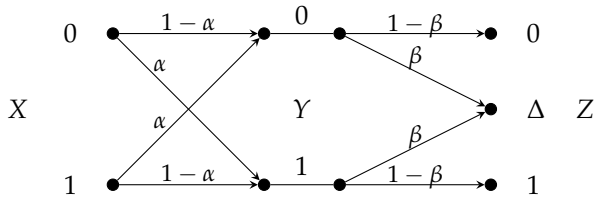


Figure 5: One BSC and one BEC in cascade.

6.13. In Figure 6 a general Z-channel is shown. Plot the capacity as a function of the error probability α.

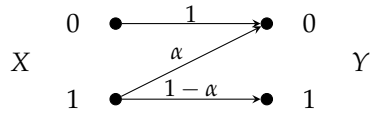
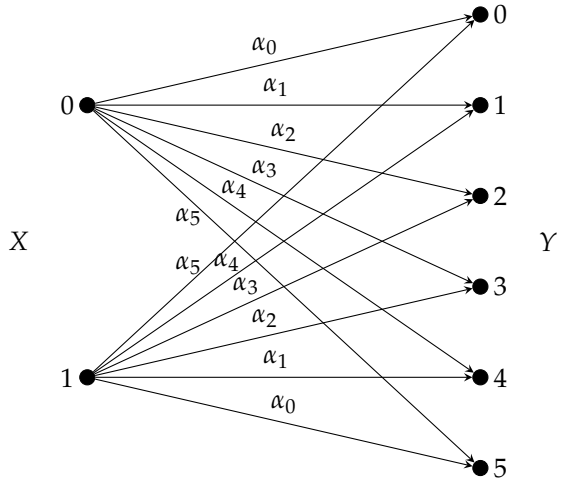


Figure 6: One BSC and one BEC in cascade.



		Y					
$P(Y\|X)$		0	1	2	3	4	5
X	0	α_0	α_1	α_2	α_3	α_4	α_5
	1	α_5	α_4	α_3	α_2	α_1	α_0

Figure 7: A channel and its probability function.

6.14. Consider the discrete memory-less channel (DMC) given in Figure 7.

(a) Show that the maximising distribution giving the capacity

$$C_6 = \max_{p(x)} I(X;Y)$$

is given by $P(X=0) = \frac{1}{2}$ and $P(X=1) = \frac{1}{2}$.
Verify that the capacity is given by

$$C_6 = 1 + H(\alpha_0 + \alpha_5, \alpha_1 + \alpha_4, \alpha_2 + \alpha_3) - H(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$$

(b) Split the outputs in two sets, $\mathcal{Y}_0 = \{0,1,2\}$ and $\mathcal{Y}_1 = \{3,4,5\}$, and construct a binary symmetric channel, i.e. a BSC with error probability $p = \alpha_3 + \alpha_4 + \alpha_5$. Denote the capacity of the corresponding BSC as C_{BSC} and show that

$$C_{\text{BSC}} \leq C_6 \leq 1$$

where C_6 is the capacity of the channel in Figure 7.

Chapter 7

7.1. Consider a linear encoder where three information bits $\boldsymbol{u} = (u_0, u_1, u_2)$ is complemented with three parity bits according to

$$v_0 = u_1 \oplus u_2$$
$$v_1 = u_0 \oplus u_2$$
$$v_2 = u_0 \oplus u_1$$

Hence, an information word $\boldsymbol{u} = (u_0, u_1, u_2)$ is encoded to the codeword $\boldsymbol{x} = (u_0, u_1, u_2, v_0, v_1, v_2)$.

(a) What is the code rate R?

(b) Find a generator matrix G.

(c) What is the minimum distance, $d_{\min}$, of the code?

(d) Find a parity check matrix H, such that $GH^T = 0$.

(e) Construct a syndrom table for decoding.

(f) Make an example where a three bit vector is encoded, transmitted over a channel and decoded.

7.2. Show that if $d_{\min} \geq \lambda + \gamma + 1$ for a linear code, it is capable of correcting λ errors and simultaneously detecting γ errors, where $\gamma > \lambda$.

7.3. Consider an (M, n) block code $\mathcal{B}$. One way to extend the code is to add one more bit such that the codeword has even Hamming weight, i.e.

$$\mathcal{B}_E = \{(y_1 \dots y_n y_{n+1}) | (y_1 \dots y_n) \in \mathcal{B} \text{ and } y_1 + \cdots + y_n + y_{n+1} = 0 \pmod 2\}$$

(a) Show that if $\mathcal{B}$ is a linear code, so is $\mathcal{B}_E$. If you instead extend the code with a bit such that the number of ones is odd, will the code still be linear?

(b) Let H be the parity check matrix for the code $\mathcal{B}$ and show that

$$H_E = \begin{pmatrix} & & & 0 \\ & H & & \vdots \\ & & & 0 \\ 1 & \cdots & 1 & 1 \end{pmatrix}$$

is the parity check matrix for the extended code $\mathcal{B}_E$.

(c) What can you say about the minimum distance for the extended code?

(d) In the childhood of computing the ASCII table consisted of seven bit vectors where an extra parity bit was appended such that the vector always had even number of ones. This was an easy way to detect errors in e.g. punch-cards. What is the parity check matrix for this code?

7.4. Plot, using e.g. MATLAB, the resulting bit error rate as a function of E_b/N_0 when using binary repetition codes of rate $R = 1/3$, $R = 1/5$ and $R = 1/7$. Compare with the uncoded case. Notice that E_b is the energy per information bit, i.e. for a rate $R = 1/N$ the energy per transmitted bit is E_b/N. The noise parameter is naturally independent of the code rate.

7.5. One of the best known asymptotic upper bound for the code rate is

$$R \leq h\Big(\frac{1}{2} - \sqrt{\delta(1-\delta)}\Big)$$

Show with a plot how this improves the result from the asymptotic Hamming bound and compare with with the Gilbert-Varshamov bound.

7.6. Verify that the free distance for the code generated by the generator matrix generator matrix

$$G(D) = \begin{pmatrix} 1 + D + D^2 & 1 + D^2 \end{pmatrix}$$

is $d_{\text{free}} = 5$. Decode the received sequence

$$\boldsymbol{r} = 01\ 11\ 00\ 01\ 11\ 00\ 01\ 00\ 10$$

7.7. A convolutional code is formed from the generator matrix

$$G(D) = \begin{pmatrix} 1+D & 1+D+D^2 \end{pmatrix}$$

(a) Derive the free distance d_{free}.

(b) Decode the received sequence

$$\boldsymbol{r} = 01\ 11\ 00\ 01\ 11\ 00\ 01\ 00\ 10$$

Assume that the encoder is started and ended in the all-zero state.

7.8. Repeat Problem 7.7 for the generator matrix

$$G(D) = \begin{pmatrix} 1+D+D^2+D^3 & 1+D+D^3 \end{pmatrix}$$

7.9. For the generator matrix in Problem 7.6, show that the generator matrix

$$G_s(D) = \begin{pmatrix} \frac{1+D+D^2}{1+D^2} & 1 \end{pmatrix}$$

will give the same code as $G(D)$.

7.10. Suppose a 4-bit CRC with generator polynomial $g(x) = x^4 + x^3 + 1$ has been used. Which, if any, of the following three messages will be accepted by the receiver?

(a) 11010111

(b) 10101101101

(c) 10001110111

7.11. Consider a data frame with six bits where we add a four bit CRC at the end, see Figure 8.

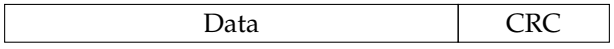


Figure 8: Six data bits and four CRC bits.

To calculate the CRC bits the following generator polynomial is used

$$g(x) = (1+x)(1+x^2+x^3) = 1+x+x^2+x^4$$

(a) Will the encoding scheme be able to detect all
- single errors?
- double errors?
- triple errors?
- quadruple errors?

(b) Assume the data vector $\boldsymbol{d} = 010111$ should be transmitted. Find the CRC bits for the frame. Then, introduce an error pattern that is detectable and show how the detection works.

Chapter 8

8.1. Derive the differential entropy for the following distributions:

(a) Rectangular distribution: $f(x) = \frac{1}{b-a}, \quad a \leq x \leq b.$

(b) Normal distribution: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty \leq x \leq \infty.$

(c) Exponential distribution: $f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$

(d) Laplace distribution: $f(x) = \frac{1}{2}\lambda e^{-\lambda|x|}, \quad -\infty \leq x \leq \infty.$

8.2. Consider the joint distribution on X and Y given by

$$f(x,y) = \alpha^2 e^{-(x+y)}$$

for $x \geq 0$ and $y \geq 0$. (Compare with Problem 3.11)

(a) Derive $P(X < 4, Y < 4)$.

(b) Derive the joint entropy.

(c) Derive the conditional entropy $H(X|Y)$.

8.3. Repeat Problem 8.2 for

$$f(x,y) = \alpha^2 2^{-(x+y)}$$

8.4. In wireless communication the attenuation due to a shadowing object is often modeled as a log-Normal random variable, $X \sim \log\mathrm{N}(\mu,\sigma)$. If the logarithm of a random variable X is normal distriibuted, i.e. $Y = \ln X \sim \mathrm{N}(\mu,\sigma)$, then X is said to be logNormal distributed. Notice that $X \in [0,\infty]$ and $Y \in [-\infty,\infty]$.

(a) Use the probability

$$P(X < a) = \int_0^a f_X(x)dx$$

to show that the density function is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$$

(b) Use the density function in (a) to find

$$E[X] = e^{\mu+\frac{\sigma^2}{2}}$$
$$E[X^2] = e^{2\mu+2\sigma^2}$$
$$V[X] = e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$$

(c) Show that the entropy is

$$H(X) = \frac{1}{2}\log 2\pi e\sigma^2 + \frac{\mu}{\ln 2}$$

8.5. Let $h(x)$ be the density function for a Normal distribution, $N(\mu, \sigma)$, and $f(x)$ any distribution with the same mean and variance. Then, show that

$$D\big(f(x)\big|\big|h(x)\big) = \log 2\pi e\sigma^2 - H_f(X)$$

8.6. Let X_1 and X_2 be two independent normal distributions random variables with distributions $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$, respectively. Construct a new random variable $X = X_1 + X_2$.

(a) What is the distribution of X?

(b) Derive the differential entropy of X.

8.7. The length X a stick that is manufactured in a poorly managed company, is uniformly distributed.

(a) The length varies between 1 and 2 meters, i.e.

$$f(x) = \begin{cases} 1, & 1 \leq x \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

Derive the differential entropy $H(X)$.

(b) The length varies between 100 and 200 cm, i.e.

$$f(x) = \begin{cases} 0.01, & 100 \leq x \leq 200 \\ 0, & \text{otherwise} \end{cases}$$

Derive the differential entropy $H(X)$.

8.8. Consider a channel with binary input with $P(X = 0) = p$ and $P(X = 1) = 1 - p$. During the transmission a uniformly distributed noise parameter Z in the interval $[0, a]$, where $a > 1$, is added to X, i.e. $Y = X + Z$.

(a) Calculate the mutual information according to

$$I(X;Y) = H(X) - H(X|Y)$$

(b) Calculate the mutual information according to

$$I(X;Y) = H(Y) - H(Y|X)$$

(c) Calculate the capacity by maximising over p.

Chapter 9

9.1. Consider an additive channel where the output is $Y = X + Z$, where the noise is normal distributed with $N(0, \sigma)$. The channel has an output power constraint $E[Y^2] \leq P$. Derive the channel capacity for the channel.

9.2. Consider a random variable X, drawn from a uniform distribution $\mathrm{U}(1)$, that is transmitted over a channel with additive noise Z, also distributed uniformly $\mathrm{U}(a)$ where $a \leq 1$. The received random variable is then $Y = X + Z$. Derive the average information obtained about X from the received Y, i.e. $I(X;Y)$.

9.3. Consider two channels, both with attenuation and Gaussian noise. The first channel has the attenuation H_1 and noise distribution $n_1 \sim N(0, \sqrt{N_1})$ and the second channel has attenuation H_2 and noise distribution $n_2 \sim N(0, \sqrt{N_2})$. Then the two channels are used in cascade, i.e. a signal X is first transmitted over the first channel and then over the second channel, see Figure 9. Assume that both channels work over the same bandwidth W.

(a) Derive an expression for the channel capacity for the cascaded channel.

(b) Denote the signal to noise ratio over the cascaded channel as SNR and the two constituent channels as SNR_1 and SNR_2, respectively. Show that

$$\text{SNR} = \frac{\text{SNR}_1 \cdot \text{SNR}_2}{\text{SNR}_1 + \text{SNR}_2}$$

Notice that the formula is equivalent to the total resistance of a parallel coupling in electronics design.

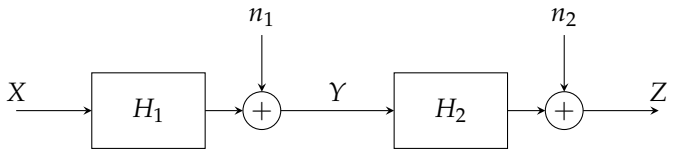


Figure 9: A channel consisting of two Gaussian channels.

9.4. Consider four independent, parallel, time discrete, additive Gaussian channels. The variance of the noise in the ith channel is $\sigma_i = i^2$, $i = 1, 2, 3, 4$. The total power of the used signals is limited by

$$\sum_{i=1}^{4} P_i \leq 17.$$

Determine the channel capacity for this parallel combination.

9.5. Consider six parallell Gaussian channels with the noise levels

$$N = (8, 12, 14, 10, 16, 6)$$

The total allowed power usage in the transmitted signal is $P = 19$.

(a) What is the capacity of the combined channel?

(b) If you must divide the power equally over the six channels, what is the capacity?

(c) If you decide to use only one of the channels, what is the maximum capacity?

Chapter 10

10.1. In a communication system a binary signalling is used, and the transmitted variable X has two equally likely amplitudes $+1$ and -1. During transmission a uniform noise is added to the signal, and the received variable is $Y = X + Z$ where $Z \sim U(\alpha)$. Derive the maximum transmitted number of bits per channel use, when

(a) $\alpha < 2$

(b) $\alpha \geq 2$

10.2. In the channel model described in Problem 10.1 consider the case when $\alpha = 4$. A hard decoding of the channel can be done by assigning

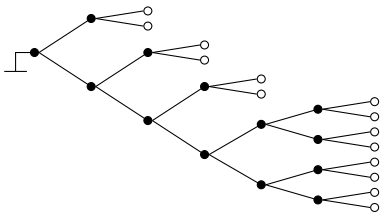
$$\tilde{Y} = \begin{cases} 1, & Y \geq 1 \\ \Delta, & -1 < Y \leq 1 \\ -1, & Y < -1 \end{cases}$$

Derive the capacity for the channel from X to $\tilde{Y}$ and compare with the result in Problem 10.1.

10.3. An additive channel, $Y = X + Z$, has the input alphabet $\mathcal{X} = \{-2, -1, 0, 1, 2\}$. The additive random variable Z is uniformly distributed over the interval $[-1, 1]$. Thus, the input is a discrete random variable and the output is a continuous random variable. Derive the capacity $C = \max_{p(x)} I(X;Y)$.

10.4. In Example 9.4 a shaping algorithm based on a binary tree construction is given. In this problem the same construction is used and the number of signal alternative expanded.

(a) Consider a tree with two more nodes as below. What is the shaping gain for this construction?



(b) Letting the shaping constellation and tree in Example 9.4 have two levels and in subproblem a have three levels. Consider the same construction with k levels and show that $L = 3$ for all $k \geq 2$.

(c) For the constellation in subproblem b show that as $k \to \infty$ the second moment is $E[X_s^2] = 17$ and thus, the asymptotic shaping gain is $\gamma_s^{(\infty)} = 0.9177$dB.
Note: It might be useful to consider the following standard sums for $|\alpha| < 1$,

$$\sum_i^\infty \alpha^i = \frac{\alpha}{1-\alpha} \quad \sum_i^\infty i\alpha^i = \frac{\alpha}{(1-\alpha)^2} \quad \sum_i^\infty i^2\alpha^i = \frac{\alpha+\alpha^2}{(1-\alpha)^3} \quad \sum_i^\infty i^3\alpha^i = \frac{\alpha+4\alpha^2+\alpha^3}{(1-\alpha)^4}$$

10.5. The shaping gain, γ_s, can be derived in two ways. First, it is the relation in power between a uniform distribution and a Gaussian distribution with equal entropies. Second, it is the relation between second moments of an N-dimensional square distribution and an N-dimensional spheric distribution, as $N \to \infty$. In this problem we will show that the results equivalent, since the spherical distribution in $N \to \infty$ dimensions, projected to one dimension is Gaussian.

(a) What is the radius in an N-dimensional sphere if the volume is one, i.e. if it is a probability distribution?

(b) If $\boldsymbol{X} = (X_1, \ldots, X_N)$ is an N-dimensional spherical (uniform) distribution, show that its projection in one dimension is

$$f_X(x) = \int_{|\tilde{\boldsymbol{x}}| \leq R^2 - x^2} d\tilde{\boldsymbol{x}} = \frac{\pi^{\frac{N-1}{2}}}{\Gamma\left(\frac{N}{2} + \frac{1}{2}\right)} \left(\frac{\Gamma\left(\frac{N}{2}+1\right)^{2/N}}{\pi} - x^2 \right)^{\frac{N-1}{2}}$$

where $\tilde{\boldsymbol{x}}$ is an $N-1$ dimensional vector.

(c) Using the first order Stirling's approximation[2]

$$\Gamma(x) \approx \left(\frac{x-1}{e}\right)^{x-1}$$

show that the result in b can be written as

$$f_X(x) \approx \left(1 + \frac{\frac{1}{2} - \pi e x^2}{\frac{N-1}{2}}\right)^{\frac{N-1}{2}}$$

for large N.

(d) Let the dimensionality N grow to infinity and use $\lim_{N\to\infty}(1+\frac{x}{N})^N = e^x$ to show that $X \sim N\left(0, \frac{1}{2\pi e}\right)$, i.e. that

$$\lim_{N\to\infty} f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}$$

where $\sigma^2 = \frac{1}{2\pi e}$.

10.6. A signal is based on an OFDM modulation with 16 sub-channels of width $\Delta f = 10$kHz. The signal power level in the whole spectra is -70dBm/Hz. On the transmission channel the noise level is constant at -140dBm/Hz, but the signal attenuation is increasing with the frequency as $|H_i|^2 = 5i + 1$dB, $i = 0, \ldots, 15$.

(a) Derive the capacity for the channel.

(b) If the required bit rate on the channel is 10^{-7}, and it is expected that the error correcting code gives a coding gain of 3dB, what is the estimated obtained bit rate for the system?

Chapter 11

11.1. Consider a k-ary source with source statistics $P(X = x) = \frac{1}{k}$. Given the Hamming distortion

$$d(x, \hat{x}) = \begin{cases} 0, & x = \hat{x} \\ 1, & x \neq \hat{x} \end{cases}$$

and that the source and destination alphabets are the same, show that the rate-distortion function is

$$R(\delta) = \begin{cases} \log(k) - \delta\log(k-1) - h(\delta), & 0 \leq \delta \leq 1 - \frac{1}{k} \\ 0, & \delta \geq 1 - \frac{1}{k} \end{cases}$$

11.2. In this problem the rate-distortion function for the exponential distribution, $X \sim \text{Exp}(\mu)$ and distortion measure $d(x, \hat{x}) = x - \hat{x}$, will be derived. The problem is divided in two parts, first it must be shown that the exponential distribution maximises the differential entropy for all one-sided distributions, and then the rate-distortion can be derived. For simplicity the derivation will be performed over the natural base in the logarithm.

[2]Also the second order approximation $\Gamma(x) \approx \sqrt{2\pi(x-1)}\left(\frac{x-1}{e}\right)^{x-1}$ can be used but the derivations and the limit value becomes a bit more complicated.

(a) In this part of the problem it will be shown that the exponential distribution maximises the entropy over all one-sided distributions. Consider a distribution $f(x)$ such that $f(x) \geq 0$, with equality for $x < 0$, and $E[X] = 1/\mu$. Then, the idea is to maximise the entropy $H(X) = -\int_0^\infty f(x) \ln f(x) dx$ with respect to the requirements

$$\begin{cases} \int_0^\infty f(x)dx = 1 \\ \int_0^\infty xf(x)dx = 1/\mu \end{cases} \tag{2}$$

- Set up a maximisation function according to Lagrange multiplier method.
- Differentiate with respect to the function f, and equal to zero. Show that this gives an optimising function on the form

  $$f(x) = e^{\alpha+\beta x}$$

- Let $g(x)$ be an arbitrary distribution with the same requirements as in (2). Show that

  $$H_g(X) \leq H_f(X)$$

  i.e. that $f(x)$ maximises the entropy.
- Show that the equation system in (2) is solved by the exponential distribution.

(b) Consider the distortion measure

$$d(x, \hat{x}) = \begin{cases} x - \hat{x}, & x \geq \hat{x} \\ \infty, & o.w. \end{cases}$$

Use the result that the exponential distribution maximises the entropy over all one-sided distributions to show that $I(X; \hat{X}) \geq -\log(\mu\delta)$ where $E[d(x, \hat{x})] \leq \delta$. Design a backward test-channel that shows the rate-distortion for the exponential distribution is

$$R(\delta) = \begin{cases} -\log(\mu\delta), & 0 \leq \delta \leq 1/\mu \\ 0, & \delta > 1/\mu \end{cases}$$

11.3. Consider a source with i.i.d. symbols generated according to Laplacian distribution,

$$f_\alpha(x) = \frac{\alpha}{2} e^{-\alpha|x|}, \quad -\infty \leq x \leq \infty$$

With the distortion measure $d(x, \hat{x}) = |x - \hat{x}|$, show that the rate-distortion function is

$$R(\delta) = \begin{cases} -\log(\alpha\delta), & 0 \leq \delta \leq \frac{1}{\alpha} \\ 0, & \delta \geq \frac{1}{\alpha} \end{cases}$$

11.4. the source variable $X \sim \mathrm{N}(0, \sqrt{2})$ is quantised with an 3-bit linear quantiser where the quantisation limits are given by the integers $\{-3, -2, -1, 0, 1, 2, 3\}$.

(a) If the reconstruction values are located at $\{-3.5, -2.5, -1.5, -0.5, 0.5, 1.5, 2.5, 3.5\}$, derive (numerically) the average distortion?

(b) Instead of the reconstruction levels above, define optimal levels. What is the average distortion for this case?

(c) Assume that the quantiser is followed by an optimal source code, what is the required number of bits per symbol?

11.5. In Example 11.4 it is shown that the optimal reconstruction levels for a 1-bit quantisation of a Gaussian variable, $X \sim \mathrm{N}(0, \sigma)$ are $\pm\sqrt{2/\pi}\sigma$. Derive the average distortion.

Chapter 12

Solutions

Chapter 1

Chapter 2

2.1. (a) From the problem we have

$$P_X(0) = P_X(1) = \frac{1}{2}$$
$$P_Y(0) = p, \quad P_Y(1) = 1 - p$$

Let $Z = X \oplus Y$. Then,

$$\begin{aligned} P_Z(0) &= P(X \oplus Y = 0) = P(X \oplus Y = 0|Y = 0)P_Y(0) + P(X \oplus Y = 0|Y = 1)P_Y(1) \\ &= P_X(0)P_Y(0) + P_X(1)P_Y(1) = \frac{1}{2}p + \frac{1}{2}(1-p) = \frac{1}{2} \\ P_Z(1) &= P(X \oplus Y = 1) = P(X \oplus Y = 1|Y = 0)P_Y(0) + P(X \oplus Y = 1|Y = 1)P_Y(1) \\ &= P_X(1)P_Y(0) + P_X(0)P_Y(1) = \frac{1}{2}p + \frac{1}{2}(1-p) = \frac{1}{2} \end{aligned}$$

where we see that Z is independent of p.

(b) Similar to (a) let

$$P_X(i) = \frac{1}{M},\ i = 0, 1, \ldots, M-1$$
$$P_Y(i) = p_i,\ i = 0, 1, \ldots, M-1; \quad \text{where } \sum_i p_i = 1$$

Then, with $Z = \sum_i X + Y \mod M$, we get

$$\begin{aligned} P_Z(i) &= P(X + Y \equiv i \mod M) = \sum_j P(X + Y \equiv i \mod M|Y = j)P_Y(j) \\ &= \sum_j P(X = \langle i - j \rangle_M)P_Y(j) = \sum_j P_X(\langle i - j \rangle_M)P_Y(j) \\ &= \frac{1}{M}\sum_j P_Y(j) = \frac{1}{M} \end{aligned}$$

where $\langle k \rangle_M$ denotes the reminder when k is divided by M. This means that when a stochastic variable X is added by a uniformly distributed variable the statistical properties of X are "destroyed".

2.2. *Alternative 1.* The Let A and B be the event that the first and the second card, respectively, is not a heart. Then the the probability that the first card is not a heart is $P(A) = 3/4$. After that there are 51 cards left where 38 are not heart, hence $P(B|A) = 38/51$. The probability for not getting any heart becomes

$$P(A, B) = P(B|A)P(A) = \frac{38}{51} \cdot \frac{3}{4} = \frac{19}{34}$$

Alternative 2. Using combinatorial principles we use that the total number of cases of two cards taken from 52 are $\binom{52}{2}$. The number of pairs with no hearts are $\binom{39}{2}$. Hence, the probability is

$$P(A,B) = \frac{\binom{39}{2}}{\binom{52}{2}} = \frac{\frac{39!}{2!37!}}{\frac{52!}{2!50!}} = \frac{39 \cdot 38}{52 \cdot 51} = \frac{19}{34}$$

2.3. In this problem we have two alternative solutions. First define X as the number of heads for person 1 and Y as the number of heads for person 2. In the first alternative, consider the probability and expand it into something we can derive,

$$\begin{aligned} P(X=Y) &= \sum_{k=0}^{n} P(X=Y|Y=k)P(Y=k) \\ &= \sum_{k=0}^{n} P(X=k)P(Y=k) = \sum_{k=0}^{n} \binom{n}{k}\left(\frac{1}{2}\right)^n \binom{n}{k}\left(\frac{1}{2}\right)^n \\ &= \frac{\sum_{k=0}^{n}\binom{n}{k}^2}{2^{2n}} = \frac{\sum_{k=0}^{n}\binom{n}{k}^2}{\left(\sum_{k=0}^{n}\binom{n}{k}\right)^2} = \frac{\binom{2n}{n}}{\sum_{k=0}^{2n}\binom{2n}{k}} \end{aligned}$$

where the last two equalities are alternative ways of writing, and we used that $\sum_k \binom{n}{k} = 2^n$ and $\sum_k \binom{n}{k}^2 = \binom{2n}{n}$.

In the second alternative, consider the total number of favourable cases related to the total number of cases. There are 2^n different binary vectors (results of n flips) resulting in a total of $2^n 2^n = 2^{2n}$ different outcomes of $2n$ tosses. Among those we need to find the total number of favourable cases. If both persons have k heads they both have $\binom{n}{k}$ different outcomes. So, in total we have $\sum_k \binom{n}{k}^2$ favourable outcomes. Therefore, we get the same result as above from

$$P(X=Y) = \frac{\text{nbr favourable cases}}{\text{nbr cases}} = \frac{\sum_{k=0}^{n}\binom{n}{k}^2}{2^{2n}}$$

2.4. Choose $\lambda = \frac{1}{2}$ and $f(\cdot) = \log(\cdot)$ to get

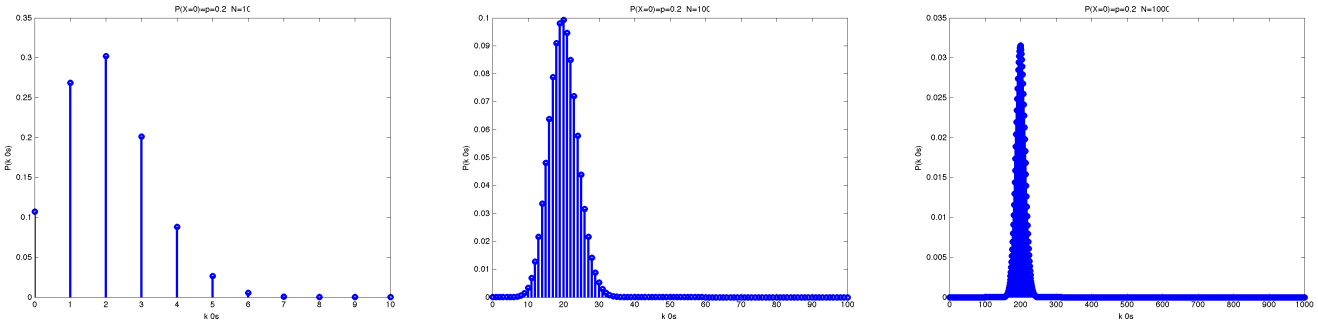
$$\begin{aligned} \frac{1}{2}\log x_1 + \frac{1}{2}\log x_2 &\leq \log\left(\frac{1}{2}x_1 + \frac{1}{2}x_2\right) \\ \log(x_1 x_2)^{\frac{1}{2}} &\leq \log\frac{x_1+x_2}{2} \\ (x_1 x_2)^{\frac{1}{2}} &\leq \frac{x_1+x_2}{2} \end{aligned}$$

where we in the last step used that the exponential function is an increasing function.

2.5. Let k be the number of 0s in the vector. Then we get the following table.

k	$\binom{10}{k}$	$P(\boldsymbol{x}\|k\ 0s) = p^k(1-p)^{(10-k)}$	$P(k\ 0s) = \binom{10}{k}P(\boldsymbol{x}\|k\ 0s)$
0	1	0.1073741824	0.1073741824
1	10	0.0268435456	0.2684354560
2	45	0.0067108864	0.3019898880
3	120	0.0016777216	0.2013265920
4	210	0.0004194304	0.0880803840
5	252	0.0001048576	0.0264241152
6	210	0.0000262144	0.0055050240
7	120	0.0000065536	0.0007864320
8	45	0.0000016384	0.0000737280
9	10	0.0000004096	0.0000040960
10	1	0.0000001024	0.0000001024

If we plot the probability for the distribution och 0s we get the picture below. We see here that the most probable *type* of sequence is not the one with only ones. The other two pictures below shows the same probability but with $N = 100$ and $N = 1000$. There we see even clearer that, with high probability, it is only a small group of sequences that will happen.



Chapter 3

3.1. Consider the function $f(x) = x - 1 - \log_b x$. For small x it will be dominated by $-\log_b x$ and for large x by x. So in both cases it will tend to infinity. Furthermore, for $x = 1$ the function will be $f(1) = 0$. The derivative of $f(x)$ in $x = 1$ is

$$\frac{\partial}{\partial x} f(x)\Big|_{x=1} = 1 - \frac{1}{x \ln b}\Big|_{x=1} = 1 - \frac{1}{\ln b} \begin{cases} < 0, & b < e \\ = 0, & b = e \\ > 0, & b > e \end{cases}$$

So, when $b = e$ there is a minimum at $x = 1$, and since it is a convex function the inequality is true. On the other hand, for $b < e$ the derivative is negative and the function must be below zero just after $x = 1$, and for $b > e$ the derivative is positive and the function must be below zero just before $x = 1$.

3.2. Use that

$$\ln x \leq x - 1, \quad x \geq 0$$

with equality for $x = 1$. Since $\frac{1}{x}$ is positive if and only if x is positive we can rewrite it as

$$\ln \frac{1}{x} \leq \frac{1}{x} - 1$$

with equality when $\frac{1}{x} = 1$, or, equivalently when $x = 1$. Changing sign on both sides gives the desired inequality.

3.3. The possible outcomes of X and Y are given in the table below:

X	1	2	3	4	5	6
Y	O	E	O	E	O	E

(a) $I(X = x; Y = y) = \log \frac{p_{X|Y}(x|y)}{p_X(x)}$
$I(X = 2; Y = \text{Even}) = \log \frac{p_{X|Y}(2|\text{Even})}{p_X(2)} = \log \frac{\frac{1}{3}}{\frac{1}{6}} = 1$
$I(X = 3; Y = \text{Even}) = \log \frac{0}{\frac{1}{6}} = -\infty$
$I(X = 2 \text{ or } X = 3; Y = \text{Even}) = \log \frac{\frac{1}{3}}{\frac{2}{6}} = 0$

(b) $I(X = 4) = -\log p_X(4) = -\log \frac{1}{6} = \log 6$
$I(Y = 0) = -\log \frac{1}{2} = \log 2 = 1$

(c) $H(X) = -\sum_{i=1}^{6} p_X(x_i) \log p_X(x_i) = -\sum_{i=1}^{6} \frac{1}{6} \log \frac{1}{6} = -6(\frac{1}{6} \log \frac{1}{6}) = \log 6$
$H(X) = H(\frac{1}{2}, \frac{1}{2}) = \log 2 = 1$

(d) $H(X|Y) = \frac{1}{2}H(X|Y = \text{Even}) + \frac{1}{2}H(X|Y = \text{Odd}) = \frac{1}{2} \log 3 + \frac{1}{2} \log 3 = \log 3$
$H(Y|X) = \frac{1}{6}H(Y|X = 1) + \frac{1}{6}H(Y|X = 2) + ... + \frac{1}{6}H(Y|X = 6) = 0 + 0 + ... + 0 = 0$
$H(X, Y) = H(X) + H(Y|X) = \log 6$

(e) $I(X; Y) = H(Y) - H(Y|X) = H(Y) = 1$

3.4. (a) The probability function for the stochastic variable Y is:

y	2	3	4	5	6	7	8	9	10	11	12
$p(y)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

(b) $H(X_1) = H(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}) = \log 6$

$H(Y) = H(\frac{1}{36}, \frac{1}{36}, \frac{2}{36}, \frac{2}{36}, \frac{3}{36}, \frac{3}{36}, \frac{4}{36}, \frac{4}{36}, \frac{5}{36}, \frac{5}{36}, \frac{6}{36}) \approx 3,2744$

(c) $I(Y; X_1) = H(Y) - H(Y|X_1) = H(Y) - H(X_2) \approx 3,2744 - \log 6 \approx 0,6894$

3.5. (a) The probability functions are:

$P(X)$	
0	$\frac{2}{3}$
1	$\frac{1}{3}$

$P(Y)$	
0	$\frac{1}{3}$
1	$\frac{2}{3}$

$P(X\|Y)$	$Y=0$	$Y=1$
$X=0$	1	$\frac{1}{2}$
$X=1$	0	$\frac{1}{2}$

$P(Y\|X)$	$Y=0$	$Y=1$
$X=0$	$\frac{1}{2}$	$\frac{1}{2}$
$X=1$	0	1

(b) $H(X) = h(\frac{1}{3}) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = \log 3 - \frac{2}{3}$

$H(Y) = h(\frac{1}{3}) = \log 3 - \frac{2}{3}$

(c) $H(X|Y) = P(Y=0)H(X|Y=0) + P(Y=1)H(X|Y=1) = \frac{1}{3}h(1) + \frac{2}{3}h(\frac{1}{2}) = \frac{2}{3}$

$H(Y|X) = P(X=0)H(Y|X=0) + P(X=1)H(Y|X=1) = \frac{2}{3}h(\frac{1}{2}) + \frac{1}{3}h(1) = \frac{2}{3}$

(d) $H(X,Y) = H(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0) = \log 3$

(e) $I(X;Y) = H(X) + H(Y) - H(X,Y) = \log 3 - \frac{2}{3} + \log 3 - \frac{2}{3} - \log 3 = \log 3 - \frac{4}{3}$

3.6. (a) The probability functions are:

$P(X)$	
A	$\frac{1}{12} + \frac{1}{6} = \frac{1}{4}$
B	$\frac{5}{45} + \frac{9}{45} = \frac{14}{45}$
C	$\frac{1}{18} + \frac{1}{4} + \frac{2}{15} = \frac{79}{180}$

$P(Y)$	
a	$\frac{1}{12} + \frac{1}{18} = \frac{5}{36}$
b	$\frac{1}{6} + \frac{1}{9} + \frac{1}{4} = \frac{19}{36}$
c	$\frac{1}{5} + \frac{2}{15} = \frac{1}{3}$

$P(X\|Y)$	a	b	c
A	$\frac{3}{5}$	$\frac{6}{19}$	0
B	0	$\frac{4}{19}$	$\frac{3}{5}$
C	$\frac{2}{5}$	$\frac{9}{19}$	$\frac{2}{5}$

$P(Y\|X)$	a	b	c
A	$\frac{1}{3}$	$\frac{2}{3}$	0
B	0	$\frac{5}{14}$	$\frac{9}{14}$
C	$\frac{10}{79}$	$\frac{45}{79}$	$\frac{24}{79}$

(b) $H(X) = H(\frac{1}{4}, \frac{14}{45}, \frac{79}{180}) \approx 1,5455$

$H(Y) = H(\frac{1}{3}, \frac{5}{36}, \frac{19}{36}) \approx 1,4105$

(c) $H(X|Y) = \sum_{i=1}^{3} P(Y=y_i)H(X|Y=y_i) \approx 1,2549$

$H(Y|X) = \sum_{i=1}^{3} P(X=x_i)H(Y|X=x_i) \approx 1,1199$

(d) $H(X,Y) = H(\frac{1}{12}, \frac{1}{6}, \frac{1}{9}, \frac{1}{5}, \frac{1}{18}, \frac{1}{4}, \frac{2}{15}) \approx 2,6654$

(e) $I(X;Y) = H(X) + H(Y) - H(X,Y) \approx 1,5455 + 1,4105 - 2,6654 \approx 0,2906$

3.7. (a)

X	$P(X)$
0	$\frac{7}{12}$
1	$\frac{5}{12}$

Y	$P(Y)$
a	$\frac{1}{3}$
b	$\frac{1}{6}$
c	$\frac{1}{2}$

$P(X\|Y)$	a	b	c
0	$\frac{1}{4}$	1	$\frac{2}{3}$
1	$\frac{3}{4}$	0	$\frac{1}{3}$

$P(Y\|X)$	a	b	c
0	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{4}{7}$
1	$\frac{3}{5}$	0	$\frac{2}{5}$

(b) $H(X) \approx 0.9799$ and $H(Y) \approx 1.4591$

(c) $H(X|Y) \approx 0.7296$ and $H(Y|X) \approx 1.2089$

(d) $H(X,Y) \approx 2.1887$

(e) $I(X;Y) \approx 0.2503$

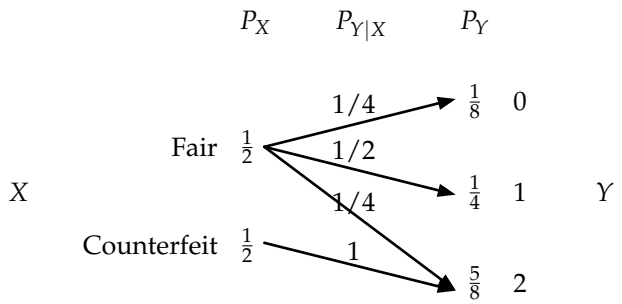


Figure 10: Probabilities or two flips with unknown coin.

3.8. Let X be the choice of coin where $P(\text{fair}) = P(\text{counterfeit}) = \frac{1}{2}$, and let Y be the number heads in two flips. The probabilities involved can be described as in Figure 10.

Hence,

$$H(Y) = H(\frac{1}{8}, \frac{1}{4}, \frac{5}{8}) = \frac{11}{4} - \frac{5}{8}\log 5$$

$$\begin{aligned} H(Y|X) &= H(Y|X = \text{fair})P(X = \text{fair}) + H(Y|X = \text{c.f.})P(X = \text{c.f.}) \\ &= \frac{1}{2}H(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}) + \frac{1}{2}H(0, 0, 1) = \frac{3}{4} \end{aligned}$$

and we conclude that

$$I(X;Y) = H(Y) - H(Y|X) = \frac{11}{4} - \frac{5}{8}\log 5 - \frac{3}{4} = 2 - \frac{5}{8}\log 5$$

3.9. (a) Fair dice:
$H_F(X) = H(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}) = -6\frac{1}{6}\log\frac{1}{6} = \log 6 \approx 2,585$
Manipulated dice:
$H_M(X) = H(\frac{1}{14}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{5}{14}) = -\frac{4}{7}\log\frac{1}{7} + \frac{1}{14}\log 14 + \frac{5}{14}\log 14 + \frac{5}{14}\log 5 \approx 2,41$

(b) $D(p||q) = \sum_x p(x)\log\frac{p(x)}{q(x)} = \frac{1}{6}\log\frac{7}{3} + \frac{4}{6}\log\frac{7}{6} + \frac{1}{6}\log\frac{7}{15} \approx 0,169$

(c) $D(q||p) = \sum_x q(x)\log\frac{q(x)}{p(x)} = \frac{1}{14}\log\frac{3}{7} + \frac{4}{7}\log\frac{6}{7} + \frac{5}{14}\log\frac{15}{7} \approx 0,178$

3.10. (a) $P_X(n) = P(\text{tail})^{n-1}P(\text{tail}) = \left(\frac{1}{2}\right)^{n-1}\left(\frac{1}{2}\right) = \left(\frac{1}{2}\right)^n$

(b) $E[X] = \sum_{n=1}^{\infty} n\left(\frac{1}{2}\right)^n = \frac{1}{2}\sum_{n=0}^{\infty} n\left(\frac{1}{2}\right)^{n-1} = \frac{\frac{1}{2}}{(1-\frac{1}{2})^2} = 2$

(c) $$\begin{aligned} H(X) &= -\sum_{n=1}^{\infty} P_X(n)\log P_X(n) = -\sum_{n=1}^{\infty}\left(\frac{1}{2}\right)^n \log\left(\frac{1}{2}\right)^n \\ &= \log 2\sum_{n=1}^{\infty} n\left(\frac{1}{2}\right)^n = E[X] = 2 \end{aligned}$$

(d) (a) $P_X(n) = pq^{n-1}$

(b) $E[X] = \sum_{n=1}^{\infty} npq^{n-1} = p\sum_{n=0}^{\infty} nq^{n-1} = \frac{p}{(1-q)^2} = \frac{1}{p}$

(c) $$H(X) = -\sum_{n=1}^{\infty} pq^{n-1}\log pq^{n-1} = -\sum_{n=1}^{\infty} pq^{n-1}\log p - \sum_{n=1}^{\infty}(n-1)pq^{n-1}\log q$$
$$= -\log p\sum_{n=1}^{\infty} pq^{n-1} - q\log q\sum_{n=0}^{\infty} npq^{n-1} = -\log p - q\log qE[X]$$
$$= \frac{-p\log p - q\log q}{p} = \frac{h(p)}{p}$$

3.11. First use that the sum over all x and y equals 1,

$$\sum_{x,y} k^2 2^{-(x+y)} = k^2\sum_x 2^{-x}\sum_y 2^{-y} = k^2 2^2 = 1 \Rightarrow k = \frac{1}{2}$$

(a)
$$P(X<4, Y<4) = \sum_{x=0}^{3}\sum_{y=0}^{3}\frac{1}{4}2^{-(x+y)} = \frac{1}{4}\Big(\sum_{x=0}^{3}2^{-x}\Big)^2 = \frac{1}{4}\Big(\frac{1-\left(\frac{1}{2}\right)^4}{1-\frac{1}{2}}\Big)^2 = \Big(\frac{15}{16}\Big)^2$$

(b)
$$H(X,Y) = -\sum_{x,y}\frac{1}{4}2^{-(x+y)}\log\frac{1}{4}2^{-(x+y)} = -\sum_{x,y}\frac{1}{4}2^{-(x+y)}\Big(\log\frac{1}{4} - (x+y)\log 2\Big)$$
$$= 2 + \sum_{x,y}x\frac{1}{4}2^{-(x+y)} + \sum_{x,y}y\frac{1}{4}2^{-(x+y)} = 2 + 2\sum_x x\frac{1}{2}2^{-x}\underbrace{\sum_y\frac{1}{2}2^{-y}}_{=1}$$
$$= 2 + 2\underbrace{\sum_x x\frac{1}{2}2^{-x}}_{=1} = 4$$

(c) To start with derive the marginals as

$$p(x) = \sum_y\frac{1}{4}2^{-(x+y)} = \frac{1}{2}2^{-x}\sum_y\frac{1}{2}2^{-y} = \frac{1}{2}2^{-x}$$
$$p(y) = \cdots = \frac{1}{2}2^{-y}$$

Since $p(x)p(y) = \frac{1}{2}2^{-x}\frac{1}{2}2^{-y} = \left(\frac{1}{2}\right)^2 2^{-(x+y)} = p(x,y)$ the variables X and Y are independent, Thus,

$$H(X|Y) = H(X) = -\sum_x\frac{1}{2}2^{-x}\log\frac{1}{2}2^{-x}$$
$$= -\sum_x\frac{1}{2}2^{-x}\Big(\log\frac{1}{2} - x\log 2\Big) = \sum_x\frac{1}{2}2^{-x} + \sum_x x\frac{1}{2}2^{-x} = 1 + 1 = 2$$

3.12. (a)

(b)

3.13. With $p(x,y) = p(x)p(y|x)$ we get

$$\begin{aligned}
D\big(p(x,y)||q(x,y)\big) &= \sum_{x,y} p(x,y)\log\frac{p(x,y)}{q(x,y)} \\
&= \sum_{x,y} p(x,y)\log\frac{p(x)}{q(x)} + \sum_{x,y} p(x,y)\log\frac{p(y|x)}{q(y|x)} \\
&= \sum_{x} p(x)\log\frac{p(x)}{q(x)} + \sum_{x,y} p(x)p(y|x)\log\frac{p(y|x)}{q(y|x)} \\
&= D\big(p(x)||q(x)\big) + \sum_{x}\Big(\sum_{y} p(y|x)\log\frac{p(y|x)}{q(y|x)}\Big)p(x) \\
&= D\big(p(x)||q(x)\big) + \sum_{x} D\big(p(y|x)||q(y|x)\big)p(x)
\end{aligned}$$

which gives the first equality. The second is obtained similarly. If X and Y are independent we have $p(y|x) = p(y)$ and $q(y|x) = q(y)$ which will give the third equality.

3.14. To simplify notations, we use the expected value,

$$\begin{aligned}
H(p,q) &= E_p\big[-\log q(x)\big] = E_p\big[-\log q(x) + \log p(x) - \log p(x)\big] \\
&= E_p\Big[\log\frac{p(x)}{q(x)}\Big] + E_p\big[-\log p(x)\big] = D\big(p(x)||q(x)\big) - H_p(X)
\end{aligned}$$

3.15. (a) Since $\alpha+\beta+\gamma = 1$, we get $\frac{\beta}{1-\alpha} + \frac{\gamma}{1-\alpha} = 1$.

$$\begin{aligned}
H(\alpha,\beta,\gamma) &= -\alpha\log\alpha - \beta\log\beta - \gamma\log\gamma \\
&= -\alpha\log\alpha - (1-\alpha)\log(1-\alpha) + (1-\alpha)\log(1-\alpha) - \beta\log\beta - \gamma\log\gamma \\
&= h(\alpha) + (1-\alpha)\Big(\log(1-\alpha) - \frac{\beta}{1-\alpha}\beta\log\beta - \frac{\gamma}{1-\alpha}\log\gamma\Big) \\
&= h(\alpha) + (1-\alpha)\Big(\Big(\frac{\beta}{1-\alpha} + \frac{\gamma}{1-\alpha}\Big)\log(1-\alpha) - \frac{\beta}{1-\alpha}\beta\log\beta - \frac{\gamma}{1-\alpha}\log\gamma\Big) \\
&= h(\alpha) + (1-\alpha)\Big(-\frac{\beta}{1-\alpha}\log\frac{\beta}{1-\alpha} - \frac{\gamma}{1-\alpha}\log\frac{\gamma}{1-\alpha}\Big) \\
&= h(\alpha) + (1-\alpha)h\Big(\frac{\beta}{1-\alpha}\Big)
\end{aligned}$$

(b) Follow the same steps as i (a)

3.16. Let the outcome of X be W and B, for white and black respectively. Then the probabilities for X conditioned on the urn, Y is as in the following table. Furthermore, since the choice of urn are equally likely the joint probability is $p(x,y) = \frac{1}{2}p(x|y)$.

$p(x\|y)$	W	B
1	4/7	3/7
2	3/10	7/10

$p(x,y)$	W	B
1	2/7	3/14
2	3/20	7/20

(a) The distribution of X is given by $P(X = \mathsf{W}) = \frac{1}{2}\cdot\frac{4}{7} + \frac{1}{2}\cdot\frac{3}{10} = \frac{61}{140}$, and the entropy $H(X) = h\big(\frac{61}{140}\big) = 0.988$.

(b) The mutual information can be derived as

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = h\big(\tfrac{61}{140}\big) + h\big(\tfrac{1}{2}\big) - H\big(\tfrac{2}{7}, \tfrac{3}{14}, \tfrac{3}{20}, \tfrac{7}{20}\big) = 0.0548$$

(c) By adding one more urn ($Y = 3$) we get the following tables (with $p(x) = 1/3$)

$p(x\|y)$	W	B
1	4/7	3/7
2	3/10	7/10
3	1	0

$p(x,y)$	W	B
1	4/21	3/21
2	1/10	7/30
3	1/3	0

Hence, $P(X = \mathsf{W}) = \frac{131}{210}$ and $P(X = \mathsf{B}) = \frac{79}{210}$, and $H(X) = h(\frac{79}{210})$. The mutual information is

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = h(\tfrac{79}{210}) + \log 3 - H(\tfrac{4}{21}, \tfrac{3}{21}, \tfrac{1}{10}, \tfrac{7}{30}, \tfrac{1}{3}) = 0.3331$$

3.17.

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

3.18. (a) The Jeffrey's divergence is

$$\begin{aligned} D_J(p||q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} + \sum_x q(x) \log \frac{q(x)}{p(x)} \\ &= \sum_x p(x) \log \frac{p(x)}{q(x)} - q(x) \log \frac{p(x)}{q(x)} = \sum_x (p(x) - q(x)) \frac{p(x)}{q(x)} \end{aligned}$$

(b) The Jensen Shannon divergence is

$$\begin{aligned} D_{JS}(p||q) &= \frac{1}{2} \sum_x p(x) \log \frac{p(x)}{\frac{p(x)+q(x)}{2}} + \frac{1}{2} \sum_x q(x) \log \frac{q(x)}{\frac{p(x)+q(x)}{2}} \\ &= \frac{1}{2} \sum_x p(x) \log p(x) - \frac{1}{2} \sum_x p(x) \log \frac{p(x)+q(x)}{2} \\ &\quad + \frac{1}{2} \sum_x q(x) \log q(x) - \frac{1}{2} \sum_x q(x) \log \frac{p(x)+q(x)}{2} \\ &= -\frac{1}{2} H(p) - \frac{1}{2} H(q) - \sum_x \frac{p(x)+q(x)}{2} \log \frac{p(x)+q(x)}{2} \\ &= H\Big(\frac{p(x)+q(x)}{2}\Big) - \frac{H(p) - H(q)}{2} \end{aligned}$$

Since $\sum_x \frac{p(x)+q(x)}{2} = \frac{\sum_x p(x) + \sum_x q(x)}{2} = 1$, the fraction $\frac{p(x)+q(x)}{2}$ is a distribution.

3.19. Use that the relative entropy is non-negative and the IT-inequality to get

$$\begin{aligned} 0 \leq D(p||q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &\leq \sum_x p(x) \Big(\frac{p(x)}{q(x)} - 1 \Big) \log e \\ &= \Big(\sum_x \frac{p^2(x)}{q(x)} - 1 \Big) \log e \end{aligned}$$

This requires that $\big(\sum_x \frac{p^2(x)}{q(x)} - 1\big) \geq 0$ which gives the assumption. The equality is given by the IT-inequality if and only if $\frac{p(x)}{q(x)} = 1$, or equivalently, if and only if $p(x) = q(x)$.

3.20. (a) The transition matrix is

$$P = \begin{pmatrix} 3/4 & 1/4 & 0 \\ 0 & 1/2 & 1/2 \\ 1/4 & 0 & 3/4 \end{pmatrix}$$

The stationary distribution is found from

$$\boldsymbol{\mu} P = \boldsymbol{\mu}$$

$$\Rightarrow \begin{cases} -\frac{1}{4}\mu_1 + \frac{1}{4}\mu_3 = 0 \\ \frac{1}{4}\mu_1 - \frac{1}{2}\mu_2 = 0 \\ \frac{1}{2}\mu_2 - \frac{1}{4}\mu_3 = 0 \end{cases}$$

Together with $\sum_i \mu_i = 1$ we get $\mu_1 = \frac{2}{5}$, $\mu_2 = \frac{1}{5}$, $\mu_3 = \frac{2}{5}$.

(b) The entropy rate is

$$\begin{aligned} H_\infty(U) &= \sum_i \mu_i H(S_i) = \frac{2}{5} h\Big(\frac{1}{4}\Big) + \frac{1}{5} h\Big(\frac{1}{2}\Big) + \frac{2}{5} h\Big(\frac{1}{4}\Big) \\ &= \frac{4}{5}\Big(2 - \frac{3}{4}\log 3\Big) + \frac{1}{5} = \frac{1}{9} - \frac{3}{4}\log 3 \approx 0.8490 \end{aligned}$$

(c) $H\Big(\frac{2}{5}, \frac{1}{5}, \frac{2}{5}\Big) = -\frac{2}{5}\log\frac{2}{5} - \frac{1}{5}\log\frac{1}{5} - \frac{2}{5}\log\frac{2}{5} = \log 5 - \frac{4}{5} \approx 1.5219$
That is, we gain in uncertainty if we take into consideration the memory of the source.

3.21. (a) The travel route follows a Markov chain according to the probability matrix

$$\Pi = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Let $\boldsymbol{\mu} = \begin{pmatrix} \mu_0 & \mu_1 & \mu_2 & \mu_3 \end{pmatrix}$ be the stationary distribution. Then, the equation system $\boldsymbol{\mu}\Pi = \boldsymbol{\mu}$ together with the condition $\sum_i \mu_i = 1$ gives the solution

$$\boldsymbol{\mu} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{2}{9} & \frac{1}{9} \end{pmatrix}$$

which is the distribution of the islands.

(b) The minimum number of bits per code symbol is entropy rate,

$$H_\infty = \tfrac{1}{3}\log 3 + \tfrac{1}{3}\log 3 + \tfrac{2}{9}\log 2 + \tfrac{1}{9}\log 1 = \tfrac{2}{9} + \tfrac{2}{3}\log 3$$

3.22. (a) With $\boldsymbol{\mu} = (\mu_2, \mu_4, \mu_6, \mu_8)$ and

$$P = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

the steady stat solution to $\boldsymbol{\mu} P = \boldsymbol{\mu}$ gives $\boldsymbol{\mu} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Hence, the entropy rate becomes

$$H_\infty(X) = \sum_{i \text{ Even}} \mu_i H(S_i) = \sum_{i \text{ Even}} \frac{1}{4} h\big(\tfrac{1}{2}\big) = 1$$

(b) With $\boldsymbol{\mu} = (\mu_1, \mu_3, \mu_5, \mu_7, \mu_9)$ and

$$P = \begin{pmatrix} 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 1/4 & 1/4 & 0 & 1/4 & 1/4 \\ 0 & 1/2 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \end{pmatrix}$$

the steady stat solution to $\boldsymbol{\mu}P = \boldsymbol{\mu}$ gives $\boldsymbol{\mu} = (\frac{1}{6}, \frac{1}{6}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6})$. Hence, the entropy rate becomes

$$H_\infty(X) = \sum_{i \text{ Odd}} \mu_i H(X_2|X_1 = i) = 4\frac{1}{6}h\Big(\frac{1}{2}\Big) + \frac{1}{3}H\Big(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\Big) = \frac{4}{6} + \frac{1}{3}\log 4 = \frac{4}{3}$$

Alternatively, one can define a weighted graph with weights according to the matrix

$$\big[W_{ij}\big] = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Then, we know that the steady state distribution is

$$\boldsymbol{\mu} = \Big[\frac{W_i}{2W}\Big] = (\tfrac{2}{12} \quad \tfrac{2}{12} \quad \tfrac{4}{12} \quad \tfrac{2}{12} \quad \tfrac{2}{12}) = (\tfrac{1}{6} \quad \tfrac{1}{6} \quad \tfrac{1}{3} \quad \tfrac{1}{6} \quad \tfrac{1}{6})$$

and the entropy rate is

$$H_\infty(X) = \log 12 - H\Big(\frac{1}{6}, \frac{1}{6}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}\Big) = \log 3 + 2 - 4\frac{1}{6}\log 6 - \frac{1}{3}\log 3 = \frac{4}{3}$$

3.23. (a) Follows from Stirling's approximation.

(b) $\frac{1}{p^{np}q^{nq}} = 2^{-n\log p^p q^q} = 2^{n(-p\log p - q\log q)} = 2^{nh(p)}$.

(c) From $e^{\frac{1}{12n}} \leq e^{\frac{1}{12}} < \sqrt{2}$.

(d) Follows from Stirling's approximation.

(e) With $12npq \geq 9$ we can use $e^{-\frac{1}{12npq}} \geq e^{-\frac{1}{9}} > \frac{\sqrt{\pi}}{2}$.

Chapter 4

4.1. (a) No

(b) Yes

(c) Yes

4.2. (a) Start from the root and expand the tree until all the codewords are reached.

(b) $H(X) = H(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{10}, \frac{1}{10}) = 2,5219$
$E(L) = \frac{1}{5} + \frac{3}{5} + \frac{3}{5} + \frac{3}{5} + \frac{4}{10} + \frac{4}{10} = 2,8$

(c) Yes, since $H(X) \leq E\big[L\big] \leq H(X) + 1$.

4.3. (a) According to Kraft's inequality we get:
$2^{-1}+2^{-2}+2^{-3}+2^{-4}+2^{-5}=\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\frac{1}{32}=\frac{31}{32}<1$
The code evidently exist and one example is $0, 10, 110, 1110, 11110$.

(b) $2^{-2}+2^{-2}+2^{-3}+2^{-3}+2^{-4}+2^{-4}+2^{-5}+2^{-5}=\frac{15}{16}<1$
One example is $00, 01, 100, 101, 1100, 1101, 11100, 11101$.

(c) $2^{-2}+2^{-2}+2^{-2}+2^{-3}+2^{-4}+2^{-4}+2^{-4}+2^{-5}=\frac{35}{32}>1$
The code doesn't exist!

(d) $2^{-2}+2^{-3}+2^{-3}+2^{-3}+2^{-4}+2^{-5}+2^{-5}+2^{-5}=\frac{25}{32}<1$
The set $00, 010, 011, 100, 1010, 10110, 10111, 11000$ contains the codewords.

4.4. i For a tree with one leaf (i.e. only the root) the statement is true.

ii Assume that the statement is true for a tree with $n-1$ leaves, i.e. $n-1$ leaves gives $n-2$ inner nodes. In a tree with n leaves consider two siblings. Their parent node is an inner node in the tree with n leaves, but it can also be viewed as a leaf in a tree with $n-1$ leaves. Thus, by expanding one leaf in a tree with $n-1$ leaves there is one new inner new and one extra leaf, and the resulting tree has n leaves and $n-2+1=n-1$ inner nodes.

4.5. Let the ith codeword length be $l_i=\log\frac{1}{q(x_i)}$. The average codeword length becomes

$$\begin{aligned}L_q&=\sum_i p(x_i)\log\frac{1}{q(x_i)}=\sum_i p(x_i)\Big(\log\frac{1}{q(x_i)}+\log p(x_i)-\log p(x_i)\Big)\\&=\sum_i p(x_i)\log\frac{p(x_i)}{q(x_i)}-\sum_i p(x_i)\log p(x_i)=D(p||q)+L_p\end{aligned}$$

where L_p is the optimal codeword length.

The mutual information is

$$I(X;Y)=D(p(x,y)||p(x)p(y))$$

This can be interpreted as follows. Consider two parallel sequences $\boldsymbol{x}$ and $\boldsymbol{y}$. Let $L_x=E_{p(x)}[\log\frac{1}{p(x)}]$ and $L_y=E_{p(y)}[\log\frac{1}{p(y)}]$ be the average codeword lengths when encoded separately. This should be compared with the case when the sequences are vied as one sequence of pairs of symbols, encoded with the joint codeword length $L_{x,y}=E_{p(x,y)}[\log\frac{1}{p(x,y)}]$. Consider the sum of the sum of the individual codeword lengths to get

$$\begin{aligned}L_x+L_y&=\sum_x p(x)\log\frac{1}{p(x)}+\sum_y p(y)\log\frac{1}{p(y)}\\&=\sum_{x,y}p(x,y)\log\frac{1}{p(x)}+\sum_{x,y}p(x,y)\log\frac{1}{p(y)}\\&=\sum_{x,y}p(x,y)\log\frac{1}{p(x)p(y)}=\sum_{x,y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}+\sum_{x,y}p(x,y)\log\frac{1}{p(x,y)}\\&=D\Big(p(x,y)\Big|\Big|p(x)p(y)\Big)+L_{x,y}=L_{x,y}+I(X;Y)\end{aligned}$$

This shows that the mutual information is the gain, in bits per symbol, we can make from considering pairs of symbols instead of assuming they are independent.

For example, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are binary sequences where $x_i = y_i$, it is enough to encode one of the sequences. Then X and Y are equally distributed, $p(x) = p(y)$, and we get

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = \sum_{x,y} p(x|y)p(y) \log \frac{p(x|y)p(y)}{p(y)^2}$$
$$= \sum_{x} p(y) \log \frac{1}{p(y)} = L_y$$

where, in the second last equality, we used that $p(x|y) = 1$ if $x = y$ and $p(x|y) = 0$ if $x \neq y$. The above derivation tells that we can gain the same amount of bits that is needed to encode sequence $\boldsymbol{y}$.

4.6. Begin with the two least probable nodes and move towards the root of the tree in order to find the optimal code (Huffman code). One such code is $11, 101, 100, 01, 001, 000$ where the codeword 11 corresponds to the random variable x_1 and 000 corresponds to x_6. Now use the path length lemma to obtain $E(L) = 1 + 0,6 + 0,4 + 0,2 + 0,4 = 2,6$. This is clearly less than 2,8 so the code in Problem 4.2 is not optimal!

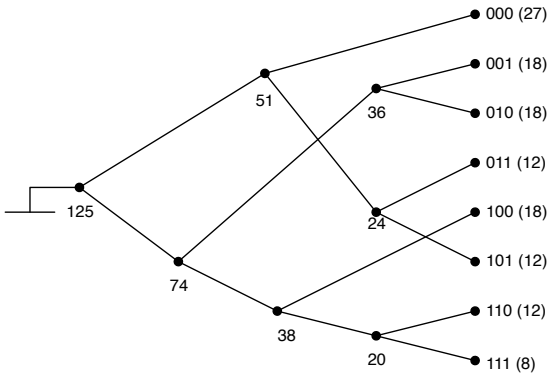
4.7. The optimal code is a Huffman code and one such example is $01, 11, 10, 001, 0001, 00001, 00000$ where the codeword 01 corresponds to the random variable x_1 and 00000 corresponds to x_7.

4.8. For the given code the probabilities and lengths of codewords is given by

$\boldsymbol{x}$	$p(\boldsymbol{x})$	$L(\boldsymbol{x})$
000	27/125	1
001	18/125	3
010	18/125	3
011	12/125	5

$\boldsymbol{x}$	$p(\boldsymbol{x})$	$L(\boldsymbol{x})$
100	18/125	3
101	12/125	5
110	12/125	5
111	8/125	5

Calculating the average codeword length gives $E[L] \approx 3.27$. Since it is more than the uncoded case the code is obviously not optimal. An optimal code can be constructed as a Huffman code. A tree is given below (labeled with the numerator of the probabilities):



The code table becomes

$\boldsymbol{x}$	$\boldsymbol{y}_H$
000	00
001	100
010	101
011	010

$\boldsymbol{x}$	$\boldsymbol{y}_H$
100	110
101	011
110	1110
111	1111

The average codeword length becomes $E[L_H] \approx 2.94$.

4.9. (a) Independent of how many nodes we start with, it will eventually come down to the case when there are four nodes left $\{x_1, x_2, x_3, x_4\}$, in which the nodes may consist of sub-trees. Assume then that x_1, with the highest probability has $p_1 > \frac{2}{5}$ and length $\ell_1 \geq 2$. Then the tree will be as in Figure 11(a). To form this tree first nodes x_3 and x_4 are merged in one subtree. Then the nodes x_1 and x_2 are merged. This second step can only be done if $p_1 \leq p_3 + p_4$. If this is not true the subtree x_3x_4 will merge with x_2 and ℓ_1 will not be 2. Thus, with the requirement that $p_1 > \frac{2}{5}$ we get $p_3 + p_4 > \frac{2}{5}$, and $p_2 \geq p_3 > \frac{1}{5}$. Then calculating the sum of the probabilities we get

$$p_1 + p_2 + p_3 + p_4 > \tfrac{2}{5} + \tfrac{1}{5} + \tfrac{2}{5} = 1$$

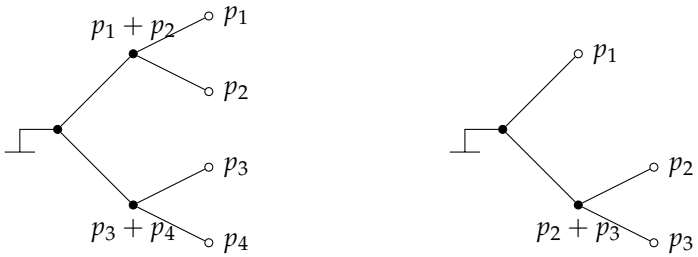
Since the sum is strictly more than 1 it is a contradiction, and we conclude that $\ell_1 < 2$, which gives the result.

In the beginning it was assumed four nodes but the actual requirement to show the contradiction comes in when there are three nodes, so the proof is valid for $n \geq 2$.

(b) Assume a Huffman tree consisting of three nodes $\{x_1, x_2, x_3\}$, where $p_1 < \frac{1}{3}$ and $\ell_1 = 1$, see Figure 11(b). From the problem formulation $p_1 \geq p_2 \geq p_3$, and we conclude that also $p_2 < \frac{1}{3}$ and $p_3 < \frac{1}{3}$. The sum of te probabilities then becomes

$$p_1 + p_2 + p_3 < \tfrac{1}{3} + \tfrac{1}{3} + \tfrac{1}{3} = 1$$

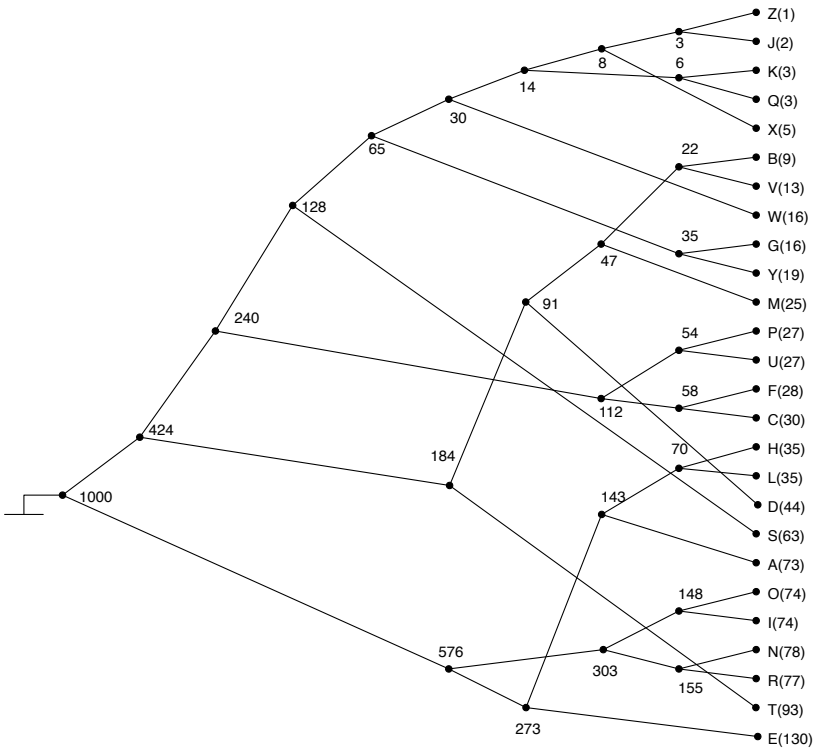
Since the sum is strictly less than 1 it is a contradiction, and we conclude that $\ell_1 \geq 1$.



(a) (a)

Figure 11: Huffman trees for Problem 4.9.

4.10. In the following tree the Huffman code of the English alphabet letters are constructed, and in the table below it is summarised as a code (reading the tree with 0 up and 1 down along the branches).



A	1101	F	00110	K	00000010	P	00100	U	00101
B	010000	G	000010	L	11001	Q	00000011	V	010001
C	00111	H	11000	M	01001	R	1011	W	000001
D	0101	I	1001	N	1010	S	0001	X	00000001
E	111	J	000000001	O	1000	T	011	Y	000011
								Z	000000000

The average codeword length becomes (from the path length lemma)

$$E[L] = \frac{1000+424+240+128+65+30+14+8+3+6+35+112+54+58+184+91+47+22+576+303+148+155+273+143+70}{1000}$$
$$\approx 4.189\text{bit/letter}$$

If all letters would have the same length it would require $\lceil \log 26 \rceil \approx 5\text{bit/letter}$. The entropy of the letters is

$$H(\tfrac{73}{1000}, \ldots, \tfrac{1}{1000}) \approx 4.162\text{bit/letter}$$

From the derivation we see that the Huffman code in this case is very close to the optimum compression, and that by using the code we gain approximately 0.8 bit per encoded letter compared to the case with equal length codewords.

4.11. (a) For the binary case an optimal code is given by a binary tree of depth 1, i.e.

X	0	1
Y	0	1

which gives average length $L_1 = 1$.

(b) For vectors of length 2, 3 and 4 the Huffman codes and average length per symbol is given by

X_2	P	Y
00	0.01	111
01	0.09	110
10	0.09	10
11	0.81	0

$\frac{1}{2}L_2 = \frac{1.2}{2} = 0.6$

X_3	P	Y
000	0.001	11111
001	0.009	11110
010	0.009	11101
011	0.081	110
100	0.009	11100
101	0.081	101
110	0.081	100
111	0.729	0

$\frac{1}{3}L_3 = \frac{1.589}{3} = 0.5297$

X_4	P	Y
0000	0.0001	1111111111
0001	0.0009	1111111110
0010	0.0009	111111110
0011	0.0081	1111110
0100	0.0009	111111101
0101	0.0081	111110
0110	0.0081	1111011
0111	0.0729	110
1000	0.0009	111111100
1001	0.0081	1111010
1010	0.0081	1111001
1011	0.0729	101
1100	0.0081	1111000
1101	0.0729	100
1110	0.0729	1110
1111	0.6561	0

$\frac{1}{4}L_4 = \frac{1.9702}{4} = 0.4925$

(c) The entropy is $H(X) = h(0.1) = 0.469$. Since the variables in the vectors are i.i.d. this is the optimal average length per symbol. In the above it is seen that already with a vector of length 4 the length is not so far away from this optima.

4.12. (a) In order to minimize the expected number of tastings we should start by tasting the one with the highest probability and then continue with same reasoning. The expected number of tastings becomes:
$1\frac{8}{23} + 2\frac{6}{23} + 3\frac{4}{23} + 4\frac{2}{23} + 5\frac{2}{23} + 5\frac{1}{23} \approx 2,39$

(b) The optimal solution is now a Huffman code. The codeword lengths are $2, 2, 2, 3, 4, 4$ (verify!). The expected number of tastings is $\approx 2,35$ (path length lemma). Here we should start with the mixture of the first and the second wine.

Chapter 5

5.1. The decoding procedure can be viewed in the following table. The colon in the *B*-buffer denotes the stop of the encoded letters for that codeword.

S-buffer	*B*-buffer	Codeword
`[IF IF =]`	`[ T:HEN T]`	(2,1,T)
`[ IF = T]`	`[H:EN THE]`	(0,0,H)
`[IF = TH]`	`[E:N THEN]`	(0,0,E)
`[F = THE]`	`[N: THEN ]`	(0,0,N)
`[ = THEN]`	`[ THEN TH:]`	(5,7,H)
`[THEN TH]`	`[EN =:  EL]`	(5,3,=)
`[ THEN =]`	`[ E:LSE E]`	(2,1,E)
`[HEN = E]`	`[L:SE ELS]`	(0,0,L)
`[EN = EL]`	`[S:E ELSE]`	(0,0,S)
`[N = ELS]`	`[E :ELSE ]`	(3,1,)
`[= ELSE ]`	`[ELSE ELS:]`	(5,7,S)
`[LSE ELS]`	`[E =:  IF ]`	(5,2,=)
`[ ELSE =]`	`[ I:F ]`	(2,1,I)
`[LSE = I]`	`[F:; ]`	(0,0,F)
`[SE = IF]`	`[;:  ]`	(0,0,;)

There are 15 codewords. In the uncoded text there are 45 letters, which corresponds to 360 bits. In the coded sequence we first have the buffer of 7 letters, which gives 56 bits. Then, each codeword requires $3 + 3 + 8 = 14$ bits. With 15 codewords we get $7 \cdot 8 + 15(3 + 3 + 8) = 266$ bits. The compression rate becomes $R = \frac{266}{360} = 0.7389$.

5.2. The decoding is done in the following table.

Index	Codeword	Dictionary (text)
1 :	$(0, t)$	t
2 :	$(0, i)$	i
3 :	$(0, m)$	m
4 :	$(0, \text{␣})$	␣
5 :	$(1, h)$	th
6 :	$(0, e)$	e
7 :	$(4, t)$	␣t
8 :	$(0, h)$	h
9 :	$(2, n)$	in
10 :	$(7, w)$	␣tw
11 :	$(9, \text{␣})$	in␣
12 :	$(1, i)$	ti
13 :	$(0, n)$	n
14 :	$(0, s)$	s
15 :	$(3, i)$	mi
16 :	$(5, .)$	th.

Hence, the text is “tim the thin twin tinsmith.”.

5.3. The decoding procedure can be viewed in the following table. The colon in the binary representation of the codeword shows where the index stops and the character code begins. This separator is not necessary in the final code string.

Index	Codeword	Dictionary	Binary
1	(0,I)	`[I]`	:01001001
2	(0,F)	`[F]`	0:01000110
3	(0,)	`[ ]`	00:00100000
4	(1,F)	`[IF]`	01:01000110
5	(3,=)	`[ =]`	011:00111101
6	(3,T)	`[ T]`	011:01010100
7	(0,H)	`[H]`	000:01001000
8	(0,E)	`[E]`	000:01000101
9	(0,N)	`[N]`	0000:01001110
10	(6,H)	`[ TH]`	0110:01001000
11	(8,N)	`[EN]`	1000:01001110
12	(10,E)	`[ THE]`	1010:01000101
13	(9,)	`[N ]`	1001:00100000
14	(0,=)	`[=]`	0000:00111101
15	(3,E)	`[ E]`	0011:01000101
16	(0,L)	`[L]`	0000:01001100
17	(0,S)	`[S]`	00000:01010011
18	(8,)	`[E ]`	01000:00100000
19	(8,L)	`[EL]`	01000:01001100
20	(17,E)	`[SE]`	10001:01000101
21	(15,L)	`[ EL]`	01111:01001100
22	(20,)	`[SE ]`	10100:00100000
23	(14,)	`[= ]`	01110:00100000
24	(4,;)	`[IF;]`	00100:00111011

In the uncoded text there are 45 letters, which corresponds to 360 bits. In the coded sequence there are in total $1 + 2 \cdot 2 + 4 \cdot 3 + 8 \cdot 4 + 8 \cdot 5 = 89$ bits for the indexes and $24 \cdot 8 = 192$ bits for the characters of the codewords. In total the code sequence is $89 + 192 = 281$ bits. The compression rate becomes $R = \frac{281}{360} = 0.7806$.

5.4. (a)

S-buffer	*B*-buffer	Codeword
`[Nat the ba]`	`[t s:]`	(8,2,s)
`[ the bat s]`	`[w:at]`	(0,0,w)
`[the bat sw]`	`[at a:]`	(5,3,a)
`[bat swat a]`	`[t M:]`	(3,2,M)
`[ swat at M]`	`[att:]`	(4,2,t)
`[at at Matt]`	`[ t:h]`	(5,1,t)
`[ at Matt t]`	`[h:e ]`	(0,0,h)
`[at Matt th]`	`[e:  g]`	(0,0,e)
`[t Matt the]`	`[ g:n]`	(4,1,g)
`[Matt the g]`	`[n:at]`	(0,0,n)
`[att the gn]`	`[at:]`	(10,1,t)

Text: 264 bits, Code: 234 bits, Rate:0.886364

(b)

S-buffer	*B*-buffer	Codeword
[Nat the ba]	[t :s]	(0,8,2)
[t the bat]	[s:wa]	(1,s)
[the bat s]	[w:at]	(1,w)
[the bat sw]	[at :]	(0,5,3)
[bat swat]	[at :]	(0,3,3)
[t swat at]	[M:at]	(1,M)
[swat at M]	[at:t]	(0,4,2)
[wat at Mat]	[t :t]	(0,5,2)
[t at Matt]	[t:he]	(0,2,1)
[at Matt t]	[h:e]	(1,h)
[at Matt th]	[e: g]	(1,e)
[t Matt the]	[:gn]	(0,4,1)
[Matt the]	[g:na]	(1,g)
[Matt the g]	[n:at]	(1,n)
[att the gn]	[at:]	(0,10,2)

Text: 264 bits, Code: 199 bits, Rate:0.7538

(c)

Index	Codeword	Dictionary	Binary
1	(0,N)	[N]	:01001110
2	(0,a)	[a]	0:01100001
3	(0,t)	[t]	00:01110100
4	(0,)	[ ]	00:00100000
5	(3,h)	[th]	011:01101000
6	(0,e)	[e]	000:01100101
7	(4,b)	[b]	100:01100010
8	(2,t)	[at]	010:01110100
9	(4,s)	[s]	0100:01110011
10	(0,w)	[w]	0000:01110111
11	(8,)	[at]	1000:00100000
12	(11,M)	[at M]	1011:01001101
13	(8,t)	[att]	1000:01110100
14	(4,t)	[t]	0100:01110100
15	(0,h)	[h]	0000:01101000
16	(6,)	[e]	0110:00100000
17	(0,g)	[g]	00000:01100111
18	(0,n)	[n]	00000:01101110
19	(2,t)	–	00010:01110100

Text: 264 bits, Code: 216 bits, Rate:0.8182

(d)

Index	Codeword	Dictionary	Binary
32		[ ]	
77		[M]	
78		[N]	
97		[a]	
98		[b]	
101		[e]	
103		[g]	
104		[h]	
110		[n]	
115		[s]	
116		[t]	
119		[w]	
256	78	[Na]	01001110
257	97	[at]	001100001
258	116	[t]	001110100
259	32	[t]	000100000
260	116	[th]	001110100
261	104	[he]	001101000
262	101	[e]	001100101
263	32	[b]	000100000
264	98	[ba]	001100010
265	257	[at]	100000001
266	32	[s]	000100000
267	115	[sw]	001110011
268	119	[wa]	001110111
269	265	[at a]	100001001
270	265	[at M]	100001001
271	77	[Ma]	001001101
272	257	[att]	100000001
273	258	[t t]	100000010
274	260	[the]	100000100
275	262	[e g]	100000110
276	103	[gn]	001100111
277	110	[na]	001101110
278	257	–	100000001

Text: 264 bits, Code: 206 bits, Rate:0.7803

5.5.

5.6. **Encoding**

step	lexicon	prefix	new symbol	codeword	
				(pointer,new symbol)	binary
0	∅	∅	T	(0,'T')	,01010100
1	T	∅	H	(0,'H')	0,01001000
2	H	∅	E	(0,'E')	00,01000101
3	E	∅	␣	(0,'␣')	00,00100000
4	␣	∅	F	(0,'F')	000,01000110
5	F	∅	R	(0,'R')	000,01010010
6	R	∅	I	(0,'I')	000,01001001
7	I	E	N	(3,'N')	011,01001110
8	EN	∅	D	(0,'D')	0000,01000100
9	D	␣	I	(4,'I')	0100,01001001
10	␣I	∅	N	(0,'N')	0000,01001110
11	N	␣	N	(4,'N')	0100,01001110
12	␣N	E	E	(3,'E')	0011,01000101
13	EE	D	␣	(9,'␣')	1001,00100000
14	D␣	I	S	(7,'S')	0111,01010011
15	IS	␣	T	(4,'T')	0100,01010100
16	␣T	H	E	(2,'E')	00010,01000101
17	HE	␣	F	(4,'F')	00100,01000110
18	␣F	R	I	(6,'I')	00110,01001001
19	RI	EN	D	(8,'D')	01000,01000100
20	END	␣I	N	(10,'N')	01010,01001110
21	␣IN	D	E	(9,'E')	01001,01000101
22	DE	E	D	(3,'D')	00011,01000100

The length of the code sequence is 268 bits. Assume that the source alphabet is ASCII, then the source sequence is of length 312 bits.

There are only ten different symbols in the sequence, therefore we can use a 10 letter alphabet, {T,H,E,-,F,R,I,N,D,S}. In that case we get $39 \cdot 4 = 156$ bits as the source sequence.

5.7. Let X describe the source, i.e. $P_X(0) = p$ and $P_X(0) = q = 1 - p$.

(a) Since nq might not be an integer we round it to $[nq]$. Then we say that a sequence of length n with a share of 1s equal to q has $[nq]$ 1s. There are $\binom{n}{[nq]}$ such sequences.

(b) To represent the sequences in (a) we need $\lfloor \log \binom{n}{[nq]} \rfloor$ bits (we use the lower integer limit $\lfloor \rfloor$ to achieve an integer number). Hence, in total we need $\frac{1}{n} \lfloor \log \binom{n}{[nq]} \rfloor$ bits/source bit.

(c) Use that $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ and $n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$ in the following derivation (approximating nq

and np as integers),

$$\begin{aligned}
\frac{1}{n}\log\binom{n}{nq} &= \frac{1}{n}\log\frac{n!}{nq!np!} = \frac{1}{n}\Big(\log n! - \log nq! - \log np!\Big)\\
&\approx \frac{1}{n}\Big(\log\sqrt{2\pi n}(\frac{n}{e})^n - \log\sqrt{2\pi nq}(\frac{nq}{e})^{nq}\Big) - \log\sqrt{2\pi np}(\frac{np}{e})^{np}\Big)\\
&= \frac{1}{n}\Big(\frac{1}{2}\log 2\pi + \frac{1}{2}\log n + n\log n - n\log e\\
&\qquad - \frac{1}{2}\log 2\pi - \frac{1}{2}\log nq - nq\log nq + nq\log e\\
&\qquad - \frac{1}{2}\log 2\pi - \frac{1}{2}\log np - np\log np + np\log e\Big)\\
&\approx \frac{1}{n}\Big(n\log n - qn\log qn - pn\log pn\Big)\\
&= \frac{1}{n}\Big(-qn\log q - qp\log p\Big) = h(q) = h(p)
\end{aligned}$$

where $\approx$ denotes approximations for large n and we used $\frac{1}{n}\log n \to 0,\ n \to \infty$. Notice, that this imply that for large n we get $\binom{n}{k} = \binom{n}{\frac{k}{n}n} \approx 2^{nh(\frac{k}{n})}$

Chapter 6

6.1. The definition of jointly typical sequences can be rewritten as

$$2^{-n(H(X,Y)+\varepsilon)} \le p(x,y) \le 2^{-n(H(X,Y)-\varepsilon)}$$

and

$$2^{-n(H(Y)+\varepsilon)} \le p(y) \le 2^{-n(H(Y)-\varepsilon)}$$

Dividing these and using the chain rule concludes the proof.

6.2. A binary sequence $\boldsymbol{x}$ of length 100 with k 1s has the probability

$$P(\boldsymbol{x}) = \Big(\frac{49}{50}\Big)^{100-k}\Big(\frac{1}{50}\Big)^{k} = \frac{49^{100-k}}{50^{100}}$$

(a) The most likely sequence is clearly the all-zero sequence with probability

$$P(00\ldots0) = \Big(\frac{49}{50}\Big)^{100} \approx 0.1326$$

(b) By definition a sequence $\boldsymbol{x}$ is ϵ-typical if

$$2^{-n(H(X)+\epsilon)} \le P(\boldsymbol{x}) \le 2^{-n(H(X)-\epsilon)}$$

or, equivalently,

$$-\epsilon \le -\frac{1}{n}\log P(\boldsymbol{x}) - H(X) \le \epsilon$$

Here,

$$H(X) = h(\frac{1}{50}) = -\frac{1}{50}\log\frac{1}{50} - \frac{49}{50}\log\frac{49}{50} = \log 50 - \frac{49}{50}\log 49 = 1 + 2\log 5 - \frac{49}{25}\log 7$$

and, for the all-zero sequence,

$$-\frac{1}{100}\log P(00\ldots0)) = -\frac{1}{100}\log\Big(\frac{49}{50}\Big)^{100} = -\log 49 + \log 50 = 1 + 2\log 5 - 2\log 7$$

Thus, we get

$$-\frac{1}{n}\log P(\boldsymbol{x}) - H(X) = 1 + 2\log 5 - 2\log 7 - 1 - 2\log 5 + \frac{49}{25}\log 7 = -\frac{1}{25}\log 7 < -\epsilon$$

and see that the all-zero sequence in not an ϵ-typical sequence.

(c) Consider again the condition for ϵ-typicality and derive

$$\begin{aligned}-\frac{1}{n}\log P(\boldsymbol{x}) - H(X) &= \frac{1}{100}\log\frac{49^{100-k}}{50^{100}} + \frac{1}{50}\log\frac{1}{50} + \frac{49}{50}\log\frac{49}{50}\\ &= \log 50 - \frac{100-k}{50}\log 7 - \log 50 + \frac{49}{25}\log 7 = -\frac{2-k}{50}\log 7\end{aligned}$$

Hence, for ϵ-typical sequences

$$\begin{aligned}-\frac{1}{50}\log 7 \le -\frac{2-k}{50}\log 7 &\le \frac{1}{50}\log 7\\ -1 \le k-2 &\le 1\\ 1 \le k &\le 3\end{aligned}$$

So, the number of ϵ-typical sequences is

$$\binom{100}{1} + \binom{100}{2} + \binom{100}{3} = 166750$$

which should be compared with the total number of sequences $2^{100} \approx 1.2677 \cdot 10^{30}$.

6.3. Consider a sequence of n cuts and let $\boldsymbol{x} = x_1x_2\ldots x_n$ be the the outcome where x_i is the part saved in cut i. If in k of the cuts we save the long part and in $n-k$ the short part, the length becomes $L_k = (\frac{2}{3})^k(\frac{1}{3})^{(n-k)} = \frac{2^k}{3^n}$. The probability for such a sequence is $P(\boldsymbol{x}) = (\frac{3}{4})^k(\frac{1}{4})^{(n-k)} = \frac{3^k}{4^n}$. On the other hand we know that the most probable sequences are the typical, represented by the set $A_\varepsilon(\boldsymbol{X})$. Hence, if we consider a typical sequence we know that the probability is bounded by

$$2^{-n(H(X)+\varepsilon)} \le P(\boldsymbol{x}) \le 2^{-n(H(X)-\varepsilon)}$$

To the first order of the exponent (assume ε very small), this gives that $P(\boldsymbol{x}) = 2^{-nH(X)}$, where $H(X) = h(\frac{1}{4})$. Combining the two expressions for the probability gives

$$3^k = 2^{2n}\cdot 2^{-nh(\frac{1}{4})} = 2^{n(2-h(\frac{1}{4}))}$$

or, equivalently,

$$k = n\frac{2-h(\frac{1}{4})}{\log 3} = n\frac{2+\frac{1}{4}\log\frac{1}{4}+\frac{3}{4}\log\frac{1}{4}+\frac{3}{4}\log 3}{\log 3} = n\frac{3}{4}$$

Going back to the remaining length we get

$$L_k = \frac{2^{n\frac{3}{4}}}{3^n} = \Big(\frac{2^{\frac{3}{4}}}{3}\Big)^n$$

and we conclude that, in average, we keep $\frac{2^{3/4}}{3}$ of the length at each cut.

6.4. (a) According to the data processing inequality we have that $I(X;Y) \geq I(X;\tilde{Y})$, where $X \to Y \to \tilde{Y}$ forms a Markov chain. Now if $\tilde{p}(x)$ maximizes $I(X;\tilde{Y})$ we have that

$$C = \max_{p(x)} I(X;Y) \geq I(X;Y)_{p(x)=\tilde{p}(x)} \geq I(X;\tilde{Y})_{p(x)=\tilde{p}(x)} = \max_{p(x)} I(X;\tilde{Y}) = \tilde{C}$$

(b) The capacity is not decreased only if we have equality in the data processing inequality, that is when $X \to \tilde{Y} \to Y$ forms the Markov chain.

6.5. (a) Since X and Z independent $H(Y|X) = H(Z|X) = H(Z) = \log 3$. The capacity becomes

$$C = \max_{p(x)} I(X;Y) = \max_{p(x)} H(Y) - \log 3 = \log 11 - \log 3 = \log \frac{11}{3}$$

(b) This is achieved for uniform Y which by symmetry is achieved for uniform X, i.e. $p(x_i) = \frac{1}{11} \quad \forall i$.

6.6. Assume that $P(X=1) = p$ and $P(X=0) = 1-p$. Then

$$\begin{cases} P(Y=1) = P(X=1)P(Z=1) = \alpha p \\ P(Y=0) = 1 - P(Y=1) = 1 - \alpha p \end{cases}$$

Then

$$I(X;Y) = H(Y) - H(Y|X) = h(\alpha p) - ((1-p)h(1) + ph(\alpha)) = h(\alpha p) - ph(\alpha)$$

Differentiating with respect to p gives us the maximising $\tilde{p} = \frac{1}{\alpha(2^{\frac{h(\alpha)}{\alpha}}+1)}$. The capacity is

$$C = h(\alpha\tilde{p}) - \tilde{p}h(\alpha) = \cdots = \log(2^{\frac{h(\alpha)}{\alpha}} + 1) - \frac{h(\alpha)}{\alpha}$$

6.7. (a) $C = \log 4 - h(\frac{1}{2}) = 2 - 1 = 1$

(b) $C = \log 4 - H(\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}) \approx 0,0817$

(c) $C = \log 3 - H(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}) \approx 0,126$

6.8. By assuming that $P(X=0) = \pi$ and $P(X=1) = 1-\pi$ we get the following:

$$\begin{aligned} H(Y) &= H(\pi(1-p-q) + (1-\pi)p, \pi q + (1-\pi)q, (1-\pi)(1-p-q) + \pi p) \\ &= H(\pi - 2p\pi - q\pi + p, q, 1 - p - q - \pi + 2p\pi + q\pi) \\ &= h(q) + (1-q)H\left(\frac{\pi - 2p\pi - q\pi + p}{(1-q)}, \frac{1-p-q-\pi+2p\pi+q\pi}{(1-q)}\right) \leq h(q) + (1-q) \end{aligned}$$

with equality if $\pi = \frac{1}{2}$, where $H(\frac{1}{2}, \frac{1}{2}) = 1$.

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) = \max_{p(x)} (H(Y) - H(Y|X)) = h(q) + (1-q) - H(p, q, 1-p-q) \\ &= (1-q)\left(1 - H\left(\frac{1-p-q}{1-q}, \frac{p}{1-q}\right)\right) \end{aligned}$$

6.9. Assume that $P(X=0) = 1-A$ and $P(X=1) = A$. Then

$$H(Y) = H\left((1-A) + \frac{A}{2}, \frac{A}{2}\right) = H\left(1 - \frac{A}{2}, \frac{A}{2}\right) = h(\frac{A}{2})$$

$$H(Y|X) = P(X=0)H(Y|X=0) + P(X=1)H(Y|X=1) = Ah(\frac{1}{2}) = A$$

and we conclude

$$C = \max_{p(x)} \Big\{ h(\tfrac{A}{2}) - A \Big\}$$

Differentiation with respect to A gives the optimal $\tilde{A} = \frac{2}{5}$.

$$C = h(\tfrac{\tilde{A}}{2}) - \tilde{A} \approx 0,322$$

6.10. By cascading two BSCs we get the following probabilities:

$$\begin{aligned}
&P(Z=0|X=0) = (1-p)^2 + p^2 \\
&P(Z=1|X=0) = p(1-p) + (1-p)p = 2p(1-p) \\
&P(Z=0|X=1) = 2p(1-p) \\
&P(Z=1|X=1) = (1-p)^2 + p^2
\end{aligned}$$

This channel can be seen as a new BSC with crossover probability $\epsilon = 2p(1-p)$. The capacity for this channel becomes $C = 1 - h(\epsilon) = 1 - h(2p(1-p))$.

6.11. (a) The channel is weakly symmetric, so we can directly state the capacity as

$$C = \log 4 - H(\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{4}, 0) = 2 - \tfrac{3}{2} = \tfrac{1}{2}$$

(b) By letting $P(X=0) = \frac{1}{6}$ and $P(X=1) = \frac{5}{6}$, the probabilities for the received symbols are $P(A) = \frac{1}{12}$, $P(B) = \frac{1}{4}$, $P(C) = \frac{1}{4}$ and $P(D) = \frac{5}{12}$. An optimal compression code is given by the following Huffman code.

Y	Z
A	000
B	001
C	01
D	1

which gives the average length $L = 1.917$ bit. As a comparison the entropy is $H(\frac{1}{12}, \frac{1}{4}, \frac{1}{4}, \frac{5}{12}) = 1.825$ bit.

6.12. The overall channel has the probabilities

$$\begin{aligned}
&P(Z=0|X=0) = (1-\alpha)(1-\beta) && P(Z=1|X=1) = (1-\alpha)(1-\beta) \\
&P(Z=\Delta|X=0) = (1-\alpha)\beta + \alpha\beta = \beta && P(Z=\Delta|X=0) = \beta \\
&P(Z=1|X=0) = \alpha(1-\beta) && P(Z=0|X=1) = \alpha(1-\beta)
\end{aligned}$$

Identifying with the channel model in Problem 6.8 with $p = \alpha(1-\beta)$ and $q = \beta$, the capacity follows from the solution.

6.13. Denote $P(X=0) = p$. Then the joint probability and the probability for Y is given by

	$P(X\|Y)$	$Y=0$	$Y=1$
X	0	p	0
	1	$(1-p)\alpha$	$((1-p)(1-\alpha))$
	$P(Y)$:	$p + (1-p)\alpha$	$(1-p)(1-\alpha)$
		$= 1 - (1-p)(1-\alpha)$	

The conditional and unconditional entropies of Y are then given by

$$H(Y|X) = H(Y|X=0)p + H(Y|X=1)(1-p) = (1-p)h(\alpha)$$
$$H(Y) = h\big((1-p)(1-\alpha)\big)$$

By using $\frac{d}{dx}h(x) = \log\frac{1-x}{x}$ the derivative of the mutual information is

$$\frac{d}{dp}I(X;Y) = \frac{d}{dp}H(Y) - H(Y|X) = \frac{d}{dp}h\big((1-p)(1-\alpha)\big) - (1-p)h(\alpha)$$
$$= -(1-\alpha)\log\frac{1-(1-p)(1-\alpha)}{(1-p)(1-\alpha)} + h(\alpha) = 0$$

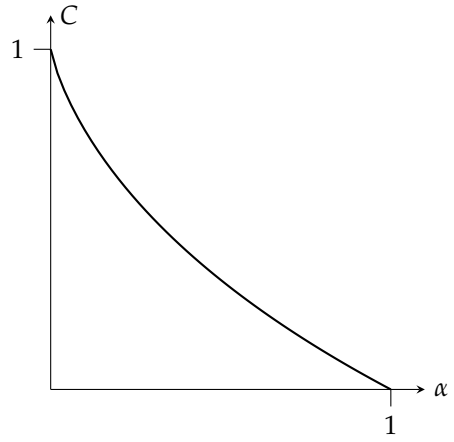
which gives

$$1-p = \frac{1}{(1-\alpha)\big(1+2^{\frac{h(\alpha)}{1-\alpha}}\big)}$$

Inserting to the mutual information gives

$$C = h\left(\frac{1}{\big(1+2^{\frac{h(\alpha)}{1-\alpha}}\big)}\right) - \frac{\frac{h(\alpha)}{1-\alpha}}{1+2^{\frac{h(\alpha)}{1-\alpha}}}$$

Here the value for $\alpha \to 1$ becomes a limit value which can be found as $C \to 0$. Then the capacity can be plotted as a function of α as shown here to the right.



6.14. (a) The mutual information between X and Y is

$$I(X;Y) = H(Y) - H(Y|X)$$
$$= H(Y) - \sum_{i=0}^{1} H(Y|x=i)P(x=i) = H(Y) - H(\alpha_0,\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5)$$

So, what is left to optimize is $H(Y)$. From the probability table we see that $p(y=j|x=0) = \alpha_j$ and $p(y=j|x=1) = \alpha_{5-j}$. If we assume that the probability of X is given by $p(x=0) = p$ and $p(x=1) = 1-p$, then the joint probability is given by $p(y=j, x=0) = p\alpha_j$ and $p(y=j, x=1) = (1-p)\alpha_{5-j}$. Hence, we can write the probability for Y as $p(y=j) = p\alpha_j + (1-p)\alpha_{5-j}$ and the entropy as

$$H(Y) = \sum_{j=0}^{5}(p\alpha_j + (1-p)\alpha_{5-j})\log(p\alpha_j + (1-p)\alpha_{5-j})$$

The corresponding derivative with respect to p is

$$\frac{\partial}{\partial p}H(Y) = \sum_{j=0}^{5}(\alpha_j - \alpha_{5-j})\Big(\log(p\alpha_j + (1-p)\alpha_{5-j}) + \frac{1}{\ln 2}\Big)$$

Then, setting $p = \frac{1}{2}$ and splitting in two sums we get

$$\frac{\partial}{\partial p}H(Y)\Big|_{p=\frac{1}{2}} = \sum_{j=0}^{2}(\alpha_j - \alpha_{5-j})\Big(\frac{1}{2\ln 2} + \log(\alpha_j + \alpha_{5-j})\Big)$$
$$+ \sum_{j=3}^{5}(\alpha_j - \alpha_{5-j})\Big(\frac{1}{2\ln 2} + \log(\alpha_j + \alpha_{5-j})\Big)$$

In the second sum replace the summation variable with $n = 5 - j$, then

$$\frac{\partial}{\partial p} H(Y)\Big|_{p=\frac{1}{2}} = \sum_{j=0}^{2} (\alpha_j - \alpha_{5-j}) \Big(\frac{1}{2 \ln 2} + \log(\alpha_j + \alpha_{5-j}) \Big) + \sum_{n=0}^{2} (\alpha_{5-n} - \alpha_n) \Big(\frac{1}{2 \ln 2} + \log(\alpha_{5-n} + \alpha_n) \Big)$$

Since $(\alpha_{5-n} - \alpha_n) = -(\alpha_n - \alpha_{5-n})$ we get two identical sums with different sign,

$$\frac{\partial}{\partial p} H(Y)\Big|_{p=\frac{1}{2}} = \sum_{j=0}^{2} (\alpha_j - \alpha_{5-j}) \Big(\frac{1}{2 \ln 2} + \log(\alpha_j + \alpha_{5-j}) \Big) - \sum_{j=0}^{2} (\alpha_j - \alpha_{5-j}) \Big(\frac{1}{2 \ln 2} + \log(\alpha_j + \alpha_{5-j}) \Big) = 0$$

and we have seen that $p = \frac{1}{2}$ maximizes $H(Y)$. (Here the maximum follows from the fact that the entropy is a concave function.)
Then, for $p = \frac{1}{2}$, we get

$$\begin{aligned} H(Y) &= -\sum_{j=0}^{5} \frac{1}{2}(\alpha_j + \alpha_{5-j}) \log \frac{1}{2}(\alpha_j + \alpha_{5-j}) \\ &= \frac{1}{2} \sum_{j=0}^{5} (\alpha_j + \alpha_{5-j}) - \frac{1}{2} \sum_{j=0}^{5} (\alpha_j + \alpha_{5-j}) \log(\alpha_j + \alpha_{5-j}) \\ &= 1 - \frac{1}{2} \Big(\sum_{j=0}^{5} (\alpha_j + \alpha_{5-j}) \log(\alpha_j + \alpha_{5-j}) + \sum_{n=0}^{2} (\alpha_n + \alpha_{5-n}) \log(\alpha_n + \alpha_{5-n}) \Big) \\ &= 1 - \sum_{j=0}^{2} (\alpha_j + \alpha_{5-j}) \log(\alpha_j + \alpha_{5-j}) = 1 + H(\alpha_0 + \alpha_5, \alpha_1 + \alpha_4, \alpha_2 + \alpha_3) \end{aligned}$$

Hence, the capacity is

$$C_6 = 1 + H(\alpha_0 + \alpha_5, \alpha_1 + \alpha_4, \alpha_2 + \alpha_3) + H(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$$

(b) The right hand inequality is straight forward since

$$C_6 \leq I(X;Y) = H(X) - H(X|Y) \leq H(X) \leq \log |\mathcal{X}| = 1$$

For the left hand inequality we first derive the capacity for the corresponding BSC. The error probability is $p = \alpha_3 + \alpha_4 + \alpha_5$, hence,

$$C_{\text{BSC}} = 1 - h(p) = 1 - H(\alpha_0 + \alpha_1 + \alpha_2, \alpha_3 + \alpha_4 + \alpha_5)$$

So, to show that $C_{\text{BSC}} \leq C_6$ we should show that

$$\begin{aligned} C_6 - C_{\text{BSC}} = 1 &+ H(\alpha_0 + \alpha_5, \alpha_1 + \alpha_4, \alpha_2 + \alpha_3) + H(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) \\ &- 1 + H(\alpha_0 + \alpha_1 + \alpha_2, \alpha_3 + \alpha_4 + \alpha_5) \end{aligned}$$

is non-negative. For this we introduce a new pair of random variables A and B with the joint distribution and marginal distributions according to

$P(A,B)$		B = 0	1	2
A	0	α_0	α_1	α_2
	1	α_5	α_4	α_3

A	$P(A)$
0	$\alpha_0 + \alpha_1 + \alpha_2$
1	$\alpha_3 + \alpha_4 + \alpha_5$

B	$P(B)$
0	$\alpha_0 + \alpha_5$
1	$\alpha_1 + \alpha_4$
2	$\alpha_2 + \alpha_3$

Then we can identify in the capacity formula above

$$\begin{aligned} C_6 - C_{\text{BSC}} &= 1 + H(B) - H(A,B) - 1 + H(A) \\ &= H(A) + H(B) - H(A,B) = I(A;B) \geq 0 \end{aligned}$$

which is the desired result. (The above inequality can also be obtained from the IT-inequality).

Chapter 7

7.1. (a) $R = \frac{3}{6}$

(b) Find the codewords for $\boldsymbol{u}_1 = (100)$, $\boldsymbol{u}_2 = (010)$ and $\boldsymbol{u}_3 = (001)$ and form the generator matrix

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

(c) List all codewords

$\boldsymbol{u}$	$\boldsymbol{x}$
000	000000
001	001110
010	010101
011	011011

$\boldsymbol{u}$	$\boldsymbol{x}$
100	100011
101	101101
110	110110
111	111000

Then we get $d_{\min} = \min_{\boldsymbol{x} \neq \boldsymbol{0}} \{w_H(\boldsymbol{x})\} = 3$

(d) From part b we note that $G = (I \; P)$. Since

$$\begin{pmatrix} I & P \end{pmatrix} \begin{pmatrix} P \\ I \end{pmatrix} = P \oplus P = 0$$

we get

$$H = (P^T \; I) = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

(e) List the most probable error patterns

$\boldsymbol{e}$	$\boldsymbol{s} = \boldsymbol{e}H^T$
000000	000
100000	011
010000	101
001000	110
000100	100
000010	010
000001	001
100100	111

where the last row is one of the weight two vectors that gives the syndrom (111).

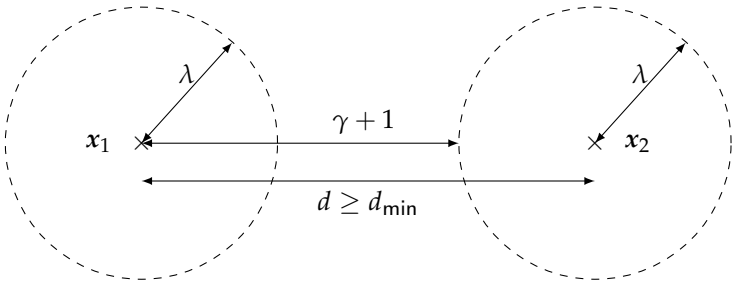
(f) One (correctable) error

$$
\begin{aligned}
&\boldsymbol{u} = 101 \\
\Rightarrow\ &\boldsymbol{x} = 101101 \\
&\boldsymbol{e} = 010000 \\
\Rightarrow\ &\boldsymbol{y} = \boldsymbol{x} \oplus \boldsymbol{e} = 111101 \\
\Rightarrow\ &\boldsymbol{s} = \boldsymbol{y}H^T = 101 \\
\Rightarrow\ &\hat{\boldsymbol{e}} = 010000 \\
\Rightarrow\ &\hat{\boldsymbol{x}} = \boldsymbol{y} \oplus \hat{\boldsymbol{e}} = 101101 \\
\Rightarrow\ &\hat{\boldsymbol{u}} = 101
\end{aligned}
$$

An uncorrectable error

$$
\begin{aligned}
&\boldsymbol{u} = 101 \\
\Rightarrow\ &\boldsymbol{x} = 101101 \\
&\boldsymbol{e} = 001100 \\
\Rightarrow\ &\boldsymbol{y} = \boldsymbol{x} \oplus \boldsymbol{e} = 100001 \\
\Rightarrow\ &\boldsymbol{s} = \boldsymbol{y}H^T = 010 \\
\Rightarrow\ &\hat{\boldsymbol{e}} = 000010 \\
\Rightarrow\ &\hat{\boldsymbol{x}} = \boldsymbol{y} \oplus \hat{\boldsymbol{e}} = 100011 \\
\Rightarrow\ &\hat{\boldsymbol{u}} = 100
\end{aligned}
$$

7.2. Consider the graphical interpretation of $\mathbb{F}_2^n$ and the two codewords $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.



A received symbol that is at Hamming distance at most λ from a codeword is corrected to that codeword. This is indicated by a sphere with radius λ around each codeword. Received symbols that lies outside a sphere are detected to be erroneous. The distance from one codeword to the sphere around another codeword is $\gamma + 1$, the number of detected errors, and the minimal distance between two codewords must be at least $\gamma + 1 + \lambda$. Hence, $d_{\min} \geq \lambda + \gamma + 1$.

7.3. (a) For the code to be linear the all-zero vector should be a codeword and the (position-wise) addition of any two codewords should again be a codeword. Since the all-zero vector is a codeword in $\mathcal{B}$ it is also a codeword in $\mathcal{B}_E$. To show that the addition of two codewords is again a codeword we need to show that the resulting vector has even weight. For this we use the position-wise AND function to get the positions in which both codewords have ones. Then if $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathcal{B}$ the weight of their sum can be written as

$$w_H(\boldsymbol{y}_1 + \boldsymbol{y}_2) = w_H(\boldsymbol{y}_1) + w_H(\boldsymbol{y}_2) - 2w_H(\boldsymbol{y}_1 \& \boldsymbol{y}_2)$$

here we notice that the first two terms are known to be even and the third term is also even since it contains the factor 2. Therefore the resulting vector is also even and we conclude that the code is even.

For the case when an extra bit is added such that the codeword has even weight the code is not linear since the all-zero vector is not a codeword.

(b) A vector $\boldsymbol{y} = (y_1 \dots y_{n+1})$ is a codeword iff $\boldsymbol{y}H_E^T = \boldsymbol{0}$. This gives

$$
\begin{aligned}
\boldsymbol{y}H_E^T &= (y_1 \dots y_n y_{n+1}) \begin{pmatrix} & & & 1 \\ & H^T & & \vdots \\ & & & 1 \\ 0 & \dots & 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} (y_1 \dots y_n)H^T & \sum_{i=1}^{n+1} y_i \end{pmatrix} = \boldsymbol{0}
\end{aligned}
$$

which gives the two conditions that $(y_1 \dots y_N) \in \mathcal{B}$ and that $w_H(y_1 \dots y_{n+1}) =$ even.

(c) Assume $\mathcal{B}$ has minimum distance d and $\mathcal{B}_E$ minimum distance d_E. If d is even then $d_E = d$, but if d odd then $d_E = d + 1$.

(d) $H = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$.

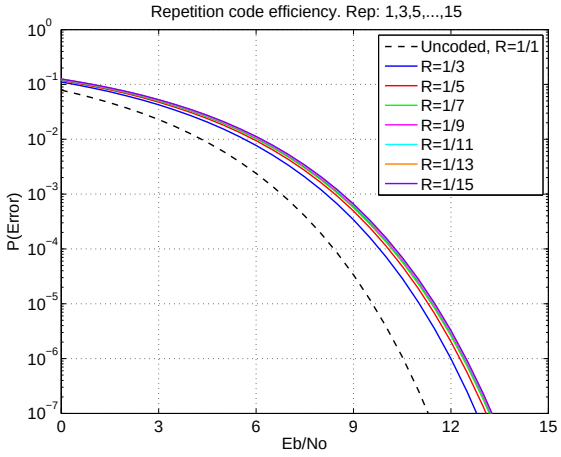
7.4. The error probability when transmitting one bit with energy E_b over a channel with Gaussian noise of level $N_0/2$ is $P_b = Q(\sqrt{2E_b/N_0})$. A repetition code with N repetitions gives the energy E_b/N per transmitted bit, and thus the error probability $P_{b,N} = Q(\sqrt{2E_b/N_0N})$. On the other hand, the redundancy of the code gives that it requires at least $i = \lceil N/2 \rceil$ errors in the codeword for the result to be erroneous. Since there are $\binom{N}{i}$ vectors with i errors, the total error probability becomes

$$P_{\text{error}} = \sum_{i=\lceil N/2 \rceil}^{N} \binom{N}{i} Q\left(\sqrt{2\frac{E_b/N}{N_0}}\right)$$

In MATLAB the Q-function can be derived from the erfc-function as

```
function Qfunc = Q(x)
Qfunc = 1/2*erfc(x/sqrt(2));
```

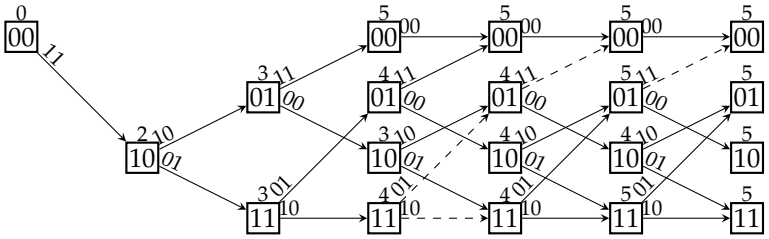
A plot of the results for $N = 3, 5, \ldots, 15$ is shown in the figure below.



Note: The fact that the error probability actually gets worse by using the repetition code might come as a surprise, especially since it is a standard example of a error correcting code. But, what the code actually does is that it prolongs the transmission time for a signal, using the same energy, and thus lowering the amplitude. The decoding of this long signal does not use the complete signal, but rather split it into pieces and sum up the result. If instead the whole signal was used the result should be roughly the same in all cases. This can be employed by using a soft decoding algorithm instead of hard decision, bit by bit.

7.5. —

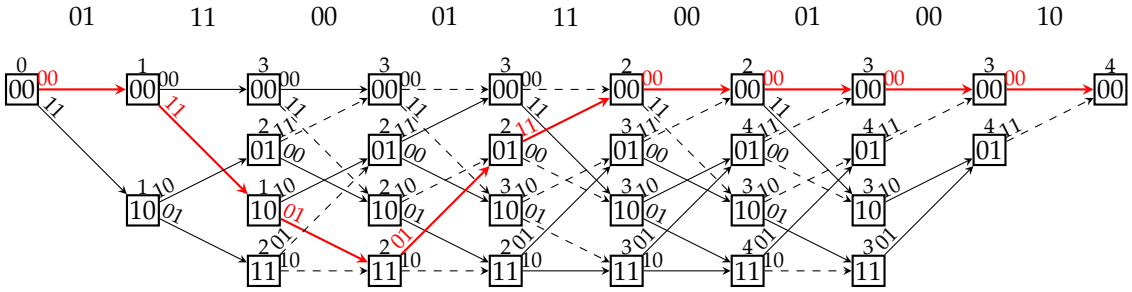
7.6. The free distance is the minimum weight of a non-zero path, starting and ending in the all-zero state. A simple and brute force method to find it is to start in the zero state and give a non-zero input. After this, follow all possible paths, counting the Hamming weight of the paths, until the zero state has the minimum commutative metric. This can be done in a tree, by expanding the minimum weight node until the zero state is minimum. It can also be done in a trellis by expanding the paths on step at a time until the zero state has a minimum weight. Below the trellis version of the algoritm is shown.



The algorithm is stopped when there are no other state with less weight than the zero state. Then it is seen that the free distance is $d_{\text{free}} = 5$. Notice that there are no branches diverging from the zero path once it has remerged. Such branches cannot become less than the metric in the zero state it emerges from.

In this case, the algorithm could have stopped already after the third step by noticing that the last step in the path going back to the zero state will add weight 2. Hence, the path up to any other state must be 2 less than the zero state at the same time instant, which is 5. There are publications of more efficient algorithms talking these things into account.

The decoding is done in a trellis comparing the received sequence with all the possible sequences of that length. The metric used in the following picture is the Hamming distance.



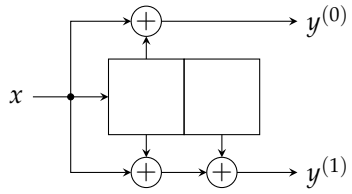
In the figure the red path is found by following the surviving branches from the end node to the start node. This corresponds to the minimum distance path, or the maximum likelihood path. In this case it is

$$\hat{\boldsymbol{v}}_1 = 00\ 11\ 01\ 01\ 11\ 00\ 00\ 00\ 00 \qquad \Rightarrow \hat{\boldsymbol{u}}_1 = 0110000$$

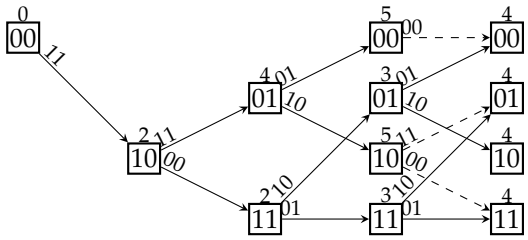
So the answer is that the most likely transmitted information sequence is $\hat{\boldsymbol{u}}_1 = 0110000$.

It is worth noticing that there are 2^7 possible information sequences, so the decoding in the trellis has compared 128 code sequences with the received sequence and sorted out the one with least Hamming distance.

7.7. The encoder circuit for this generator matrix is

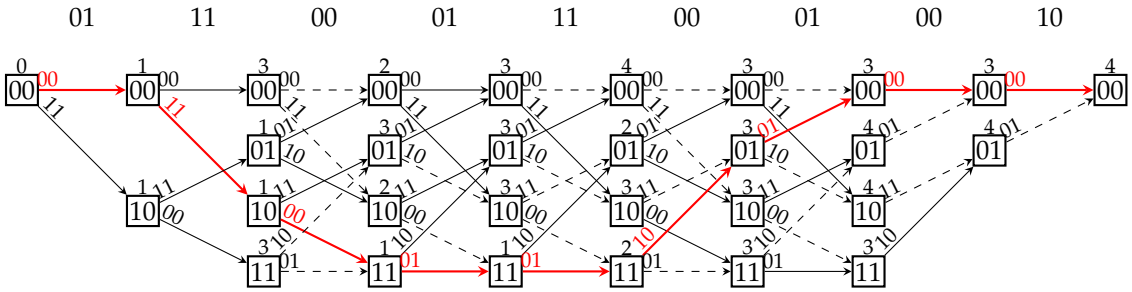


Following the same structure and methods as in Problem 7.6, the free distance is derived from the following trellis.



$\Rightarrow d_{\text{free}} = 4$

Decoding is done as follows.



Hence, the most likely code sequence is $\hat{\boldsymbol{v}} = 00\ 11\ 00\ 01\ 01\ 10\ 01\ 00\ 00$ and the corresponding information sequence is $\hat{\boldsymbol{u}} = 0111100$.

7.8. —

7.9. In Problem 7.6 the generator matrix

$$G(D) = \begin{pmatrix} 1 + D + D^2 & 1 + D^2 \end{pmatrix}$$

was specified. To show that they generate the same code, we should show that a codeword generated by one matrix also can be generated by the other. Their relation is $G_s(D) = \frac{1}{1+D^2} G(D)$.

First assume the code sequence $v_1(D)$ is generated by $G_s(D)$ from the information sequence $u_1(D)$ as $v_1(D) = u_1(D)G_s(D) = u_1(D)\frac{1}{1+D^2}G(D)$. Thus, $v_1(D)$ is also generated by $G(D)$ from the sequence $\tilde{u}_1(D) = \frac{u_1(D)}{1+D^2}$. Similarly, if a code sequence $v_2(D)$ is generated gy $G(D)$ from $u_2(D)$, then it is also generated by $G_s(D)$ from $\tilde{u}_2(D) = (1 + D^2)u_2(D)$. That is, any codeword generated by $G_s(D)$ can also be generated by $G(D)$, and vice versa, and the sets of codewords, i.e. the codes, are equivalent.

7.10. (a)

$$\frac{x^7 + x^6 + x^4 + x^2 + x + 1}{x^4 + x^3 + 1} = x^3 + 1 + \frac{x^2 + x}{x^4 + x^3 + 1}$$

remainder $\neq 0$, so no acceptance.

(b)

$$\frac{x^{10} + x^8 + x^6 + x^5 + x^3 + x^2 + 1}{x^4 + x^3 + 1} = x^6 + x^5 + \frac{x^3 + x^2 + 1}{x^4 + x^3 + 1}$$

remainder $\neq 0$, so no acceptance.

(c)

$$\frac{x^{10}+x^6+x^5+x^4+x^2+x+1}{x^4+x^3+1} = x^6+x^5+x^4+x^3+x^2+x+1$$

remainder $= 0$, so acceptance.

7.11.

Chapter 8

8.1. According to the definition of differential entropy ($H(X) = -\int f(x)\log f(x)\,dx$) we get that:

(a)

$$\begin{aligned} H(X) &= -\int_a^b f(x)\log f(x)\,dx = -\int_a^b \frac{1}{b-a}\log\left(\frac{1}{b-a}\right)dx \\ &= \left[\frac{x}{b-a}\log(b-a)\right]_a^b = \log(b-a) \end{aligned}$$

(b)

$$\begin{aligned} H(X) &= -\int_{-\infty}^{\infty} f(x)\log f(x)\,dx \\ &= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right]dx \\ &= \log\sqrt{2\pi\sigma^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx \\ &\quad + \frac{\log e}{2\sigma^2}\int_{-\infty}^{\infty}(x-\mu)^2\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx \\ &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{\log e}{2\sigma^2}\sigma^2 = \frac{1}{2}\log(2\pi e\sigma^2) \end{aligned}$$

(c)

$$\begin{aligned} H(X) &= -\int_0^{\infty} f(x)\log f(x)\,dx = -\int_0^{\infty}\lambda e^{-\lambda x}\log\left(\lambda e^{-\lambda x}\right)dx \\ &= -\int_0^{\infty}\lambda e^{-\lambda x}\left(\log\lambda - \lambda x\log e\right)\,dx = -\log\lambda + \lambda\log e\int_0^{\infty}x\lambda e^{-\lambda x}\,dx \\ &= -\log\lambda + \log e = \log\frac{e}{\lambda} \end{aligned}$$

(d)

$$\begin{aligned} H(X) &= -\int_{-\infty}^{\infty}\frac{1}{2}\lambda e^{-\lambda|x|}\log\left(\frac{1}{2}\lambda e^{-\lambda|x|}\right)dx \\ &= -\left[\frac{1}{2}\int_{-\infty}^{0}\lambda e^{\lambda x}\log\left(\frac{\lambda}{2}e^{\lambda x}\right)dx + \frac{1}{2}\int_0^{\infty}\lambda e^{-\lambda x}\log\left(\frac{\lambda}{2}e^{-\lambda x}\right)dx\right] \\ &= -\left[\int_0^{\infty}\left(\lambda e^{-\lambda x}\log\left(\frac{\lambda}{2}\right) + \lambda e^{-\lambda x}\log e(-\lambda x)\right)\,dx\right] \\ &= -\left[\log\left(\frac{\lambda}{2}\right) - \lambda\log e\int_0^{\infty}x\lambda e^{-\lambda x}\,dx\right] = \log\frac{2e}{\lambda} \end{aligned}$$

8.2. First derive α from $\int f(x,y)dxdy = 1$,

$$\iint_0^{\infty} f(x,y)dxdy = \alpha^2\int_0^{\infty}e^{-x}dx\int_0^{\infty}e^{-y}dy = \alpha^2 \Rightarrow \alpha = 1$$

(a) The probability that both X and Y are limited by 4 is

$$P(X<4,Y<4)=\iint_0^4 e^{-(x+y)}dxdy=\Big(\int_0^4 e^{-x}dx\Big)^2=\Big(\Big[-e^{-x}\Big]_0^4\Big)^2$$
$$=\big(1-e^{-4}\big)^2=1-2e^{-4}+e^{-8}\approx 0.9637$$

(b) Since $f(x,y)=e^{-(x+y)}=e^{-x}e^{-y}=f(x)f(y)$, the variables X and Y are independent and identically distributed., and they both have the same entropy

$$H(X)=-\int_0^\infty e^{-x}\log e-xdx=\log e\int_0^\infty xe^{-x}dx=\log e[-(1+x)e^{-x}]_0^\infty=\log e$$

The joint entropy is

$$H(X,Y)=H(X)+H(Y)=2H(X)=2\log e=\log e^2$$

(c) Since X and Y are independent $H(X|Y)=H(X)=\log e$.

8.3. First get α from

$$\iint_0^\infty f(x,y)dxdy=\Big(\alpha\int_0^\infty 2^{-x}dx\Big)^2=\Big(\alpha\Big[-\frac{2^{-x}}{\ln 2}\Big]_0^\infty\Big)^2=\Big(\frac{\alpha}{\ln 2}\Big)^2=1\Rightarrow\alpha=\ln 2$$

(a) The probability is

$$P(X<4,Y<4)=\iint_0^4\ln^2 2\,2^{-(x+y)}dxdy=\Big(\ln 2\int_0^4 2^{-x}dx\Big)^2$$
$$=\Big(\ln 2\Big[\frac{-2^{-x}}{\ln 2}\Big]_0^4\Big)^2=(1-2^{-4})^2=\frac{225}{256}\approx 0.88$$

(b) Since X ad Y are i.i.d. the joint entropy is

$$H(X,Y)=2H(X)=\log\Big(\frac{e}{\ln 2}\Big)^2\approx 3.94$$

where

$$H(X)=-\int_0^\infty\alpha 2^{-x}\log\alpha 2^{-x}dx=-\int_0^\infty\alpha 2^{-x}\Big(\log\alpha-x\Big)dx$$
$$=\alpha\int_0^\infty x2^{-x}dx-\log\alpha=\ln 2\Big[-\frac{(1+x\ln 2)2^{-x}}{\ln^2 2}\Big]_0^\infty$$
$$=\frac{1}{\ln 2}-\log(\ln 2)=\log\frac{e}{\ln 2}\approx 1.97$$

(c) Since X and Y are independent $H(X|Y)=H(X)=\log\frac{e}{\ln 2}$.

8.4. (a) Assign $Y=\ln X$, which is $N(\mu,\sigma)$ distributed, then $X=e^Y$. Then,

$$P(X<a)=P\big(e^Y<a\big)=P\big(Y<\ln a\big)$$
$$=\int_{-\infty}^{\ln a}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-\mu)^2}{2\sigma^2}}dy=\left[\begin{array}{l}x=e^y\ \Rightarrow\ y=\ln x\\ dy=\frac{1}{x}dx\end{array}\right]$$
$$=\int_0^a\frac{1}{x\sqrt{2\pi\sigma^2}}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}dx$$

which means $f_X(x)=\frac{1}{x\sqrt{2\pi\sigma^2}}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$.

(b) The mean, second order moment and variance can be found as

$$\begin{aligned}
E[X] &= \int_0^\infty \frac{x}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} dx = \left[\begin{array}{l} y = \ln x \;\Rightarrow\; x = e^y \\ dy = \frac{1}{x}dx \end{array} \right] \\
&= \int_{-\infty}^\infty \frac{e^y}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy \\
&= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-(\mu+\sigma^2))^2}{2\sigma^2}} e^{\mu+\frac{\sigma^2}{2}} dy = e^{\mu+\frac{\sigma^2}{2}} \\
E[X^2] &= \int_0^\infty \frac{x^2}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} dx = \left[\begin{array}{l} y = \ln x \;\Rightarrow\; x = e^y \\ dy = \frac{1}{x}dx \end{array} \right] \\
&= \int_{-\infty}^\infty \frac{e^{2y}}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy \\
&= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-(\mu+2\sigma^2))^2}{2\sigma^2}} e^{2\mu+2\sigma^2} dy = e^{2\mu+2\sigma^2} \\
V[X] &= E[X^2] - E[X]^2 = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2} = e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right)
\end{aligned}$$

(c) The entropy is derived by using the same change of variables, $y = \ln x$,

$$\begin{aligned}
H(X) &= -\int_0^\infty \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} \log\Big(\frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}\Big) dx \\
&= -\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \log\Big(\frac{e^{-y}}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}\Big) dx \\
&= \log e \int_{-\infty}^\infty \frac{y}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} - \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \log\Big(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}\Big) dx \\
&= \frac{E[Y]}{\ln 2} + H(Y) = \frac{\mu}{\ln 2} + \frac{1}{2}\log 2\pi e\sigma^2
\end{aligned}$$

8.5. Consider

$$\begin{aligned}
D\big(f(x)\big|\big|h(x)\big) &= \int f(x)\log\frac{f(x)}{h(x)}dx \\
&= -\int f(x)\log h(x)dx + \int f(x)\log f(x)dx \\
&= \log 2\pi e\sigma^2 - H_f(X)
\end{aligned}$$

where we used that $-\int f(x)\log h(x)dx = \log 2\pi e\sigma^2$.

8.6. (a) The sum of two normal variables is normal distributed with $N(\mu_1+\mu_2, \sqrt{\sigma_1^2+\sigma_2^2})$.

(b) According to Problem 8.1 the entropy becomes $\frac{1}{2}\log 2\pi e(\sigma_1^2+\sigma_2^2)$.

8.7. The differential entropy for a uniformly distributed variable between a and b is $H(X) = \log(b-a)$.

(a) $H(X) = \log(2-1) = \log 1 = 0$

(b) $H(X) = \log(200-100) = \log 100 \approx 6,644$

8.8. From the problem we have that $P(X=0) = p$, $P(X=1) = 1-p$ and that $f(z) = \frac{1}{a}$, $0 \le z \le a$, where $a > 1$. This gives that the conditional density for y becomes

$$\begin{aligned}
f(y|X=0) &= \frac{1}{a}, \quad 0 \le y \le a \\
f(y|X=1) &= \frac{1}{a}, \quad 1 \le y \le a+1
\end{aligned}$$

which gives the density for y as

$$f(y) = \sum_x f(y|X=x)P(X=x) = \begin{cases} p\frac{1}{a}, & 0 \le y \le 1 \\ \frac{1}{a}, & 1 \le y \le a \\ (1-p)\frac{1}{a}, & a \le y \le a+1 \end{cases}$$

(a)
$$\begin{aligned} H(X) &= h(p) \\ H(X|Y) &= \underbrace{H(X|0 \le y \le 1)}_{=0} P(0 \le y \le 1) \\ &\quad + \underbrace{H(X|1 \le y \le a)}_{=h(p)} P(1 \le y \le a) \\ &\quad + \underbrace{H(X|a \le y \le a+1)}_{=0} P(a \le y \le a+1) \\ &= -h(p) \int_1^a \frac{1}{a} \log \frac{1}{a} dy = h(p) \frac{a-1}{a} \\ I(X;Y) &= H(X) - H(X|Y) = \frac{1}{a} h(p) \end{aligned}$$

(b)
$$\begin{aligned} H(Y) &= -\int_0^1 \frac{p}{a} \log \frac{p}{a} dy - \int_1^a \frac{1}{a} \log \frac{1}{a} dy - \int_a^{a+1} \frac{1-p}{a} \log \frac{1-p}{a} dy \\ &= -\frac{p}{a} \log \frac{p}{a} - (a-1)\frac{1}{a} \log \frac{1}{a} - \frac{1-p}{a} \log \frac{1-p}{a} \\ &= \frac{1}{a} h(p) + \log a \\ H(X|Y) &= \sum_x \underbrace{H(Y|X=x)}_{=\log a} P(X=x) = \log a \\ I(X;Y) &= H(Y) - H(Y|X) = \frac{1}{a} h(p) \end{aligned}$$

(c) $C = \max_p I(X;Y) = \frac{1}{a}$ for $p = \frac{1}{2}$.

Chapter 9

9.1. The capacity of this additive white Gaussian noise channel with the output power constraint $E[Y^2] \le P$ is

$$\begin{aligned} C &= \max_{f(X):E[Y^2] \le P} I(X;Y) = \max_{f(X):E[Y^2] \le P} (H(Y) - H(Y|X)) \\ &= \max_{f(X):E[Y^2] \le P} (H(Y) - H(Z)) \end{aligned}$$

Here the maximum differential entropy is achieved by a normal distribution and the power constraint on Y is satisfied if we choose the distribution of X as $N(0, P-\sigma)$. The capacity is

$$C = \frac{1}{2} \log (2\pi e(P - \sigma + \sigma)) - \frac{1}{2} \log (2\pi e(\sigma)) = \frac{1}{2} \log (2\pi eP) - \frac{1}{2} \log (2\pi e\sigma) = \frac{1}{2} \log (\frac{P}{\sigma})$$

9.2. —

9.3. (a) The received power is

$$P_Z = |H_2|^2 P_Y = |H_1|^2 |H_2|^2 P_X$$

and the received noise is Gaussian with variance

$$N = N_1 |H_1|^2 + N_2$$

Hence, an equivalent channel model from X to Z has the attenuation $H_1 H_2$ and additive noise with distribution $n \sim N(0, \sqrt{N_1|H_1|^2 + N_2})$. That means the capacity becomes

$$C = W \log\Big(1 + \frac{|H_1|^2 |H_2|^2 P_X}{W(N_1|H_1|^2 + N_2)}\Big)$$

(b) From the problem we get the SNRs for the two channels

$$\text{SNR}_1 = \frac{|H_1|^2 P_X}{WN_1} = \frac{P_Y}{WN_1}$$
$$\text{SNR}_1 = \frac{|H_2|^2 P_Y}{WN_2} = \frac{P_Z}{WN_2}$$

Then the total SNR can be expressed as

$$\text{SNR} = \frac{|H_1|^2 |H_2|^2 P_X}{W(N_1|H_1|^2 + N_2)} = \frac{\frac{|H_1|^2 P_X |H_2|^2 P_Y}{WN_1 WN_2}}{\frac{P_Y}{W^2 N_1 N_2} W(N_2 + N_1|H_1|^2)}$$
$$= \frac{\frac{|H_1|^2 P_X}{WN_1} \frac{|H_2|^2 P_Y}{WN_2}}{\frac{P_Y}{WN_1} + \frac{|H_1|^2 P_Y}{WN_2}} = \frac{\text{SNR}_1 \cdot \text{SNR}_2}{\text{SNR}_1 + \text{SNR}_2}$$

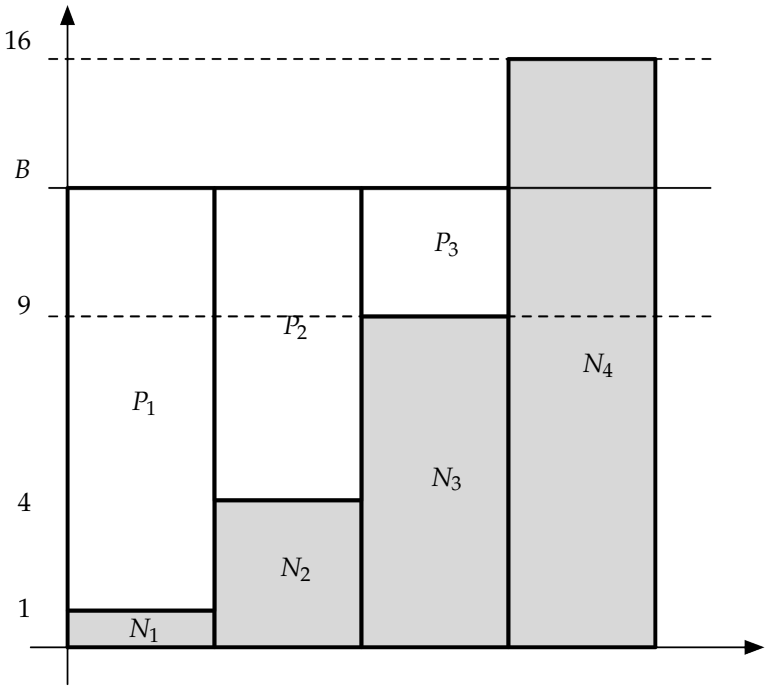
Notice, that by considering the invers of the SNR, the noise to signal ratio, the derivations can be considerably simplified,

$$\frac{1}{\text{SNR}} = \frac{W(N_1|H_1|^2 + N_2)}{|H_1|^2 |H_2|^2 P_X} = \frac{WN_1}{|H_2|^2 P_X} + \frac{WN_2}{|H_1|^2 |H_2|^2 P_X} = \frac{1}{\text{SNR}_1} + \frac{1}{\text{SNR}_2}$$

which is equivalent to the desired result.

9.4. We can use the total power $P_1 + P_2 + P_3 + P_4 = 17$ and for the four channels the noise power is $N_1 = 1$, $N_2 = 4$, $N_3 = 9$, $N_4 = 16$. Let $B = P_i + N_i$ for the used channels. Since $(16-1) + (16-4) + (16-9) > 17$ we should not use channel four when reaching capacity. Similarly, since $(9-1) + (9-4) < 17$ we should use the rest of the three channels. These tests are marked as dashed lines in the figure below. Hence, $B = P_1 + 1 = P_2 + 4 = P_3 + 9$, which leads to $B = \frac{1}{3}(P_1 + P_2 + P_3 + 14) = \frac{1}{3}(17 + 14) = \frac{31}{3}$. The capacity becomes

$$C = \sum_{i=1}^{3} \frac{1}{2} \log\Big(1 + \frac{P_i}{N_i}\Big) = \sum_{i=1}^{3} \frac{1}{2} \log \frac{B}{N_i} = \frac{1}{2} \log \frac{\frac{31}{3}}{1} + \frac{1}{2} \log \frac{\frac{31}{3}}{4} + \frac{1}{2} \log \frac{\frac{31}{3}}{9}$$
$$= \frac{3}{2} \log 31 - \frac{5}{2} \log 3 - 1 \approx 2.4689$$



9.5. (a) Use the water filling algorithm to derive the capacity. When a sub-channel is deleted ($P_i = 0$) the total number of sub-channel is changed and the power distribution has to be recalculated. We get the following recursion:

1. Iteration 1
   $B = B - N_i = \frac{1}{6}(\sum_i N_i + P) = 14.17$
   $P_i = (6.17, 2.17, 0.17, 4.17, -1.83, 8.17)$
   Sub-channel 5 should not be used, $P_5 = 0$.
2. Iteration 1
   $B = \frac{1}{5}(\sum_{i \neq 5} N_i + P) = 13.8$
   $P_i = B - N_i = (5.80, 1.80, -0.20, 3.80, 0, 7.80)$
   Sub-channel 3 should not be used, $P_3 = 0$.
3. Iteration 1
   $B = \frac{1}{4}(\sum_{i \neq 3,5} N_i + P) = 13.75$
   $P_i = B - N_i = (5.75, 1.75, 0, 3.75, 0, 7.75)$
   All remaining sub-channels can be used.

The capacities in the sub-channels are

$$C_i = \frac{1}{2}\log\Big(1 + \frac{P_i}{N_i}\Big) = (0.39, 0.10, 0, 0.23, 0, 0.60)$$

and the total capacity $C = \sum_i C_i = 1.32$ bit/transmission.

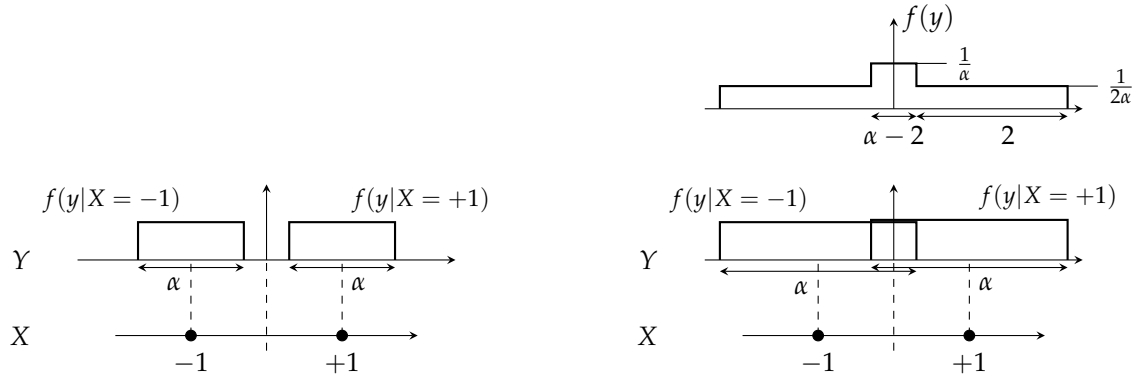
(b) If the power is equally distributed over the sub-channels we get $P_i = 19/6 = 3.17$. That gives the capacities

$$\begin{aligned} C &= \sum_i \frac{1}{2}\log\Big(1 + \frac{19/6}{N_i}\Big) \\ &= 0.24 + 0.17 + 0.15 + 0.20 + 0.13 + 0.31 = 1.19 \text{ bit/transmission} \end{aligned}$$

(c) When using only one sub-channel the capacity is maximised if we take the one with least noise, $N = 6$. This gives the capacity $C = \frac{1}{2}\log 2(1 + 19/6) = 1.03$ bit/transmission.

Chapter 10

10.1. In the following figure the resulting distributions are depicted.



(a) For $\alpha < 2$ the left figure describes the received distribution. Since the density functions $f(y|X=1)$ and $f(y|X=-1)$ are non-overlapping, the transmitted value can directly be determined from the received Y. Hence, $I(X;Y) = 1$.

The result can also be found from the following derivations:

$$
\begin{aligned}
H(Y|X=i) &= -\int_{-\alpha/2}^{\alpha/2} \frac{1}{\alpha} \log \frac{1}{\alpha} dx = \log \alpha \\
H(Y|X) &= \sum \frac{1}{2} H(Y|X=i) = \log \alpha \\
H(Y) &= -2 \int_{-\alpha/2}^{\alpha/2} \frac{1}{2\alpha} \log \frac{1}{2\alpha} dx = \log 2\alpha = 1 + \log \alpha \\
I(X;Y) &= H(Y) - H(Y|X) = 1 \text{ b/tr}
\end{aligned}
$$

(b) For $\alpha \geq 2$ there is an overlap between $f(y|X=1)$ and $f(y|X=-1)$ as shown in the right figure. Still, $H(Y|X) = \log \alpha$, but since the distribution of Y depends on the amount of the overlap, we need to rederive the entropy,

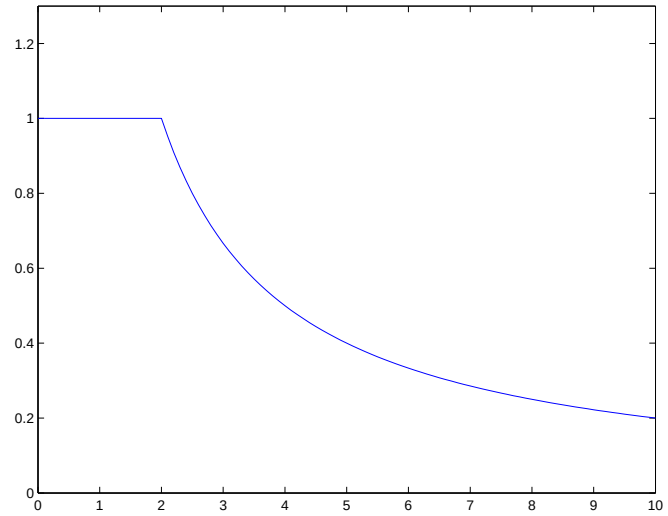
$$
\begin{aligned}
H(Y) &= -\int_{-1-\frac{\alpha}{2}}^{1-\frac{\alpha}{2}} \frac{1}{2\alpha} \log \frac{1}{2\alpha} dx - \int_{1-\frac{\alpha}{2}}^{-1+\frac{\alpha}{2}} \frac{1}{\alpha} \log \frac{1}{\alpha} dx - \int_{-1+\frac{\alpha}{2}}^{1+\frac{\alpha}{2}} \frac{1}{2\alpha} \log \frac{1}{2\alpha} dx \\
&= 2\frac{1}{2\alpha} \log(2\alpha) 2 + \frac{1}{\alpha} \log(\alpha)(\alpha - 2) \\
&= \frac{2}{\alpha} + \log \alpha
\end{aligned}
$$

Thus, $I(X;Y) = \frac{2}{\alpha}$ b/tr.

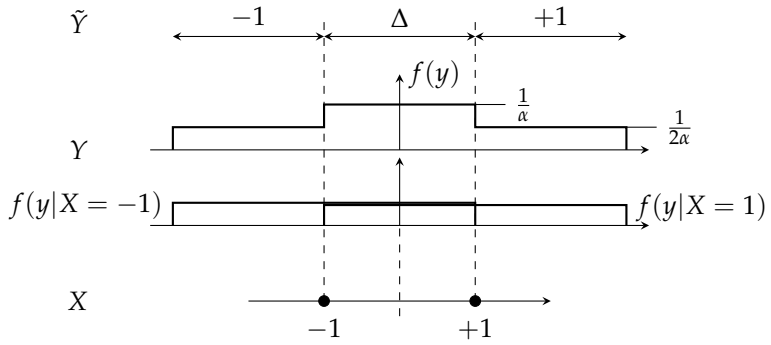
Summarising, the mutual information becomes

$$
I(X;Y) = \min\left\{1, \frac{1}{\alpha}\right\}, \quad \alpha > 0
$$

which is plotted below.



10.2. For $\alpha = 4$ the mutual information is $I(X;Y) = 2/4 = 1/2$. We get the following density functions, where also the intervals for hard decoding is shown.



The probability for overlap is $P(\Delta|X = i) = 1/2$, and the resulting DMC channel is the binary erasure channel. Hence, the capacity is

$$C_{\text{BEC}} = 1 - \frac{1}{2} = \frac{1}{2}$$

In most cases it is beneficially to use the the soft information, the value of the received symbol instead of the hard decoding, since it should grant some extra information. E.g. in the case of binary transmission and Gaussian noise it is a difference if the received symbol is 3 or 0.5. But in the case here we have uniform noise. Then we get three intervals where for $\tilde{Y} = -1$ it is certain that $X = -1$ and for $\tilde{Y} = 1$ it is certain that $X = 1$. When $\tilde{Y} = \Delta$ the two possible transmitted alternatives are equally likely and we get no information at all. Since the information is either complete or none, there is no difference between the two models.

As a comparison, for $\alpha > 2$ the probability for the overlapped interval is $P(\Delta|X = i) = \frac{\alpha-2}{\alpha} = 1 - \frac{2}{\alpha}$. Thus, the capacity for the BEC is $C_{\text{BEC}} = \frac{2}{\alpha}$, which is the same as for the continuous case.

10.3. The mutual information is $I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(X+Z|X) = H(Y) - H(Z)$, where $H(Z) = \log(1-(-1)) = \log 2$. Since Y ranges from -3 to 3 with uniform weights $p_{-2}/2$ for $-3 \leq Y \leq -2$, $(p_{-2} + p_{-1})/2$ for $-2 \leq Y \leq -1$ etc the maximum of $H(Y)$ is obtained for a uniform Y. This can be achieved if the distribution of X is $(\frac{1}{3}, 0, \frac{1}{3}, 0, \frac{1}{3})$. Now $H(Y) = \log(3-(-3)) = \log 6$.

We conclude that $C = \log 6 - \log 2 = \log 3$.

10.4.

10.5.

10.6. —

Chapter 11

11.1.

11.2.

11.3. Follows directly from Problem 11.2.

11.4.

11.5.

$$E[d(x, x_Q)] = \frac{\sigma^2}{\pi}(\pi - 2)$$

Chapter 12