LUND UNIVERSITY
Electrical and Information Technology

Information Theory

Solutions to problems

Problems

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 $Z = X + Y \mod 2$. Show that Z is independent of Y for all values of p.

(b) Let X be a stochastic variable uniformly distributed over $\{1, 2, \dots, M\}$. Let Y be independent of X, with an arbitrary probability function over $\{1, 2, \dots, M\}$. 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. The random variable X denotes the outcome of a roll with a five sided fair die and Y the outcome from a roll with an eight sided fair die.

(a) What is the distribution of $Z_a = X + Y$

(b) What is the distribution of $Z_b = X - Y$

(c) What is the distribution of $Z_c = |X - Y|$

2.5. Flip a fair coin until Heads comes up and denote the number of flips by X.

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

2.6. Let X be Poisson distributed, $X \sim \text{Po}(\lambda)$, see Appendix **??**. Show that the expectation and variance are $E\big[X\big] = V\big[X\big] = \lambda$.

2.7. Let X be Exponentially distributed, $X \sim \text{Exp}(\lambda)$, see Appendix **??**. Show that the expectation and variance are $E\big[X\big] = \frac{1}{\lambda}$ and $V\big[X\big] = \frac{1}{\lambda^2}$.

2.8. Show that the second order moment around a point c is minimised by the variance, i.e. that

$$E\big[(X - c)^2\big] \geq E\big[(X - m)^2\big]$$

with equality if and only if $c = m$, where $m = E[X]$.

2.9. 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 number of zeros.

2.10. An urn has 10 balls, seven white and three black. Six times after each other a ball is drawn from the urn. What is the probability of the number of black balls drawn in the series, if

(a) the ball is replaced in the urn after each draw?

(b) drawn balls are not replaced?

2.11. Use Jensen's inequality to show

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

Hint: The logarithm is a concave function.

2.12. On some occasions in the text book, Stirling's approximation is used to relate the binomial function with the binary entropy, defined in Chapter **??**. There are different versions of this approximation in the literature, with different accuracy (and difficulty). Here, one of the basic versions is derived.

(a) Consider the logarithm of the faculty function

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

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

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

or, equivalently,

$$n! \approx e\sqrt{n}\Big(\frac{n}{e}\Big)^n$$

(b) To improve the approximation for large n, use the limit value

$$\lim_{n \to \infty} n! \Big(\frac{e}{n}\Big)^n \frac{1}{\sqrt{n}} = \sqrt{2\pi}$$

Show how this gives the more common version of Stirling's approximation

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

(c) Use the result in (b) to estimate the approximation error in (a) for large n.

2.13. A Markov process is defined by the state transition graph in Figure 1.

(a) Give the state transition matrix, P.

(b) Derive the steady state distribution, $\boldsymbol{\pi} = (\pi_0\ \pi_1\ \pi_2\ \pi_3)$.

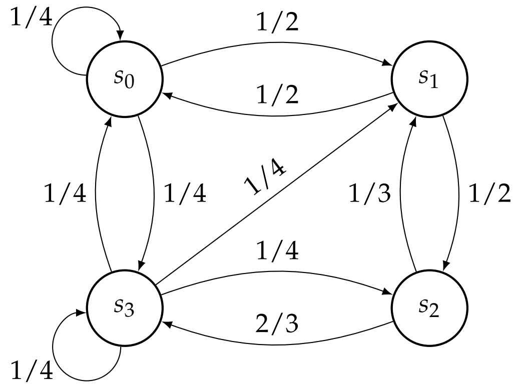


Figure 1: A state transition graph for a Markov process.

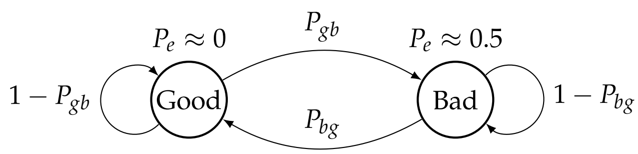


Figure 2: Gilbert-Elliott channel model for bursty noise.

2.14. Often in communication systems the transmission is distorted by bursty noise. One way to model the noise bursts is through the so called Gilbert-Elliott channel model. It consists of a time discrete Markov model with two states, *Good* and *Bad*. In the *Good* state the transmission is essentially error free, while in the *Bad* state the error probability is high, e.g. 0.5. The probability for transition from *Good* to *Bad* is denoted by P_{gb}, and from *Bad* to *Good* is P_{gb}, see Figure 2.

(a) Derive the steady state distribution for the Markov model.

(b) What is the expected time duration for a burst?

(c) What is the expected time between two consecutive bursts?

2.15. Consider an infinite random walk process with states s_k, $k \geq 0$. If the process is in state s_k it will take one step backwards to s_{k-1} with probability q or one step forward to s_{k+1} with probability p. For state s_0 the step backwards leads back to itself. The graph of the Markov process is presented in Figure 3. Assume that $p < q$ and $q = 1 - p$, and derive the steady state distribution.

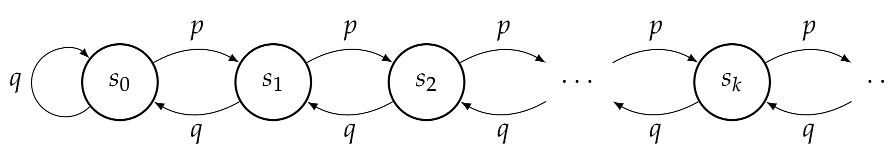


Figure 3: Markov chain for an infinite state random walk.

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 natural base, i.e. when $b = e$.

3.2. Use the IT-inequality to show that, for all positive x, $\ln x \geq 1 - \frac{1}{x}$ with equality if and only if $x = 1$.

3.3. The outcome of a throw with a fair die is denoted by X. Then, 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.

(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

$P(X,Y)$

X	$Y = a$	$Y = b$	$Y = c$
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) $I(X, Y)$

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

$P(X,Y)$

X	$Y = a$	$Y = b$	$Y = c$
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. In an experiment there are 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 you get about the identity of the coin by studying the number of Heads from the flips?

3.8. An urn has 18 balls; ten blue, five red and three green. Someone draws one ball from the urn and puts it in a box without looking. Let the random variable X denote the colour fo this first ball. Next, you draw a ball from the urn and let Y denote the colour of this second ball.

(a) What is the uncertainty of X?
(b) What is the uncertainty of Y if you first open the box to get the colour of the first ball?
(c) What is the uncertainty of Y if you do not open the box?
(d) Assume that you do not open the box. How much information about X do you get from Y?

3.9. Consider two dice where the first has equal probability for all six numbers and the second has a small weight close to the surface of number 1. Let X be the outcome of a roll 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 die and the manipulated die, respectively?
(b) What is $D(p||q)$?
(c) What is $D(q||p)$?

3.10. The joint distribution of X and Y is given by

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

(a) Determine k.
(b) Derive $P(X < 4, Y < 4)$
(c) Derive the joint entropy.
(d) Derive the conditional probability $H(X|Y)$.

3.11. The two distributions $p(x,y)$ and $q(x,y)$ are defined over the same set of outcomes. Verify that

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

and that, if X and Y are independent,

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

3.12. 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

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

3.13. (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

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

3.14. 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.

3.15. Show that

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

3.16. 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, 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 given two symmetric measures 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

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

3.17. 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.

3.18. A Markov source with output symbols $\{A, B, C\}$, is characterised by the graph in Figure 4.

(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?

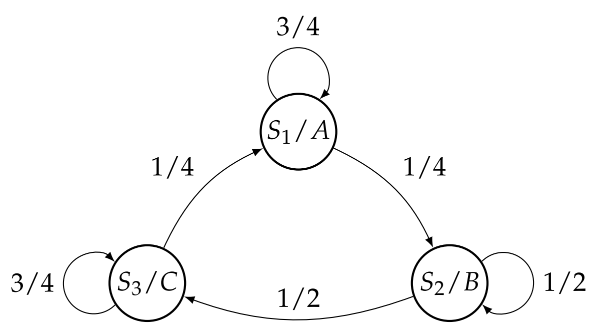


Figure 4: A Markov graph for the source in Problem 3.18

3.19. The engineer Inga is going to spend her vacation in an archipelago with four main islands. The islands are connected with four different boat lines, and one sightseeing tour around the largest island, see the map in Figure 5.

To avoid planning the vacation route too much, she decides to take a boat every day. She will choose one of the boat lines 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 each of 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).

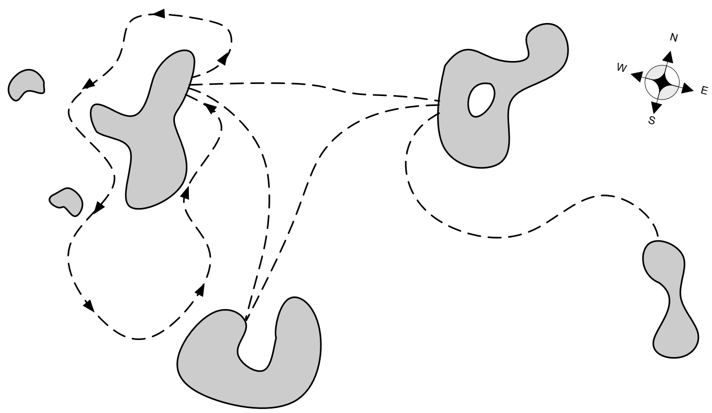


Figure 5: A map over the islands and the boat connections.

3.20. A man climbs an infinitely long ladder. At each time instant he tosses a coin. If he gets Head he takes a step up on the ladder but if he gets Tail he drops down to the ground (step 0). The coin is counterfeit with $P(\text{Head}) = p$ and $P(\text{Tail}) = 1 - p$. The sequence of where on the ladder the man stands forms a Markov chain.

(a) Construct the state transition matrix for the process and draw the state transition graph.

(b) What is the entropy rate of the process?

(c) After the man has taken many steps according to the process you call him and ask if he is on the ground. What is the uncertainty about his answer? If he answers that he is not on the ground, what is the uncertainty of which step he is on? (You can trust that he is telling the truth.)

3.21. Four points are written on the unit circle, see Figure 6. A process moves from the current point to one of its neighbours. If the current point is $\Phi = \varphi_i$, the next point is chosen with probabilities

$$\Phi^+ = \begin{cases} \varphi_i + \frac{\pi}{2}, & p_+ = \frac{2\pi - \varphi_i}{2\pi} \\ \varphi_i - \frac{\pi}{2}, & p_- = \frac{\varphi_i}{2\pi} \end{cases} \tag{2}$$

Derive the entropy rate for the process.

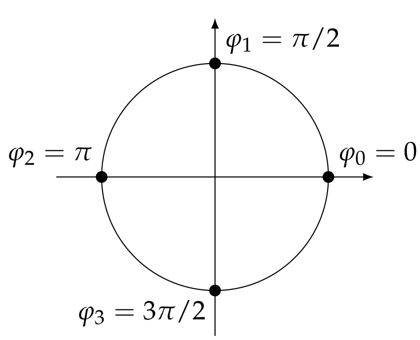


Figure 6: Four points on the unit circle

3.22. Consider a discrete stationary (time invariant) random process, $X_1 X_2 \dots X_n$, where $X_i \in \{x_1, x_2, \dots, x_k\}$ and k finite. Define the two entropy functions

$$\begin{aligned} H_n(X) &= \tfrac{1}{n} H(X_1 \dots X_n) \\ H(X|X^n) &= H(X_n|X_1 \dots X_{n-1}) \end{aligned}$$

As $n \to \infty$ these functions approaches the entropy rate functions $H_\infty(X)$ and $H(X|X^\infty)$. In the course it has been shown that these limits exists and are equal, $H_\infty(X) = H(X|X^\infty)$. In this problem the same result is derived in an alternative way.

(a) Show that

$$H_n(X) \geq H(X|X^n)$$

(b) Show that $H_n(X)$ is a decreasing function, i.e. that

$$H_n(X) \leq H_{n-1}(X)$$

Remark: This shows $0 \leq H_\infty(X) \leq H_n(X) \leq H_{n-1}(X) \leq H(X) \leq \log k$, and that the limit exists.

(c) Show that for any fixed integer μ, in $1 \leq \mu < n$,

$$H_n(X) \leq \tfrac{\mu}{n} H_\mu(X) + \tfrac{n-\mu}{n} H(X|X^\mu)$$

(d) Use the results above to show that for all $\mu \ll n$

$$H(X|X^\infty) \leq H_\infty(X) \leq H(X|X^\mu)$$

Remark: By letting $\mu \to \infty$ this gives $H_\infty(X) = H(X|X^\infty)$.

Chapter 4

4.1. Consider the code $\{0, 01, 10\}$.

(a) Is it non-singular?
(b) Is it uniquely decodable?
(c) Is it a prefix code?

4.2. For each of the following sets of codeword lengths decide if there exists a binary prefix 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) $\{2, 3, 3, 3, 4, 5, 5, 5\}$

4.3. In the following tables the probabilities for a random variable X is given together with a binary prefix 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 $L = E[\ell]$.
(c) Is it possible that the code is optimal for X?
(d) Is the code optimal for X?

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

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

4.6. Construct optimal codes for the random variables with the following distributions. Also derive the average codeword lengths and the entropies.

(a)

$x:$	x_1	x_2	x_3	x_4
$p(x):$	$1/4$	$1/4$	$3/8$	$1/8$

(b)

$x:$	x_1	x_2	x_3	x_4	x_5
$p(x):$	$1/3$	$1/6$	$1/6$	$2/9$	$1/9$

(c)

$x:$	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

4.7. 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 compared to the uncoded case.

x	y
000	0
001	100
010	110
011	11100

x	y
100	101
101	11101
110	11110
111	11111

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

x	Nbr
A	73
B	9
C	30
D	44
E	130

x	Nbr
F	28
G	16
H	35
I	74
J	2

x	Nbr
K	3
L	35
M	25
N	78
O	74

x	Nbr
P	27
Q	3
R	77
S	63
T	93

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

Construct a binary Huffman code for the 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?

4.9. 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. symbols, $\boldsymbol{X}^{(n)} = X_1 X_2 \dots X_n$. 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.

4.10. A *unary* source code is a mapping from the positive integers to binary vectors such that the codeword for the integer n consists of $n-1$ zeros followed by a one, i.e.

n	y
1	1
2	01
3	001
4	0001
5	00001
$\vdots$	$\vdots$

This is clearly a prefix code since there is only one 1, located at the end of each codeword.

On Wikipedia (`https://en.wikipedia.org/wiki/Unary_coding`)[1] it is stated that the unary code is optimal for the probability function

$$P(n) = (k-1)k^{-n}, \quad n = 1, 2, \dots \text{ and } k \geq 1.61803\dots$$

The aim of this problem is to clarify if this is correct. However, we will only consider probability functions where k is an integer, i.e. $k = 2, 3, 4, \dots$ is considered.

(a) Is it true that for any integer $k \geq 2$ the function $P(n) = (k-1)k^{-n}$, $n = 1, 2, 3, \dots$ is a probability function?

(b) Is it true that for $k = 2$ the unary code is optimal? What is the average codeword length?

(c) Is it true that for any integer $k = 2, 3, 4, \dots$ the unary code is optimal? What is the average codeword length?

4.11. A random variable X has k outcomes and probabilities $p_i = P(X = x_i)$. To construct a source code for the variable, order the outcomes according to the probabilities, with the highest to the left, i.e. the vector becomes $x_1, x_2, \dots, x_k$ where $p_1 \geq p_2 \geq \dots \geq p_k$. Split the vector in two parts such that their sums are as close as possible. That is, find an index q such that

$$\left| \sum_{i=1}^{q} p_i - \sum_{j=q+1}^{k} p_j \right|$$

is as small as possible. Label the left part with 0 and the right part with 1. Continue the procedure iteratively in the same way with each of the parts until there are only two outcomes in the vector. The labeling constitute the codeword for each outcome.

[1] You can of course not use any formulas found on Wikipedia, or elsewhere on Internet, without verifying it.

The described method is often referred to as Fano coding. For each of the probability vectors below find the Fano code and state if it is also a Huffman code and/or if it is optimal.

$$P_a = \left(\frac{4}{10}, \frac{3}{10}, \frac{2}{10}, \frac{1}{10}\right)$$

$$P_b = \left(\frac{6}{21}, \frac{5}{21}, \frac{4}{21}, \frac{3}{21}, \frac{2}{21}, \frac{1}{21}\right)$$

$$P_c = \left(\frac{15}{43}, \frac{7}{43}, \frac{7}{43}, \frac{7}{43}, \frac{7}{43}\right)$$

4.12. Use arithmetic coding to compress the string $\boldsymbol{x} = \text{d a c b d a a a d a}$, if the distribution is given by

x :	a	b	c	d
$p(x)$:	0.5	0.3	0.1	0.1

4.13. Use the same distribution as in Problem 4.12 and decode the following codeword from an arithmetic encoder.

$$y = 01111000111011100$$

4.14. Construct a binary random vector $\boldsymbol{x} = x_1 x_2 \dots x_n$ of i.i.d. symbols with the distribution

x :	0	1
$p(x)$:	1/4	3/4

Compress the vector using arithmetic coding for $n = 10\,000$. What is the average codeword length per symbol? Compare with the entropy.

Note that this problem is best suited for computer implementation, and that the intervals need to be scaled to avoid numerical problems.

Chapter 5

5.1. Given the same source sequence as in Problem 4.12, i.e. $\boldsymbol{x} = \text{d a c b d a a a d a}$.

(a) Use a two-pass Huffman coding of the sequence. That is, estimate the distribution from the sequence, construct a Huffman code and encode the sequence.

(b) Compare the result in (a) with the case when the true distribution is given, see Problem 4.12.

5.2. Use the one sweep Huffman algorithm to encode the sequence in Problem 5.1.

5.3. 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?

5.4. Compress the following text string using LZ77 with $S = 16$ and $B = 8$,

I scream, you scream, we all scream for ice cream.

Initialise the encoder with the beginning of the string. Give the codewords and the compression ratio.

5.5. 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	$(0, \llcorner\!\lrcorner)$
5	$(1, h)$
6	$(0, e)$
7	$(4, t)$
8	$(0, h)$
9	$(2, n)$
10	$(7, w)$
11	$(9, \llcorner\!\lrcorner)$
12	$(1, i)$
13	$(0, n)$
14	$(0, s)$
15	$(3, i)$
16	$(5, .)$

Decode to get the text back.

5.6. 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?

5.7. Compress the text

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

and calculate the compression rate using for the following settings.

(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.

5.8. 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?

5.9. 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?

5.10. Decode the following codeword sequence using the LZW algorithm,

$$\boldsymbol{y} = 102, 114, 101, 115, 104, 108, 121, 32, 256, 105, 101, 100, 263, 257, 259, 263, 108, 258, 104$$

The algorithm is initialised with the ASCI table from index 1 to 255, where the following is interesting for the problem:

Ind:	32	100	101	102	104	105	108	114	115	121
Letter:	space	d	e	f	h	i	l	r	s	y

Chapter 6

6.1. Consider a binary memory-less source where $P(0) = p$ and $P(1) = q = 1 - p$. For large n, the number of 1s in a sequence of length n tends to nq.

(a) How many sequences of length n has the number of ones equal to nq?

(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) equals the entropy, $h(q) = h(p)$.

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

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

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

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

(a) What is the probability for the most probable vector?

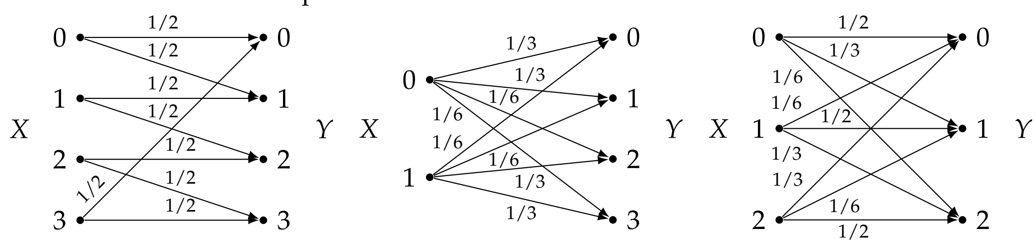
(b) Is the most probable vector ϵ-typical?

(c) How many ϵ-typical vectors are there?

6.4. 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 during a long sequence of splits?

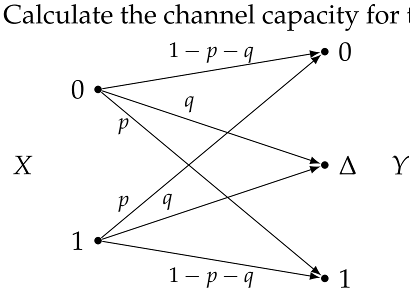
Hint: Consider the distribution of saved parts for the most common type of sequence.

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

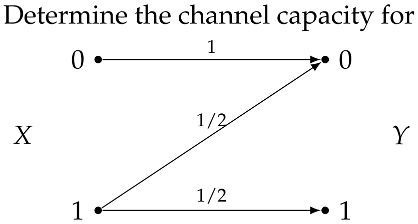


(a) Noisy typewriter (b) Soft decoding (c) 3-ary channel

6.6. Calculate the channel capacity for the extended binary erasure channel shown below.



6.7. Determine the channel capacity for the following channel.



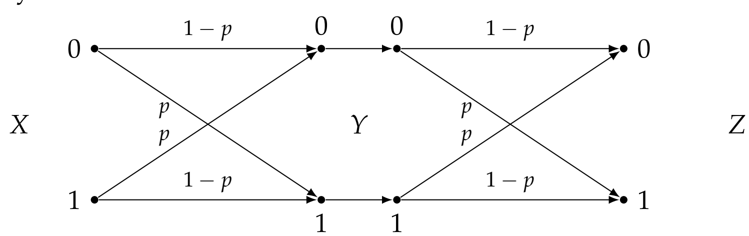
Hint: $D(h(p)) = D(p)\log(\frac{1-p}{p})$.

6.8. The random variable $X \in \{0, 1, \ldots, 14\}$ is transmitted over an additive channel,

$$Y = X + Z, \qquad \text{mod } 15$$

where $p(Z=1) = p(Z=2) = p(Z=3) = \frac{1}{3}$. What is the capacity for the channel and for what distribution $p(x)$ is it reached?

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



6.10. A discrete memoryless channel is shown in Figure 7.

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

(b) Assume that the probability for X is given by $P(X = 0) = 1/6$ and $P(X = 1) = 5/6$, and that the source is memoryless. Find an optimal code to compress the sequence Y. What is the average codeword length?

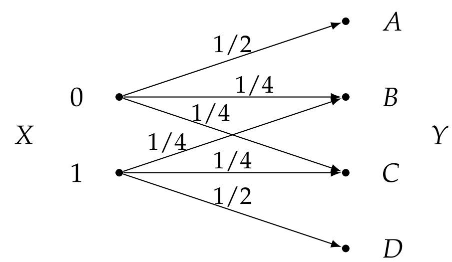


Figure 7: A discrete memoryless channel

6.11. Two channels are cascaded, one BSC and one BEC, according to Figure 8. Derive the channel capacity.

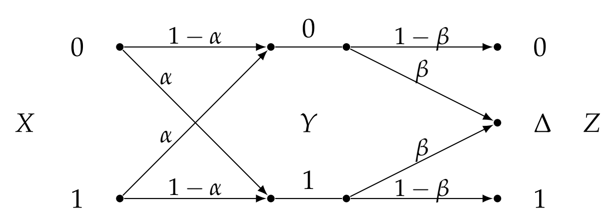


Figure 8: One BSC and one BEC in cascade.

6.12. Use two discrete memoryless channels in parallel, see Figure 9. That is, the transmitted symbol X is transmitted over two channels and the receiver gets two symbols, Y and Z. The two channels work independently in parallel, i.e. $P(y|x)$ and $P(z|x)$ are independent distributions, and hence, $P(y|x,z) = P(y|x)$ and $P(z|x,y) = P(z|x)$). However, it does not mean Y and Z are independent, so in general $p(y|z) \neq p(y)$.

(a) Consider the information about X the receiver gets by observing Y and Z, and show that

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

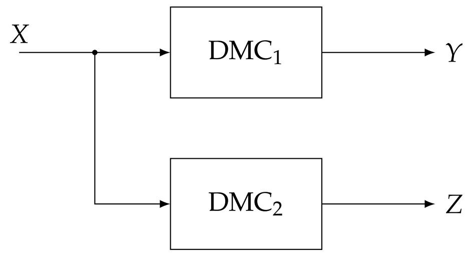


Figure 9: Two DMC used in parallel.

(b) Consider the case when both the channels are BSC with error probability p, and let $P(X = 0) = P(X = 1) = \frac{1}{2}$. Use the result in a and show that

$$
\begin{aligned}
I(X;Y,Z) &= H(Y,Z) - 2h(p) \\
&= p^2 \log \frac{2p^2}{p^2 + (1-p)^2} + (1-p)^2 \log \frac{2(1-p)^2}{p^2 + (1-p)^2} \\
&= \Big(p^2 + (1-p)^2\Big)\left(1 - h\left(\frac{p^2}{p^2 + (1-p)^2}\right)\right)
\end{aligned}
$$

Hint: Consider the distribution $P(y,z|x)$ to get $P(y,z)$.

One interpretation of this channel is that the transmitted symbol is sent twice over a BSC.

6.13. In Figure 10 a general Z-channel is shown. Plot the capacity as a function of the error probability α.

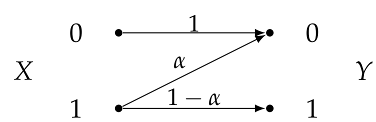


Figure 10: One BSC and one BEC in cascade.

6.14. In Figure 11 A discrete memory-less channel is given.

(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, $\{0, 1, 2\}$ and $\{3, 4, 5\}$, and construct a binary symmetric channel with error probability $p = \alpha_3 + \alpha_4 + \alpha_5$. Denote the capacity of the corresponding BSC as C_{BSC} and show that

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

where C_6 is the capacity of the channel in Figure 11.

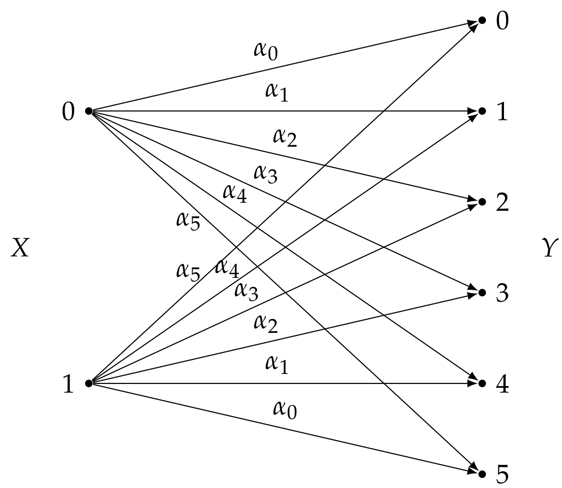


Figure 11: A channel and its probability function.

Chapter 7

7.1. In a coding scheme, three information bits $\boldsymbol{u} = (u_0, u_1, u_2)$ are appended with three parity bits according to

$$v_0 = u_1 + u_2$$
$$v_1 = u_0 + u_2$$
$$v_2 = u_0 + u_1$$

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 syndrome 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. One way to extend the the code $\mathcal{B}$ is to add one more bit such that the codeword has even Hamming weight, i.e.

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

(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?

7.4. In the early days of computers 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?

7.5. 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.

7.6. 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 $G(D)$.

Data	CRC

Figure 12: Six data bits and four bits CRC.

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 a four bit CRC is added at the end, see Figure 12.

To calculate the CRC bits the following generator polynomial is used

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

(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: $f(x) = \frac{1}{b-a}, \quad a \le x \le b.$
(b) Normal distribution: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty \le x \le \infty.$
(c) Exponential distribution: $f(x) = \lambda e^{-\lambda x}, \quad x \ge 0.$
(d) Laplace distribution: $f(x) = \frac{1}{2}\lambda e^{-\lambda|x|}, \quad -\infty \le x \le \infty.$

8.2. The joint distribution on X and Y is given by

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

for $x \ge 0$ and $y \ge 0$. (Compare with Problem 3.10)

(a) Determine α.
(b) Derive $P(X < 4, Y < 4)$.
(c) Derive the joint entropy.
(d) 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 can be modeled as a log-Normal random variable, $X \sim \text{logN}(\mu, \sigma)$. If the logarithm of a random variable X is normal distributed, i.e. $Y = \ln X \sim \text{N}(\mu, \sigma)$, 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 X and Y be two independent equally distributed random variables and form $Z = X + Y$. Derive the mutual information $I(X; Z)$ if

(a) X and Y are Gaussian with zero mean and unit variance, i.e. $X, Y \sim \text{N}(0, 1)$.

(b) X and Y are uniformly distributed between $-\frac{1}{2}$ and $\frac{1}{2}$, i.e. $X, Y \sim \text{U}(-\frac{1}{2}, \frac{1}{2})$.
Hint: $\int t \ln t \, dt = \frac{t^2}{2} \ln t - \frac{t^2}{4}$.

8.6. Show that for a continuous random variable X

$$E\big[(X - \alpha)^2\big] \geq \frac{1}{2\pi e} 2^{2H(X)}$$

for any constant α.
Hint: Use that $E\big[(X - \alpha)^2\big]$ is minimised for $\alpha = E\big[X\big]$.

8.7. The vector $X_1, X_2, \ldots, X_n$ consists of n i.i.d. Gaussian distributed random variables. Their sum is denoted $Y = \sum_i X_i$. Derive the information $I(X_k; Y)$ for the case when

(a) $X_i \sim \text{N}(0, 1)$

(b) $X_i \sim \mathrm{N}(m_i, \sigma_i)$

8.8. For a one-dimensional discrete random variable over a finite interval the uniform distribution maximises the entropy. In this problem it will be shown that this is a more general rule. Consider a finite region, $\mathcal{R}$, in N dimensions.

(a) Assume that $\boldsymbol{X} = X_1, \ldots, X_N$ is discrete valued N-dimensional random vector with probability function $p(\boldsymbol{x})$ such that

$$\sum_{\boldsymbol{x}\in\mathcal{R}} p(\boldsymbol{x}) = 1$$
$$p(\boldsymbol{x}) = 0, \boldsymbol{x} \notin \mathcal{R}$$

where $\mathcal{R}$ has finite number of outcomes, $\sum_{\boldsymbol{x}\in\mathcal{R}} 1 = k$. Show that the uniform distribution maximises the entropy over all such distributions.

(b) Assume that $\boldsymbol{X} = X_1, \ldots, X_N$ is continuous valued N-dimensional random vector with density function $f(\boldsymbol{x})$ such that

$$\int_{\mathcal{R}} f(\boldsymbol{x}) d\boldsymbol{x} = 1$$
$$f(\boldsymbol{x}) = 0, \boldsymbol{x} \notin \mathcal{R}$$

where $\mathcal{R}$ has finite volume, $\int_{\mathcal{R}} 1 d\boldsymbol{x} = V$. Show that the uniform distribution maximises the differential entropy over all such distributions.

8.9. A 2-dimensional uniform distribution is defined over the shaded area shown in Figure 13. Derive

(a) $H(X,Y)$

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

(c) $I(X;Y)$

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

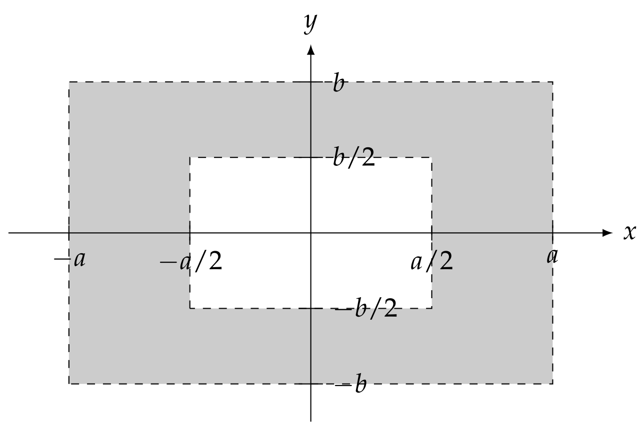


Figure 13: A 2-dimension uniform distribution.

8.10. If $h(x)$ is the density function for a Normal distribution, $\mathrm{N}(\mu, \sigma)$, and $f(x)$ any distribution with the same mean and variance, show that

$$D\big(f(x)\big|\big|h(x)\big) = H_h(X) - H_f(X)$$

8.11. The two independent Gaussian random variables X_1 and X_2 are distributed according to $\mathrm{N}(\mu_1, \sigma_1)$ and $\mathrm{N}(\mu_2, \sigma_2)$. Construct a new random variable $X = X_1 + X_2$.

(a) What is the distribution of X?

(b) Derive the differential entropy of X.

8.12. Let X_1 and X_2 be two Gaussian random variables, $X_1 \sim \mathrm{N}(\mu_1, \sigma_1)$ and $X_1 \sim \mathrm{N}(\mu_1, \sigma_1)$, with density functions $f_1(x)$ and $f_2(x)$.

(a) Derive $D(f_1||f_2)$.

(b) Let $X = X_1 + X_2$ with density function $f(x)$. Derive $D(f||f_2)$.

8.13. A random length, X, is uniformly distributed between 1 and 2 meters. Derive the differential entropy if the length is considered distributed according to the two cases below.

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

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

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

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

(c) Explain the difference.

8.14. (a) Let $X \sim \mathrm{N}(\mathbf{0}, \Lambda)$ be an n-dimensional Gaussian vector with zero mean and covariance matrix $\Lambda = E[\boldsymbol{X}\boldsymbol{X}^T]$. Use the chain rule for differential entropy to show that

$$H(\boldsymbol{X}) \leq \frac{1}{2} \log\Big((2\pi e)^n \prod_i \sigma_i^2 \Big)$$

where $\sigma_i^2 = E[X_i^2]$ is the variance for the ith variable.

(b) Use the result in (a) to show Hadamad's inequality for covariance matrices, i.e. that

$$|\Lambda| \leq \prod_i \sigma_i^2$$

Chapter 9

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

9.2. A band-limited Gaussian channel has bandwidth $W = 1$ kHz. The transmitted signal power is limited to $P = 10$ W and the noise on the channel is distributed according to $\mathrm{N}(0, \sqrt{N_0/2})$, where $N_0 = 0.01$ W/Hz. What is the channel capacity?

9.3. A random variable X is drawn from a uniform distribution $\mathrm{U}(1)$, and transmitted over a channel with additive noise Z, 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. $I(X;Y)$.

9.4. A channel consists of two channels, both with attenuation and Gaussian noise. The first channel has the attenuation H_1 and noise distribution $n_1 \sim \mathrm{N}(0, \sqrt{N_1/2})$ and the second channel has attenuation H_2 and noise distribution $n_2 \sim \mathrm{N}(0, \sqrt{N_2/2})$. 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 14. 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

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

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

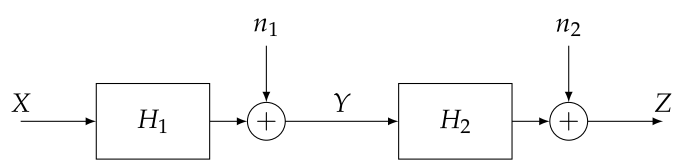


Figure 14: A channel consisting of two Gaussian channels.

9.5. A wide-band channel is split in 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

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

Derive the channel capacity.

9.6. A channel consists of 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?

9.7. An OFDM channel with five sub-channels, each occupying a bandwidth of 10 kHz. Due to regulations the allowed power level in the utilised band is -60 dBm/Hz. Each of the sub-channels have separate attenuation and noise levels according to Figure 15. Notice that the attenuation $|G_i|^2$ is given in dB and the noice $N_{0,i}$ in dBm/Hz, where i is the sub-channel. Derive the total capacity for the channel.

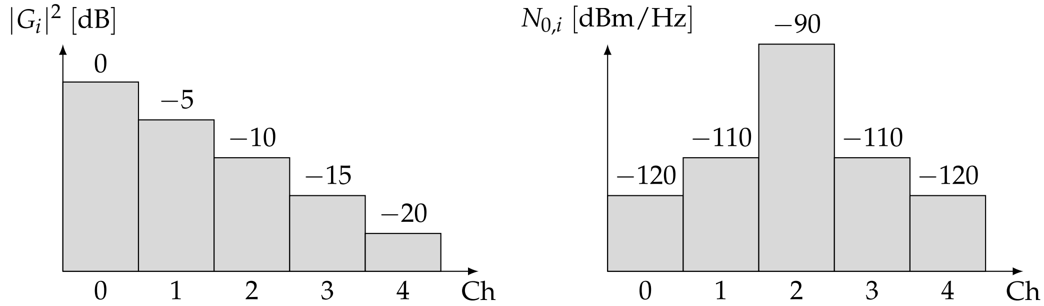


Figure 15: Attenuation and noise in the five sub-channels.

9.8. A wide-band Gaussian channel is split in four sub-channels, each with bandwidth $W_\Delta = 1$ kHz. The attenuations and noise parameters are

$$|H_i|^2 = \begin{pmatrix} -36 & -30 & -16 & -21 \end{pmatrix} \quad \text{[dB]}$$
$$N_0 = \begin{pmatrix} -108 & -129 & -96 & -124 \end{pmatrix} \quad \text{[dBm/Hz]}$$

With the total allowed power on the channel $P = -50$ dBm, what is the highest possible bit rate on the channel?

9.9. An OFDM modulation scheme is used to split a wide-band channel into 10 independent sub-bands with AWGN (Additive White Gaussian Noise). The channel parameters for the noise level N_0 and the signal attenuation $|H_i|^2$ is shown in Figure 16. The total power in the transmitted signal is allowed to be $P = -20$ dBm and the carrier spacing is $W_\Delta = 10$ kHz.

(a) Derive the capacity for the system.

(b) If the transmitted power is distributed evenly over the sub-channels, what is the capacity for the system?

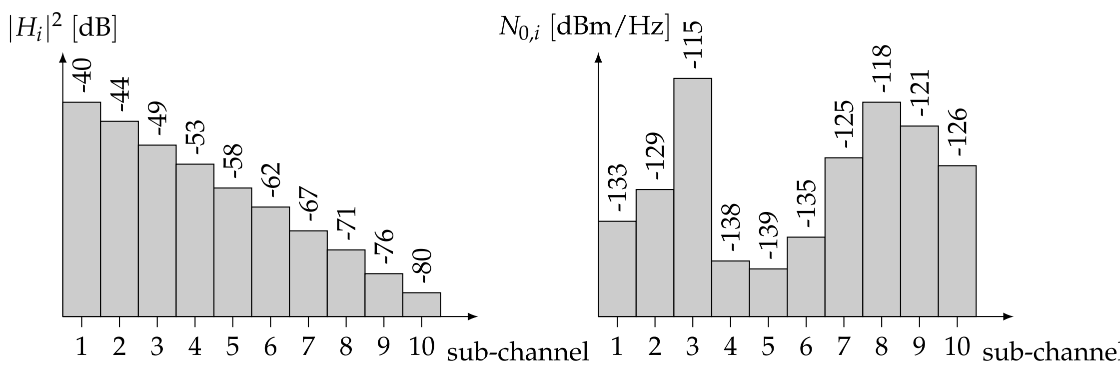


Figure 16: Channel parameters for an OFDM communication system.

9.10. A 3×3 MIMO system with maximum transmit power $P = 10$, and the noise at each receive antenna Gaussian with variance $N = 1$, has the following attenuation matrix

$$H = \begin{pmatrix} 0.05 & 0.22 & 0.73 \\ 0.67 & 0.08 & 0.60 \\ 0.98 & 0.36 & 0.45 \end{pmatrix}$$

Derive the channel capacity for the system. What is the capacity achieving distribution on the transmitted vector $\boldsymbol{X}$?

Hint: The singular value decomposition of H is given by

$$U = \begin{pmatrix} -0.36 & 0.88 & 0.30 \\ -0.59 & 0.03 & -0.81 \\ -0.72 & -0.47 & 0.51 \end{pmatrix}, \ S = \begin{pmatrix} 1.51 & 0 & 0 \\ 0 & 0.60 & 0 \\ 0 & 0 & 0.19 \end{pmatrix}, \ V = \begin{pmatrix} -0.74 & -0.65 & -0.16 \\ -0.26 & 0.05 & 0.97 \\ -0.62 & 0.76 & -0.20 \end{pmatrix}$$

9.11. The 5×4 MIMO attenuation matrix

$$H = \begin{pmatrix} 0.76 & 0.40 & 0.61 & 0.90 & 0.42 \\ 0.46 & 0.73 & 0.97 & 0.23 & 0.73 \\ 0.61 & 0.96 & 0.63 & 0.10 & 0.94 \\ 0.36 & 0.88 & 0.78 & 0.83 & 0.30 \end{pmatrix}$$

has the singular value decomposition $H = USV^T$ where

$$\text{diag}(S) = \begin{pmatrix} 2.85 & 0.89 & 0.46 & 0.30 \end{pmatrix}$$

What is the channel capacity for the system if $P = 5$ and $N = 2$?

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 $Z \sim \text{U}(\alpha)$, $E[Z] = 0$. 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 $X \in \{-2, -1, 0, 1, 2\}$ and Z is uniformly distributed $Z \sim \text{U}(-1, 1)$. Derive the capacity.

10.4. A communication scheme uses 4-PAM system, meaning there are four different signal alternatives differentiated by their amplitudes, see Figure 17. During the transmission there is a noise, Z, added to the signal so the received signal is $Y = X + Z$. The noise has the distribution as viewed in Figure 18.

(a) Assuming the signal alternatives are equally likely, how much information can be transmitted in each transmission (channel use)?

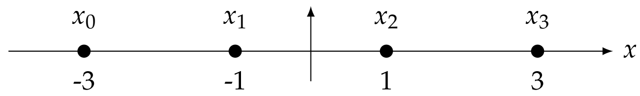


Figure 17: 4-PAM signalling.

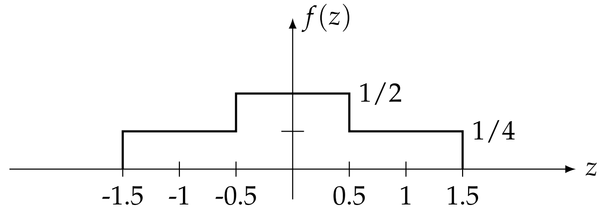


Figure 18: Density function of the additive noise.

(b) What is the capacity for the transmission, i.e. what is the maximum mutual information using the given signal alternativess and the noise. For what distribution on X can it be obtained? How is the average power of the transmitted signal affected by the optimisation?

10.5. Consider a 4-PAM communication system with the signal alternatives

$$s_x(t) = A_x\sqrt{E_g}\phi(t), \quad x = 0,1,2,3$$

where $A_x \in \{-3,-1,1,3\}$ are the amplitudes and $\phi(t)$ a normalised basis function. During the transmission the noise Z is added with the distribution $Z \sim \mathrm{N}(0,\sqrt{N_0/2})$. At the receiver each signal is decoded back to $\{0,1,2,3\}$ according to the decision regions in Figure 19.

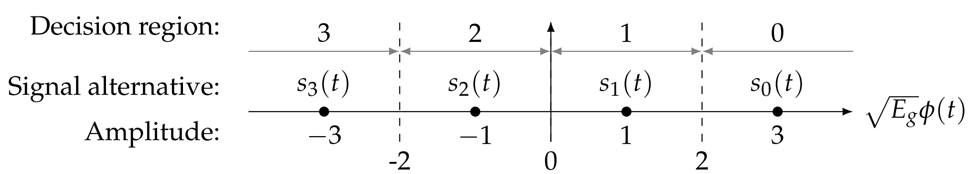


Figure 19: 4-PAM modulation and decision regions.

Denote the probability for the receiver to make an erroneous decision by

$$\varepsilon = P\big(Z > \sqrt{E_g}\big)$$

It can be assumed that the probabilities of making errors to non-neighbouring decision regions are negligible, i.e.

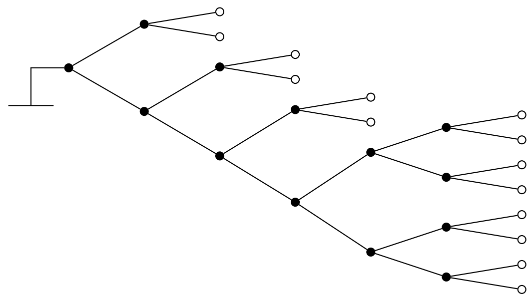
$$P\big(Z > 3\sqrt{E_g}\big) \approx 0$$

(a) Construct a corresponding DMC?

(b) Assume that the symbols are transmitted with equal probability. Express the maximum information transmission per channel use for the DMC, in terms of ε and the binary entropy function. Sketch a plot for different ε.

10.6. In Example **??** 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) Below is a tree given with two additional nodes compared with the example. What is the shaping gain for this construction?



(b) Letting the shaping constellation and tree in Example **??** 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 $\gamma_s^{(\infty)} = 0.9177\text{dB}$.
Note: It might be useful to consider the following standard sums for $|\alpha| < 1$,

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

10.7. The maximum 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 uniform distribution and an N-dimensional spheric uniform distribution, as $N \to \infty$. This will give the maximum shaping gain since the most efficient way to pack the points is as a sphere. In this problem it will be shown that the results are equivalent, since the spherical distribution projected to one dimension becomes Gaussian as $N \to \infty$.

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

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

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

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

(c) Using the first order Stirling's approximation

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

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 $X \sim N(0, \sqrt{\frac{1}{2\pi e}})$, i.e. that

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

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

10.8. Assume a transmission system uses a bandwidth of $W = 10$ kHz an the allowed signalling power is $P = 1$ mW. The channel noise can be assumed to be Gaussian with $N_0 = 1$ nW/Hz.

(a) Derive the channel capacity for the system.

(b) Give an estimate of the achievable bit rate when PAM modulation is used, if the targeted error probability is $P_e = 10^{-6}$.

10.9. Sketch a plot for the ration between the estimated achievable bit rate and the capacity for a channel with Gaussian noise. The SNR $= \frac{P}{WN_0}$ should be in the range from -10 dB to 30 dB. Make plots for different allowed bit error probabilities, e.g. $P_e \in \{10^{-3}, 10^{-6}, 10^{-9}\}$.

10.10. 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 -70 dBm/Hz. On the transmission channel the noise level is constant at -140 dBm/Hz, but the signal attenuation is increasing with the frequency as $|H_i|^2 = 5i + 10$ dB, $i = 0, \dots, 15$.

(a) Derive the capacity for the channel.

(b) If the required error rate on the channel is 10^{-6}, and it is expected that the error correcting code gives a coding gain of 3 dB, what is the estimated obtained bit rate for the system?

10.11. Consider a frequency divided system like OFDM, where the total bandwidth of $W = 10$ MHz, is split in ten equally wide independent sub-bands. In each of the sub-bands an M-PAM modulation should be used and the total allowed power is $P = 1$ W. On the sub-bands the noise to attenuation ratios are given by the vector

$$\frac{N_{o,i}}{|G_i|^2} = \begin{pmatrix} -53 & -43 & -48 & -49 & -42 & -54 & -45 & -45 & -52 & -49 \end{pmatrix} \text{ [dBm/Hz]}$$

where $N_{0,i}$ is the noise on sub-band i and G_i the signal attenuation on sub-band i. The system is supposed to work at an average error probability of $P_e = 10^{-6}$. The aim of this problem is to maximise the total information bit rate R_b for the system.

(a) Show that the total information bit rate for the system can be maximised using the water filling procedure.

(b) Derive the maximum information bit rate for the system.

(c) Assume that you add an error correcting code in each sub-band, with a coding gain of $\gamma_c = 3$ dB. How does that influence the maximum information bit rate?

Chapter 11

11.1. Consider a k-ary source with 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 (Hint: Fano's lemma can be useful.)

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

11.2. In this problem it is shown that the exponential distribution maximises the entropy over all one-sided distribution with fixed mean. The idea is to consider a general distribution $f(x)$ such that $f(x) \geq 0$, with equality for $x < 0$, and $E[X] = 1/\lambda$. The result is obtained by maximising the entropy $H(X) = -\int_0^\infty f(x) \ln f(x) dx$ with respect to the requirements

$$\begin{cases} \int_0^\infty f(x) dx = 1 \\ \int_0^\infty x f(x) dx = 1/\lambda \end{cases} \tag{3}$$

For simplicity the derivation will be performed over the natural base in the logarithm.

(a) Set up a maximisation function in f according to Lagrange multiplier method. (Notice that you will need two side conditions). Differentiate with respect to the function f, and equal to zero. Show that this gives an optimising function on the form

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

and that the equation system in (3) is solved by the exponential distribution.

(b) Let $g(x)$ be an arbitrary distribution with the same requirements as in (3). Show that

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

i.e. that $f(x) = \lambda e^{-\lambda x}$ maximises the entropy.

11.3. To derive the rate distortion function for the exponential distribution, $X \sim \text{Exp}(\lambda)$, i.e. $f(x) = \lambda e^{-\lambda x}$, first introduce a distortion measure and bound the mutual information. Then define a test channel from $\hat{X}$ to X to show that the bound can be obtained with equality. Use the natural logarithm in the derivations.

(a) 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, see Problem 11.2, to show that $I(X; \hat{X}) \geq -\ln(\lambda\delta)$ where $E[d(x, \hat{x})] \leq \delta$.

(b) Define a test channel from $\hat{X}$ to X in order to show equality in the bound for the mutual information. To derive the distribution on X, first show that the Laplace transform of the density function for the exponential distribution, $T \sim \text{Exp}(\lambda)$ where $f_T(t) = \lambda e^{-\lambda t}$, is

$$E\big[e^{sT}\big] = \frac{1}{1 + s/\lambda}$$

Derive the distribution on X and argue that the rate distortion function is

$$R(\delta) = \begin{cases} -\ln(\lambda\delta), & 0 \le \delta \le 1/\lambda \\ 0, & \delta > 1/\lambda \end{cases}$$

11.4. A source is given with i.i.d. symbols generated according to Laplacian distribution,

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

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

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

11.5. The source variable $X \sim \text{N}(0, \sqrt{2})$ 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?

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

11.7. A random variable X is Gaussian distributed with zero mean and unit variance, $X \sim \text{N}(0, 1)$. Let the outcome from this variable be quantised by a uniform quantiser with eight intervals of width Δ. That is, the quantisation intervals are given by the limits $\{-3\Delta, -2\Delta, -\Delta, 0, \Delta, 2\Delta, 3\Delta\}$ and the reconstruction values $x_Q \in \{-\frac{7\Delta}{2}, -\frac{5\Delta}{2}, -\frac{3\Delta}{2}, -\frac{\Delta}{2}, \frac{\Delta}{2}, \frac{3\Delta}{2}, \frac{5\Delta}{2}, \frac{7\Delta}{2}\}$. Notice that since the Gaussian distribution has infinite width, i.e. $-\infty \le x \le \infty$, the outer most intervals also have infinite width. That means any value exceeding the upper limit 3Δ will be reconstructed to the highest reconstruction value $\frac{7\Delta}{2}$, and vice versa for the lowest values. If Δ is very small the outer regions, the clipping regions, will have high probabilities and the quantisation error will be high. On the other hand, if Δ is large, so the clipping region will have very low probability, the quantisation intervals will grow and this will also give high quantisation error.

(a) Sketch a plot of the squared quantisation error when varying the quantisation interval Δ.

(b) Find the Δ that minimises the quantisation error

Note: The solution requires numerical calculation of the integrals.

11.8. Use Jensen's inequality to show that the arithmetic mean exceeds the geometric mean, i.e. that

$$\frac{1}{n}\sum_i a_i \geq \Big(\prod_i a_i\Big)^{1/n} \tag{4}$$

for any set of positive real numbers $a_1, a_2, \ldots, a_n$.

Solutions

Chapter 2

2.1. (a) From the problem formulation

$$P(X=0) = P(X=1) = \tfrac{1}{2} \text{ and } P(Y=0) = p, P(Y=1) = 1-p$$

First derive the unconditional probability of Z as

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

Then the conditional probability

$$P(Z=i|Y=j) = P(X \oplus Y = i|Y=j) = P(X = i \oplus j) = \tfrac{1}{2}.$$

That is, $P(Z=i|Y=j) = P(Z=i)$ for all $i,j \in \{0,1\}$.

(b) Similar to (a), for all $i,j \in \{0,1,\ldots,M-1\}$ let

$$P(X=i) = \frac{1}{M} \text{ and } P(Y=j) \quad = p_j, \text{ where } \sum_i p_i = 1$$

To simplify notations, let $\langle k\rangle_M$ denotes the reminder when k is divided by M. That is, if $\langle k\rangle_M = a$ then $k \equiv a, \mod M$. The unconditional probability of Z is

$$\begin{aligned} P(Z=i) &= \sum_j P(Z=i|Y=j)P(Y=j) = \sum_j P(\langle X+Y\rangle_M = i|Y=j)P(Y=j) \\ &= \sum_j P(X = \langle i-j\rangle_M)P(Y=j) = \tfrac{1}{M}\sum_j P(Y=j) = \tfrac{1}{M} \end{aligned}$$

When Y is known the probability becomes

$$P(Z=i|Y=j) = P(\langle X+Y\rangle_M = i|Y=j) = P(X = \langle i-j\rangle_M) = \tfrac{1}{M}$$

which shows that Z is independent of Y.

2.2. *Alternative 1*. Let A and B be the events that the first is not heart and the second card is not a heart, respectively. 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, so $P(B|A) = 38/51$. 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 combinatorics, the total number of cases of two cards taken from 52 is $\binom{52}{2}$, and the number of pairs with no hearts is $\binom{39}{2}$. 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,

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

where in the last equality it is used that $\sum_k \binom{n}{k}^2 = \binom{2n}{n}$.

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. Both X and Y are uniformly distributed with probability functions

$$p_X(k) = \tfrac{1}{5}, \quad k = 1, \dots, 5 \text{ and } p_Y(k) = \tfrac{1}{8}, \quad k = 1, \dots, 8$$

Then $p_{Z_a}(k)$ is described by the convolution of $p_X(k)$ and $p_Y(k)$. Since the uniform distribution is symmetric $-Y$ has the probability function

$$p_{-Y}(k) = \tfrac{1}{8}, \quad k = -8, \dots, -1$$

and $P_{Z_b}(k)$ is described by the convolution of $p_X(k)$ and $p_{-Y}(k)$. Finally $p_{Z_c}(k)$ is obtained by folding the negativ axis on to the positiv for $p_{Z_b}(k)$. The distributions are shown in the tables below and in Figure 20.

k	$p_{Z_a}(k)$
2	0.0250
3	0.0500
4	0.0750
5	0.1000
6	0.1250
7	0.1250
8	0.1250
9	0.1250
10	0.1000
11	0.0750
12	0.0500
13	0.0250

k	$p_{Z_b}(k)$
-7	0.0250
-6	0.0500
-5	0.0750
-4	0.1000
-3	0.1250
-2	0.1250
-1	0.1250
0	0.1250
1	0.1000
2	0.0750
3	0.0500
4	0.0250

k	$p_{Z_c}(k)$
0	0.1250
1	0.2250
2	0.2000
3	0.1750
4	0.1250
5	0.0750
6	0.0500
7	0.0250

Alternatively, the distributions can be derived by first considering tables of the functions, as shown below, and then counting the number of occurrences.

Z_a	1	2	3	4	5	6	7	8
1	2	3	4	5	6	7	8	9
2	3	4	5	6	7	8	9	10
3	4	5	6	7	8	9	10	11
4	5	6	7	8	9	10	11	12
5	6	7	8	9	10	11	12	13

Z_b	1	2	3	4	5	6	7	8
1	0	-1	-2	-3	-4	-5	-6	-7
2	1	0	-1	-2	-3	-4	-5	-6
3	2	1	0	-1	-2	-3	-4	-5
4	3	2	1	0	-1	-2	-3	-4
5	4	3	2	1	0	-1	-2	-3

Z_c	1	2	3	4	5	6	7	8
1	0	1	2	3	4	5	6	7
2	1	0	1	2	3	4	5	6
3	2	1	0	1	2	3	4	5
4	3	2	1	0	1	2	3	4
5	4	3	2	1	0	1	2	3

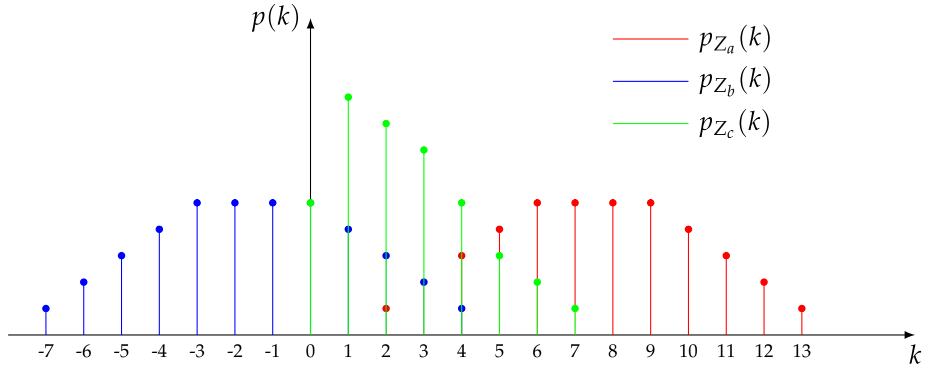


Figure 20: Probability functions for Z_a, Z_b and Z_c.

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

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

2.6. With $p(k) = \dfrac{\lambda^k e^{-\lambda}}{k!}$, $k = 0, 1, 2, \ldots$,

$$E\big[X\big] = \sum_{k=0}^{\infty} k\frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=1}^{\infty} k\frac{\lambda^k e^{-\lambda}}{k!} = \lambda \underbrace{\sum_{k=1}^{\infty} \frac{\lambda^{k-1} e^{-\lambda}}{(k-1)!}}_{=1} = \lambda$$

$$E\big[X^2\big] = \sum_{k=0}^{\infty} k^2\frac{\lambda^k e^{-\lambda}}{k!} = \lambda \sum_{k=1}^{\infty} k\frac{\lambda^{k-1} e^{-\lambda}}{(k-1)!} = \lambda\Bigg(\underbrace{\sum_{k=1}^{\infty}(k-1)\frac{\lambda^{k-1} e^{-\lambda}}{(k-1)!}}_{=\lambda} + \underbrace{\sum_{k=1}^{\infty}\frac{\lambda^{k-1} e^{-\lambda}}{(k-1)!}}_{=1}\Bigg) = \lambda^2 + \lambda$$

$$V\big[X\big] = E\big[X^2\big] - E\big[X\big]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

2.7. With $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$,

$$E\big[X\big] = \int_0^{\infty} x\lambda e^{-\lambda x} dx = \lambda\Bigg(\underbrace{\Big[x\frac{e^{-\lambda x}}{-\lambda}\Big]_0^{\infty}}_{=0} - \int_0^{\infty} \frac{e^{-\lambda x}}{-\lambda} dx\Bigg) = \frac{1}{\lambda}\underbrace{\int_0^{\infty} \lambda e^{-\lambda x} dx}_{=1} = \frac{1}{\lambda}$$

$$\begin{aligned}
E\big[X^2\big] &= \int_0^{\infty} x^2\lambda e^{-\lambda x} dx = \lambda\Bigg(\underbrace{\Big[x^2\frac{e^{-\lambda x}}{-\lambda}\Big]_0^{\infty}}_{=0} - \int_0^{\infty} 2x\frac{e^{-\lambda x}}{-\lambda} dx\Bigg) \\
&= 2\int_0^{\infty} x e^{-\lambda x} dx = \frac{2}{\lambda}\underbrace{\int_0^{\infty} x\lambda e^{-\lambda x} dx}_{=1/\lambda} = \frac{2}{\lambda^2}
\end{aligned}$$

$$V\big[X\big] = E\big[X^2\big] - E\big[X\big]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

2.8. This can be solved in many ways. Here two alternatives are given. For the first alternative, set the derivative of $E\big[(X-c)^2\big] = E\big[X^2\big] - 2mc + c^2$ equal to zero,

$$\frac{\partial}{\partial c}E\big[(X-c)^2\big] = -2m + 2c = 0$$

which gives $c = m$ as an optima. Since the second derivative is $\frac{\partial^2}{\partial c^2}E\big[(X-c)^2\big] = 2$ is positive it is a minimum. Hence,

$$E\big[(X-c)^2\big] \geq E\big[(X-m)^2\big]$$

A second alternative can be done as follows,

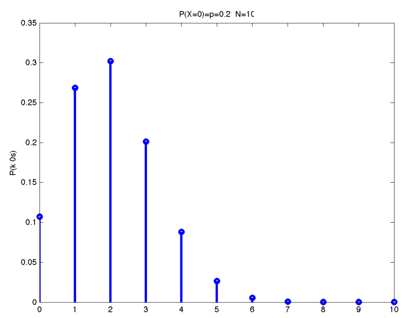
$$\begin{aligned}
E\big[(X-c)^2\big] - E\big[(X-m)^2\big] &= E\big[X^2\big] - 2E\big[X\big]c + c^2 - E\big[X^2\big] + 2E\big[X\big]m - m^2 \\
&= E\big[X^2\big] - 2mc + c^2 - E\big[X^2\big] + 2m^2 - m^2 \\
&= m^2 - 2mc + c^2 = (m-c)^2 \geq 0
\end{aligned}$$

with equality if and only if $c = m$. The expresion is equivalent to $E\big[(X-c)^2\big] \geq E\big[(X-m)^2\big]$ with equality if and only if $c = m$.

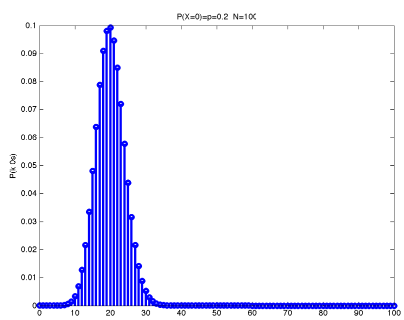
2.9. Let k be the number of 0s in the vector. Then we get the following table.

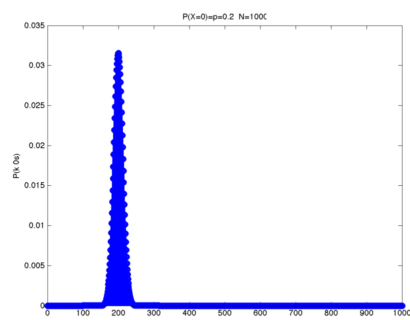
k	$\binom{10}{k}$	$P(\boldsymbol{x}\|k\,0s) = p^k(1-p)^{(10-k)}$	$P(k\,0s) = \binom{10}{k}P(\boldsymbol{x}\|k\,0s)$
0	1	0.1073741824	0.1073741824
1	10	0.0268435456	0.2684354560
2	45	0.0067108864	0.3019898880
3	120	0.0016777216	0.2013265920
4	210	0.0004194304	0.0880803840
5	252	0.0001048576	0.0264241152
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 two next pictures show 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.





2.10. (a) Since the drawn ball is replaced the six results are independent, and all hace $P(\text{black}) = p = 0.3$ and $P(\text{white}) = 1 - p = 0.7$. The resulting distribution is

$$P(k\,\text{black}) = \binom{6}{k}(1-p)^{6-k}p^k$$

which gives

k	$P(k\,\text{black})$
0	0.1176
1	0.3025
2	0.3241
3	0.1852
4	0.0595
5	0.0102
6	0.0007

(b) Draw *all* balls from the urn and place them on a line. If there are k black balls among the six first in the row, then there are $3-k$ among the last four. The first six balls can be drawn in $\binom{6}{k}$ different ways. For each of them the last four balls can be arranged in $\binom{4}{3-k}$ ways. That is, there are $\binom{6}{k}\binom{4}{3-k}$ alternatives to get k black balls in the six first draws. In total there are $\binom{10}{3} = 120$ different ways to arrange the ten balls, so the probability of having k black balls in the first six draws is

$$P(k\,\text{black}) = \frac{\binom{6}{k}\binom{4}{3-k}}{120}$$

which gives

k	$\binom{6}{k}\binom{4}{3-k}$	$P(k\,\text{black})$
0	4	0.0333
1	36	0.3
2	60	0.5
3	20	0.1667

2.11. Choose $p_1 = p_2 = \frac{1}{2}$ and $f(\cdot) = \log(\cdot)$ to get

$$\begin{aligned}
\frac{1}{2}\log x_1 + \frac{1}{2}\log x_2 &\leq \log\left(\frac{1}{2}x_1 + \frac{1}{2}x_2\right) \\
\log(x_1x_2)^{\frac{1}{2}} &\leq \log\frac{x_1+x_2}{2} \\
(x_1x_2)^{\frac{1}{2}} &\leq \frac{x_1+x_2}{2}
\end{aligned}$$

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

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

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

2.12. (a) Let $y = \ln n! = \sum_{k=1}^{n} \ln k$. However, by using trapezoid approximation the integral

$$\int_1^n \ln x dx \approx \sum_{k=1}^{n} -\frac{1}{2}\ln 1 - \frac{1}{2}\log n = \sum_{k=1}^{n} -\frac{1}{2}\ln n$$

The integral can also be derived by using $\frac{\partial}{\partial x} x \ln x = \ln x + 1$,

$$\int_1^n \ln x dx = \Big[x \ln x - x\Big]_1^n = n \ln n - n + 1$$

Thus,

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

or, equivalently,

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

By applying the exponential function on both sides the required result is obtained.

(b) As n tends to infinity

$$\lim_{n\to\infty} \frac{n!}{e\sqrt{n}\left(\frac{n}{e}\right)^n} = \lim_{n\to\infty} \frac{1}{e} \frac{n! e^n}{n^n \sqrt{n}} = \frac{\sqrt{2\pi}}{e}$$

Hence, the factor e in the result should be replaced with $\sqrt{2\pi}$ to get

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

which achieves equality in the limit as $n \to \infty$.

(c) The error made in (a) is correction factor

$$\frac{\sqrt{2\pi}}{e} \approx 0.92$$

So, the error is about 10%. For large n this is often not severe. For example, the approximations for 50! becomes $3.3 \cdot 10^{64}$ or $3.0 \cdot 10^{64}$, and 100! becomes $1.0 \cdot 10^{158}$ or $9.4 \cdot 10^{157}$.

2.13. (a) The state transition matrix is

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

(b) To derive the steady state distribution start with the matrix

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

Replace the first column with ones to represent the equation $\sum_k \pi_k = 1$,

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

Then $\boldsymbol{\pi} B = (1\,0\,0\,0)$, and consequently

$$\boldsymbol{\pi} = (1\,0\,0\,0) B^{-1} = \begin{pmatrix} \frac{4}{15} & \frac{4}{15} & \frac{1}{5} & \frac{4}{15} \end{pmatrix}$$

2.14. (a) The state transition matrix is

$$P = \begin{pmatrix} 1 - P_{gb} & P_{gb} \\ P_{bg} & 1 - P_{bg} \end{pmatrix}$$

To solve the equation system

$$\begin{cases} \boldsymbol{\pi} P = \boldsymbol{\pi} \\ \sum_i \pi_i = 1 \end{cases}$$

start to derive

$$P - I = \begin{pmatrix} -P_{gb} & P_{gb} \\ P_{bg} & -P_{bg} \end{pmatrix}$$

Then, replace the first column with the all-one vector,

$$A = \begin{pmatrix} 1 & P_{gb} \\ 1 & -P_{bg} \end{pmatrix}$$

and solve $\boldsymbol{\pi} A = (1 \ \ 0)$ for $\boldsymbol{\pi}$ as

$$\boldsymbol{\pi} = (1 \ \ 0) A^{-1} = (1 \ \ 0) \begin{pmatrix} \frac{P_{bg}}{P_{bg}+P_{gb}} & \frac{P_{gb}}{P_{bg}+P_{gb}} \\ \frac{1}{P_{bg}+P_{gb}} & \frac{-1}{P_{bg}+P_{gb}} \end{pmatrix} = \begin{pmatrix} \frac{P_{bg}}{P_{bg}+P_{gb}} & \frac{P_{gb}}{P_{bg}+P_{gb}} \end{pmatrix}$$

which is the steady state solution.

(b) The average time of a burst is given by the average length of staying in the state *Bad*. If the first decision is to leave state, the burst length will be one. If the first decision is t stay in the state and the second to leave the burst length is two. Similarly, if there are $k-1$ consecutive decisions to stay followed by one to leave, the burst length will be k. Hence,

$$E\left[L_{\text{Bad}}\right] = \sum_{k=0}^{\infty} k(1 - P_{bg})^{k-1} P_{bg} = P_{bg} \sum_{k=0}^{\infty} k(1 - P_{bg})^{k-1} = \frac{1}{P_{bg}}$$

(c) The average time between two consecutive bursts is given by the average length of staying in state *Good*,

$$E\left[L_{\text{Good}}\right] = \sum_{k=0}^{\infty} k(1 - P_{gb})^{k-1} P_{gb} = \frac{1}{P_{gb}}$$

2.15. At steady state the state probabilities can be written as

$$\begin{cases} \pi_0 &= \pi_0 q + \pi_1 p \\ \pi_1 &= \pi_0 p + \pi_2 q \\ \pi_2 &= \pi_1 p + \pi_3 q \\ & \vdots \\ \pi_k &= \pi_{k-1} p + \pi_{k+1} q \\ & \vdots \end{cases}$$

The first equation gives

$$\pi_1 = \frac{1}{q}(\pi_0 - \pi_0 q) = \frac{1-q}{q}\pi_0 = \frac{p}{q}\pi_0$$

With this at hand the second equation can be rewritten as

$$\pi_2 = \frac{1}{q}(\pi_1 - \pi_0 p) = \Big(\frac{p}{q^2} - \frac{p}{q}\Big)\pi_0 = \frac{p(1-q)}{q^2}\pi_0 = \frac{p^2}{q^2}\pi_0$$

Similarly, the third equation gives

$$\pi_3 = \frac{1}{q}(\pi_2 - \pi_1 p) = \Big(\frac{p^2}{q^3} - \frac{p^2}{q^2}\Big)\pi_0 = \frac{p^2(1-q)}{q^3}\pi_0 = \frac{p^3}{q^3}\pi_0$$

To show that this pattern continues assume that up to state s_k, $\pi_k = \frac{p^k}{q^k}\pi_0$. Then, the equation for π_k gives

$$\pi_{k+1} = \frac{1}{q}(\pi_k - \pi_{k-1} p) = \Big(\frac{p^k}{q^{k+1}} - \frac{p^k}{q^k}\Big)\pi_0 = \frac{p^k(1-q)}{q^{k+1}}\pi_0 = \frac{p^{k+1}}{q^{k+1}}\pi_0$$

That is, $\pi_k = \big(\frac{p}{q}\big)^k\pi_0$, for $k = 0,1,2,\dots$. Finally, to derive π_0 use $\sum_k \pi_k = 1$ and $p < q$ to get

$$1 = \sum_{k=0}^{\infty}\pi_k = \pi_0\sum_{k=0}^{\infty}\Big(\frac{p}{q}\Big)^k = \pi_0\frac{1}{1-\frac{p}{q}} \quad\Rightarrow\quad \pi_0 = 1 - \frac{p}{q}$$

Concluding, the steady state distribution is given by

$$\pi_k = \Big(1 - \frac{p}{q}\Big)\Big(\frac{p}{q}\Big)^k$$

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. According to the IT-inequality

$$\ln x \le x - 1, \quad x \ge 0$$

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

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

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

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

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

$P(X|Y)$

X	$Y=a$	$Y=b$	$Y=c$
0	$\frac{1}{4}$	1	$\frac{2}{3}$
1	$\frac{3}{4}$	0	$\frac{1}{3}$

$P(Y|X)$

X	$Y=a$	$Y=b$	$Y=c$
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) $I(X;Y) \approx 0.2503$

3.6. (a) The probability functions are:

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

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

$P(X|Y)$

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

$P(Y|X)$

X	$Y=a$	$Y=b$	$Y=c$
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. 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 21.

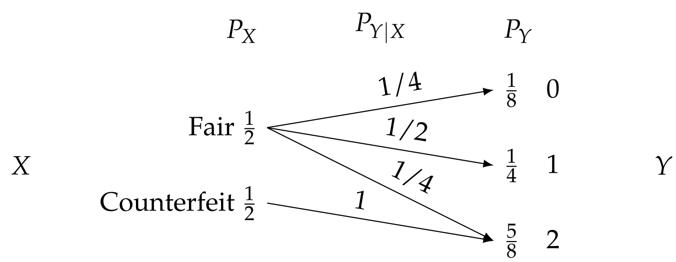


Figure 21: Probabilities or two flips with unknown coin.

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

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

3.8. (a) $H(X) = H(\frac{10}{18}, \frac{5}{18}, \frac{3}{18}) \approx 1.4153$ bit

(b) $H(Y|X) = H(Y|X = b)P(X = b) + H(Y|X = r)P(X = r) + H(Y|X = g)P(X = g)$

$= H(\frac{9}{17}, \frac{5}{17}, \frac{3}{17})\frac{10}{18} + H(\frac{10}{17}, \frac{4}{17}, \frac{3}{17})\frac{5}{18} + H(\frac{10}{17}, \frac{5}{17}, \frac{2}{17})\frac{3}{18} \approx 1.4100$ bit

(c) If X is not known the probabilities of Y are based on the original set of outcomes, i.e. the same as for X. To see this first derive $p(x, y) = p(x)p(y|x)$ in the table below

$P(X, Y)$

X	$Y = b$	$Y = r$	$Y = g$
b	$\frac{10}{18}\frac{9}{17}$	$\frac{10}{18}\frac{5}{17}$	$\frac{10}{18}\frac{3}{17}$
r	$\frac{5}{18}\frac{10}{17}$	$\frac{5}{18}\frac{4}{17}$	$\frac{5}{18}\frac{3}{17}$
g	$\frac{3}{18}\frac{10}{17}$	$\frac{3}{18}\frac{5}{17}$	$\frac{3}{18}\frac{2}{17}$

To get the probabilities of Y use $P(Y = c) = \sum_x P(X = x, Y = c)$, i.e. sum vertically in the table. Similarly, if the table is summed horizontally we get $P(X)$. Both variants give the same results,

$P(X) = P(Y) = (\frac{10}{18}, \frac{5}{18}, \frac{3}{18})$

In other words,

$H(Y) = H(\frac{10}{18}, \frac{5}{18}, \frac{3}{18}) \approx 1.4153$ bit

(d) $I(X; Y) = H(Y) - H(Y|X) = 0.0052$ bit

Naturally, it can also be derived as $I(X; Y) = H(X) - H(X|Y)$ where $P(X|Y)$ is derived from the joint distribution above. Then, in this case, $H(X|Y) = H(Y|X)$ and the result is the same.

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) Use that the sum over all x and y equals 1,

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

(b)

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

(c)

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

$$= 2 + \sum_{x,y} x\frac{1}{4}2^{-(x+y)} + \sum_{x,y} y\frac{1}{4}2^{-(x+y)} = 2 + 2\sum_x x\frac{1}{2}2^{-x}\underbrace{\sum_y \frac{1}{2}2^{-y}}_{=1}$$

$$= 2 + 2\underbrace{\sum_x x\frac{1}{2}2^{-x}}_{=1} = 4$$

(d) To start with derive the marginals as

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

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,

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

3.11. 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.

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

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

3.13. (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)

3.14. 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. Since the choice of urn are equally likely the joint probability is $p(x,y) = \frac{1}{2}p(x|y)$.

$P(X\|Y)$		
X	$Y=1$	$Y=2$
W	$4/7$	$3/10$
B	$3/7$	$7/10$

$P(X,Y)$		
X	$Y=1$	$Y=2$
W	$4/14$	$3/20$
B	$3/14$	$7/20$

(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 $p(x) = 1/3$)

$P(X\|Y)$			
X	$Y=1$	$Y=2$	$Y=3$
W	$4/7$	$3/10$	1
B	$3/7$	$7/10$	0

$P(X,Y)$			
X	$Y=1$	$Y=2$	$Y=3$
W	$4/21$	$3/30$	$1/3$
B	$3/21$	$7/30$	0

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

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

3.15.

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

3.16. (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.

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

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

This requires that $\big(\sum_x \frac{p^2(x)}{q(x)} - 1\big) \ge 0$ 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 $p(x) = q(x)$.

3.18. (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

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

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

3.19. (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

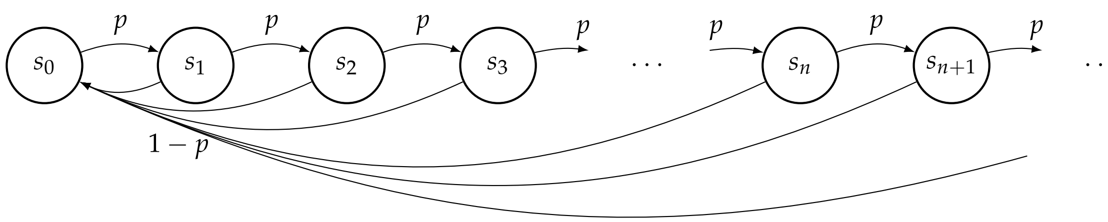
$$\boldsymbol{\mu} = \begin{pmatrix}\frac{1}{3} & \frac{1}{3} & \frac{2}{9} & \frac{1}{9}\end{pmatrix}$$

which is the distribution of the islands.

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

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

3.20. (a) Let the state of the Markov process be the step on ladder. Then the (infinite) state transition graph for the process is



This gives the transition matrix

$$P = \begin{pmatrix} 1-p & p & 0 & 0 & 0 & \cdots \\ 1-p & 0 & p & 0 & 0 & \cdots \\ 1-p & 0 & 0 & p & 0 & \cdots \\ \vdots & & & & \ddots & \ddots \end{pmatrix}$$

(b) Letting S denote the state and S^+ the state at the next time instant. At each state the entropy $H(S^+|S) = h(p)$. With $\boldsymbol{\pi} = \pi_0, \pi_1, \pi_2, \ldots,$ denoting the steady state distribution, the entropy rate is

$$H_\infty(S) = \sum_S H(S^+|S)P(S) = \sum_{i=0}^{\infty} h(p)\pi_i = h(p)$$

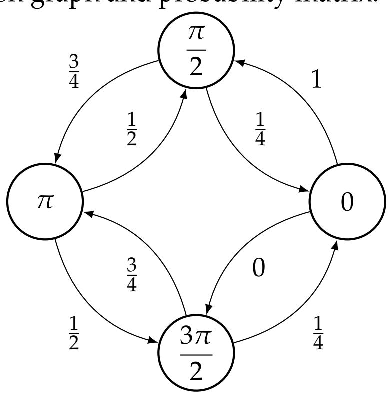
(c) In this problem we need the steady state distribution $\boldsymbol{\pi}$. From (a) we get that $\pi_n = \pi_{n-1}p = \pi_{n-2}p^2 = \pi_0 p^n$ for $n = 0, 1, 2, \ldots$. With $1 = \sum_i \pi_i = \pi_0 \sum_i p^i = \pi_0 \frac{1}{1-p}$ we conclude $\pi_n = (1-p)p^n$. The uncertainty that the man is on the ground is then

$$H(S = 0) = h(p)$$

To get the uncertainty of the step when the man is not on the ground we first need the corresponding probability as $\nu_n = \frac{\pi_n}{1-\pi_0} = (1-p)p^{n-1}$ for $n = 1, 2, \ldots$. Hence the uncertainty is

$$\begin{aligned} H(N) &= -\sum_{i=1}^{\infty}(1-p)p^{i-1}\log(1-p)p^{i-1} = -\sum_{j=0}^{\infty}(1-p)p^j\log((1-p)p^j) \\ &= -(1-p)\log(1-p)\sum_{j=0}^{\infty}p^j - p(1-p)\log p\sum_{j=0}^{\infty}jp^{j-1} \\ &= -\log(1-p) - \frac{p}{1-p}\log p = \frac{h(p)}{1-p} \end{aligned}$$

3.21. Since the process has unit memory it can be modeled as a Markov process with the following state transition graph and probability matrix.



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

Exchange the last column i $P - I$ with the all-one vector to get

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

Then, solve the equation $\boldsymbol{\pi}A = \begin{pmatrix} 0 & 0 & 0 & 1 \end{pmatrix}$ to get the steady state solution,

$$\boldsymbol{\pi} = \begin{pmatrix} 0 & 0 & 0 & 1 \end{pmatrix} A^{-1} = \begin{pmatrix} \frac{1}{8} & \frac{5}{16} & \frac{3}{8} & \frac{3}{16} \end{pmatrix}$$

The entropy rate becomes

$$\begin{aligned} H_\infty(X) &= \sum_{\varphi} H(\Phi^+|\Phi = \varphi)\pi_\varphi \\ &= 0\frac{1}{8} + h(\tfrac{1}{4})\frac{5}{16} + h(\tfrac{1}{2})\frac{3}{8} + h(\tfrac{1}{4})\frac{3}{16} \\ &= \frac{3}{8} + \frac{1}{2}h(\tfrac{1}{4}) \approx 0.78 \text{ bit/symbol} \end{aligned}$$

3.22. (a)

$$\begin{aligned} H(X_1\ldots X_n) &= \sum_{i=1}^{n} H(X_i|X_1\ldots X_{i-1}) = \sum_{i=1}^{n} H(X_n|X_i\ldots X_{n-1}) \\ &\geq \sum_{i=1}^{n} H(X_n|X_1\ldots X_{n-1}) = nH(X_n|X_1\ldots X_{n-1}) \end{aligned}$$

Dividing with n gives $H_n(X) \geq H(X|X^n)$.

(b)

$$\begin{aligned}
H(X_1 \dots X_n) &= H(X_1 \dots X_{n-1}) + H(X_n | X_1 \dots X_{n-1}) \\
&\leq H(X_1 \dots X_{n-1}) + H(X_n | X_2 \dots X_{n-1}) \\
&= H(X_1 \dots X_{n-1}) + H(X_{n-1} | X_1 \dots X_{n-2}) \\
&\leq H(X_1 \dots X_{n-1}) + \frac{1}{n-1} H(X_1 \dots X_{n-1}) \\
&= \frac{n}{n-1} H(X_1 \dots X_{n-1})
\end{aligned}$$

Dividing with n gives $H_n(X) \leq H_{n-1}(X)$.

(c)

$$\begin{aligned}
H(X_1 \dots X_n) &= \sum_{i=1}^{n} H(X_i | X_1 \dots X_{i-1}) \\
&= \sum_{i=1}^{\mu} H(X_i | X_1 \dots X_{i-1}) + \sum_{i=\mu+1}^{n} H(X_i | X_1 \dots X_{i-1}) \\
&\leq \sum_{i=1}^{\mu} H(X_i | X_1 \dots X_{i-1}) + \sum_{i=\mu+1}^{n} H(X_i | \underbrace{X_{i-\mu+1} \dots X_{i-1}}_{\mu-1}) \\
&= H(X_1 \dots X_\mu) + \sum_{i=1}^{\mu} H(X_\mu | X_1 \dots X_{\mu-1}) \\
&= H(X_1 \dots X_\mu) + (n-\mu) H(X_\mu | X_1 \dots X_{\mu-1})
\end{aligned}$$

Dividing with n gives

$$\begin{aligned}
H_n(X) &= \frac{1}{n} H(X_1 \dots X_n) \\
&\leq \frac{1}{n} H(X_1 \dots X_\mu) + \frac{n-\mu}{n} H(X_\mu | X_1 \dots X_{\mu-1}) \\
&= \frac{\mu}{n} H_\mu(X) + \frac{n-\mu}{n} H(X | X^\mu)
\end{aligned}$$

(d) The left hand side inequality follows directly from (a) as $n \to \infty$.
With $\mu \ll n$, $\lim_{n\to\infty} \frac{\mu}{n} H_\mu(X) = 0$ and $\lim_{n\to\infty} \frac{n-\mu}{n} H(X|X^\mu) = H(X|X^\mu)$. This gives that the right hand side inequality follows from (c) as $n \to \infty$.

Chapter 4

4.1. (a) Yes, all codewords are different.

(b) No, for example 0100 can be decoded as 0,10,0 or 01,0,0.

(c) No, 0 is a prefix to 01.

4.2. (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.

4.3. (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 $H(X)\leq E\big[L\big]\leq H(X)+1$.

(d) 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 is not optimal!

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

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

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. There are more than one alternative for each of the distributions. For (a) two codes are given but the (b) and (c) only one each. For a given distribution the average codeword length is always the same for all alternatives.

(a)

x	y	$y^{(2)}$
x_1	00	00
x_2	10	010
x_3	01	1
x_4	11	011

$L = 2$ bit
$H(X) = 1.9056$ bit

(b)

x	y
x_1	00
x_2	01
x_3	010
x_4	11
x_5	011

$L = 2 + \frac{5}{18} \approx 2.2778$ bit
$H(X) = 2.2244$ bit

(c)

x	y
x_1	00
x_2	10
x_3	010
x_4	110
x_5	111
x_6	0110
x_7	0111

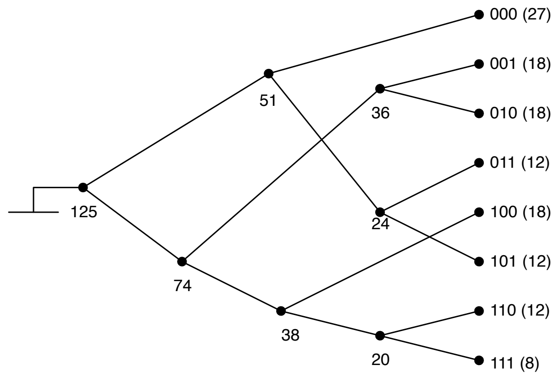
$L = 2.6$ bit
$H(X) = 2.5435$ bit

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

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

$\boldsymbol{x}$	$p(\boldsymbol{x})$	$\ell_{\boldsymbol{x}}$
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):

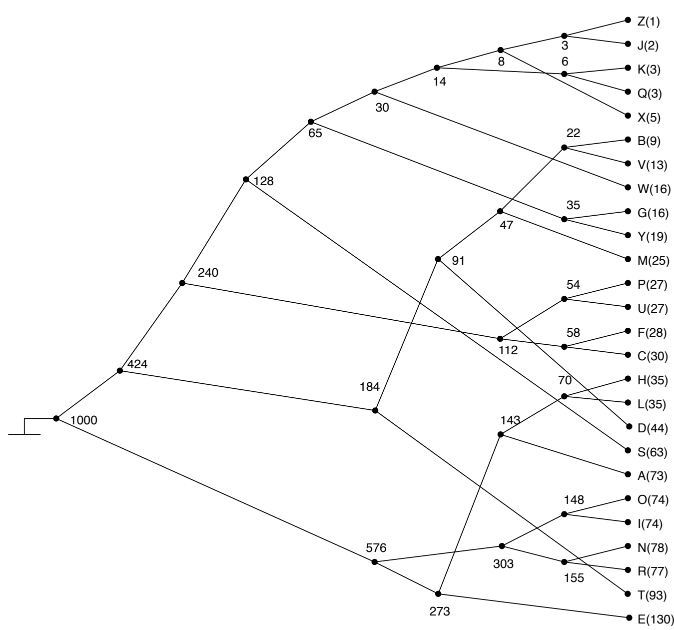


The code table becomes

x	y_H	x	y_H
000	00	100	110
001	100	101	011
010	101	110	1110
011	010	111	1111

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

4.8. 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).



x	y
A	1101
B	010000
C	00111
D	0101
E	111

x	y
F	00110
G	000010
H	11000
I	1001
J	000000001

x	y
K	00000010
L	11001
M	01001
N	1010
O	1000

x	y
P	00100
Q	00000011
R	1011
S	0001
T	011

x	y
U	00101
V	010001
W	000001
X	00000001
Y	000011
Z	000000000

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

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

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

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

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.

4.9. (a) For the binary case an optimal code is given by a binary tree of depth 1, i.e. the root and two leaves, which gives the 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.

4.10. (a) For $P(n)$ to be a probability function it must be positive and sum to 1. Here, it is clear that $P(n) \geq 0$ for all n, and since $1/k < 1$ the sum becomes

$$\sum_{n=1}^{\infty}(k-1)k^{-1} = (k-1)\sum_{n=1}^{\infty}\left(\tfrac{1}{k}\right)n = (k-1)\frac{\frac{1}{k}}{1-\frac{1}{k}} = \frac{k-1}{k-1} = 1$$

Hence, it is a probability function.

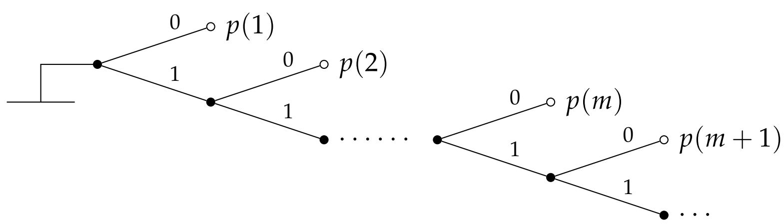
(b) With $k = 2$ we get $P(n) = \left(\frac{1}{2}\right)^n$. By considering the optimal codeword lengths

$$l_n^{(opt)} = -\log P(n) = -\log\left(\tfrac{1}{2}\right)^n = n$$

we see that this is an integer for each number n. It is also the same as the codeword lengths for the unary code, and we conclude that it is optimal for this case. The entropy is in that case equal to the average codeword length

$$H(X) = L = \sum_{n=1}^{\infty} n\left(\tfrac{1}{2}\right)^n = \tfrac{1}{2}\sum_{n=1}^{\infty} n\left(\tfrac{1}{2}\right)^{n-1} = \frac{\frac{1}{2}}{\left(1-\frac{1}{2}\right)^2} = 2$$

(c) For a general k the optimal codeword length $l_n^{(opt)} = -\log P(n)$ is typically not integers and can therefore not be used to construct an optimal code. It also means that the average length of an optimal code will not equal the entropy. Our next attempt is then to show that the code satisfies Huffman's algorithm, which will produce an optimal code. Then write the code in a tree,



Consider then the sub-tree stemming from level m (the tree containing the leaves $p(m+1)$, $p(m+2)$, etc. The root node of this tree has the probability

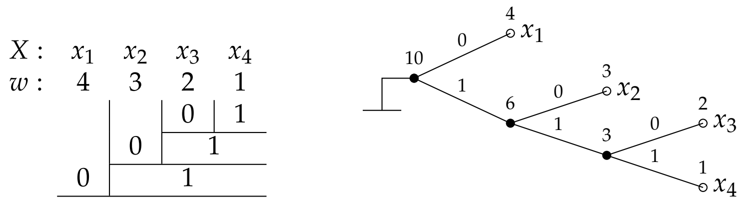
$$\begin{aligned} r(m) &= \sum_{n=m+1}^{\infty}(k-1)\left(\tfrac{1}{k}\right)^n = (k-1)\left(\tfrac{1}{k}\right)^{m+1}\sum_{l=0}^{\infty}\left(\tfrac{1}{k}\right)^l \\ &= (k-1)\left(\tfrac{1}{k}\right)^{m+1}\frac{1}{1-\frac{1}{k}} = \left(\tfrac{1}{k}\right)^m \leq (k-1)\left(\tfrac{1}{k}\right)^m = p(m) \end{aligned}$$

Hence, among the nodes $p(1), p(2), \ldots, p(m)$ and $r(m)$, the two least probable are $p(m)$ and $r(m)$. Merging those two nodes in a tree will give one step further up in the tree. After $m-2$ more similar merges, according to the Huffman algorithm, the unary code has been constructed. Hence, for $p(n)$ as in the problem, the unary code is a Huffman code and, hence, it is optimal. The corresponding codeword length given by

$$L = \sum_{n=1}^{\infty} n(k-1)\left(\tfrac{1}{k}\right)^n = (k-1)\left(\tfrac{1}{k}\right)\sum_{n=1}^{\infty} n\left(\tfrac{1}{k}\right)^{n-1} = (k-1)\frac{\frac{1}{k}}{\left(1-\frac{1}{k}\right)^2} = \frac{k}{k-1}$$

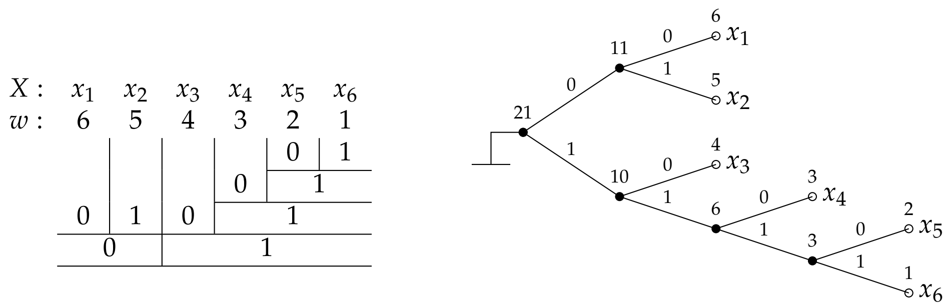
4.11. For simplicity, the common denominator in the probabilities, for each sub-problem, is dropped and the numerator is used as weight in the algorithm.

(a) In the first example the weights for the outcomes are $w(x_1) = 4$, $w(x_2) = 3$, $w(x_3) = 2$ and $w(x_4) = 1$. The first split separates $\{x_1\}$ in one part and $\{x_2, x_3, x_4\}$ in the other. The first set is marked with 0 and the second with 1. The second set is split again into $\{x_2\}$ and $\{x_3, x_4\}$. Finally the last part is split into $\{x_3\}$ and $\{x_4\}$. Since al sets now contain only one outcome each there is no more splitting. By marking the subsets in each split by 0 and 1, a code is obtained. Below, to the left, the propcedure is shown. To the right the corresponding code tree is shown.



Since the merging of the leafs in the tree follows the Huffman procedure it is a Huffman code, and hence optimal.

(b) In the second example the weights are $w(x_1, x_2, x_3, x_4, x_5, x_6) = (6, 5, 4, 3, 2, 1)$. Following the same procedure as in (a), we get

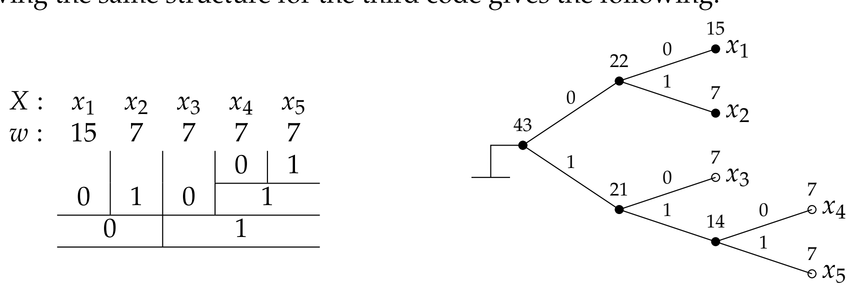


When constructing a Huffman code, first three leaves x_5 and x_6 are merged, then $\{x_5 x_6\}$ and x_4 are merged. After this the nodes in the algorithm are $(x_1, x_2, x_3, \{x_4 x_5 x_6\})$ with weights $(6, 5, 4, 6)$. So in the next step in the Huffman procedure the nodes x_2 and x_3 are merged. this is not the case in the tree above, and hence the code is not a Huffman code. Continuing the Huffman procedure results in the code tabulated below.

X	w	Y
x_1	6	10
x_2	5	01
x_3	4	00
x_4	3	110
x_5	2	1110
x_6	1	1111

The average codeword length for the Huffman code is 51/21, and according to the path length lemma the codeword length for the Fano code is $L_F = \frac{21+11+10+6+3}{21} = \frac{51}{21}$. Hence the code is optimal.

(c) Following the same structure for the third code gives the following.



The average codeword length is $L_F = \frac{43+22+21+14}{43} = \frac{100}{43}$. When constructing a Huffman code the nodes x_4 and x_5 are merged in the first step. In the second step x_2 and x_3 are merged, which is not the case in the tree for the Fano code. Hence the obtained code is not a Huffman code. In the following table a Huffman code is shown.

X	w	Y
x_1	15	0
x_2	7	100
x_3	7	101
x_4	7	110
x_5	7	111

The average codeword length is $L_H = \frac{99}{43}$. Hence, the Fano code is neither a Huffman code nor optimal.

4.12. In the following table the intervals are shown. The expanded sub-intervals are shown in bald font.

x:	d	a	c	b	d	a	a	a	d	a
d	**1.0**	1.00	0.950	0.9450	**0.94400**	0.94400	0.943925	0.943887	**0.9438688**	0.94386875
c	**0.9**	0.99	**0.945**	0.9445	**0.94385**	0.94398	0.943917	0.943884	**0.9438669**	0.94386856
b	0.8	0.98	**0.940**	**0.9440**	0.94370	0.94397	0.943910	0.943880	0.9438650	0.94386837
a	0.5	**0.95**	0.925	**0.9425**	0.94325	**0.94392**	**0.943887**	**0.943869**	0.9438594	**0.94386781**
	0.0	**0.90**	0.900	0.9400	0.94250	**0.94385**	**0.943850**	**0.943850**	0.9438500	**0.94386687**

The codeword length is $\ell_{\boldsymbol{x}} = \lceil -\log(0.94386781 - 0.94386687) \rceil + 1 = 22$. The mid value of the last interval is $\zeta = 0.94386734$, where the decimal part of the binary form gives the codeword as

$$y = 1111000110100001010010$$

Note: As a comparison, without any compression the code symbols can be encoded two with bits each. That means the algorithm in this case gives an expansion instead of compression. If the sequence would be longer the frequency of the letters would correspond better with the distribution, and the result would approach the entropy $H(X) = 1.6855$ bit.

4.13. The decimal number that points out the last sub-interval is

$$\widehat{\zeta} = \{0.01111000111011100\}_2 = 0.47238159$$

The procedure for decoding is the same as for encoding, for each level expand the sub-interval containing the number $\widehat{\zeta}$. The expanded sub-interval decides the symbol in the decoded vector. In this case the problem does not state the length of the x vector, so the stoping criteria for the expansion is when the codeword length for the sub-interval equals the codeword length of the given codeword. Then the following intervals are obtained, together with the derived codeword length and the decoded vector,

d	1.0	**0.50**	0.500	0.4750	**0.47250**	0.47250	0.472450	0.472413
c	0.9	**0.45**	0.495	**0.4725**	**0.47225**	0.47247	0.472442	0.472409
b	0.8	0.40	0.490	**0.4700**	0.47200	**0.47245**	0.472435	0.472405
a	**0.5**	0.25	**0.475**	0.4625	0.47125	**0.47237**	**0.472413**	**0.472394**
	0.0	0.00	**0.450**	0.4500	0.47000	0.47225	**0.472375**	**0.472375**
$\widehat{\ell}$:	2	6	7	10	13	15	16	17
$\widehat{x}$:	a	d	a	c	d	b	a	a

The decoded vector is $\widehat{x} = adacdbaa$.

4.14. As noted in the problem, this should solved by programming. To implement the algorithm it can aslo be noted that after updating the intervals the old interval will not be used anymore. That means only one occasion of the interval vector is needed. One such example is given below as a MATLAB implementation:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Arithmetic coding for long binary vector
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
p=1/4; % P(X=0)
N=10000; % Length of vector
%%%%%%%%%%%%%% Create vectors
x=[rand(1,N)>p]; % vector of N binary symbols (as logical)
y=''; % Code symbols
F=[0 p 1]; % Cumulative probabilities
Int=F; % Current intervals
%%%%%%%%%%%%%% Run algorithm
for t=1:N, % position in x
  if x(t), % x=1
    Int=Int(2)+(Int(3)-Int(2))*F;
  else % x=0
    Int=Int(1)+(Int(2)-Int(1))*F;
  end
  Scale=(1==1); % True
  while Scale, % Scale the intervals
    if Int(1)>=0.5, % Interval above 0.5
      Int=2*Int-1;
      y=[y '1'];
    elseif Int(3)<0.5, % Interval below 0.5
      Int=2*Int;
      y=[y '0'];
    else
      Scale=(1==0); % False
    end
  end
end
EndEll=ceil(-log2(Int(3)-Int(1)))+1;
xi=(Int(3)+Int(1))/2;
y=[y cast(float2bin(xi,0,EndEll)+'0','char')];
```

In the program the function `float2bin` is used. This, and its inverse function, are found at MathWorks, `https://se.mathworks.com/matlabcentral/answers/25549`, as:

```
function Abin = float2bin(Afloat,NumInt,NumFrac)
Abin=fix(rem(Afloat*pow2(-(NumInt-1):NumFrac),2));
```

```
function Afloat = bin2float(Abin,NumInt,NumFrac)
Afloat=Abin*pow2([NumInt-1:-1:0 -(1:NumFrac)].');
```

The program is run 10 000 times to get the codeword lengths in plotted in Figure 22. The average codeword length per symbol for the plotted values is $L = 0.8114$ codeword bit per input bit. This should be compared with the entropy for the source, $H(X) = h(\frac{1}{4}) = 0.8113$ bit, and the upper bound on the average codeword length $H(X) + \frac{2}{10\,000} = 0.8115$.

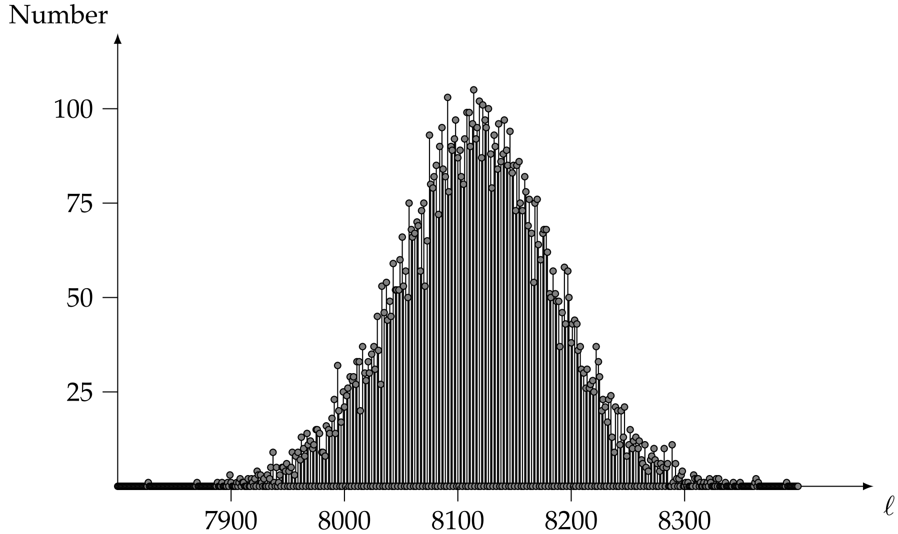


Figure 22: The resulting codeword lengths from 10 000 runs of arithmetic coding of binary vectors of length 10 000 bits.

Chapter 5

5.1. (a) The estimated distribution and a corresponding Huffman code is given by the following table.

x	$q(x)$	y
a	$5/10$	0
b	$1/10$	100
c	$1/10$	101
d	$3/10$	11

Then the encoded sequence is

$$\boldsymbol{y} = 110101100110001 10$$

which gives the average codeword length per source symbol $L_x = 1.7$ bit.

(b) The given distribution together with a Huffman code is given in the following table.

x	$p(x)$	y
a	0.5	0
b	0.3	10
c	0.1	110
d	0.1	111

Then the encoded sequence is

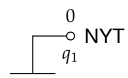
$$\boldsymbol{y} = 1110110101110001110$$

In this case the average codeword length per source symbol is $L_x = 1.9$ bit.

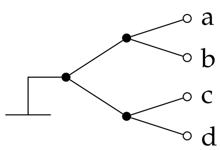
Note: The source sequence is generated according to the distribution in the problem, but it is too short to give an accurate estimation of the distribution. If the length of the sequence would be longer the codeword length for both (a) and (b) would approach the entropy $H(X) = 1.6855$, as the estimated distribution approaches the true distribution. The codeword lengths can be compared with the case when no compression is done, i.e. 2 bits per symbol. Both versions in this problem give a slight compression, while the result in Problem 4.12 does not. Again, for a longer sequence this would also approach the entropy.

5.2. Assume the alphabet is $\{a, b, c, d\}$. From the beginning there are no letters in the tree, just the NYT-node. That is,

NYT
q_1
0



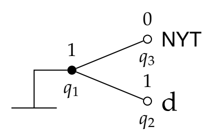
The code for the unused symbols in the NYT-node has to be determined according to some rule agreed by the transmitter and sender. In this case we use alphabetically ordered from top and down ina binary tree that is as full as possible. So when all letters are theer it is



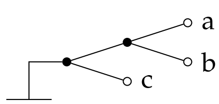
Adding symbol by symbol from the sequence gives:

'd': Expand the NYT-node and add a leaf for d

	d	NYT
q_1	q_2	q_3
1	1	0

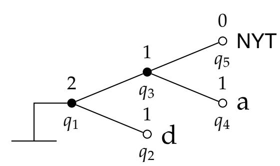


The new NYT-tree is

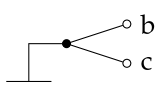


'a': Expand the NYT-node and add a leaf for a. Since $w_2 < w_1$ this can be done without rearangements.

	d		a	NYT
q_1	q_2	q_3	q_4	q_5
2	1	1	1	0

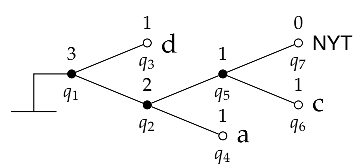


The new NYT-tree is

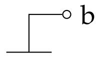


'c': Again it is a new symbnol that should be insterted into the tree. This time the path from the NYT-node to the root passes q_3 where $w_3 \not< w_2$, and the tree needs to be rearranged. Here it means swapping the sub-trees q_3 and q_2. Then, the NYT-node can be expanded and a leaf for c added.

		d	a		c	NYT
q_1	q_2	q_3	q_4	q_5	q_6	q_7
3	2	1	1	1	1	0

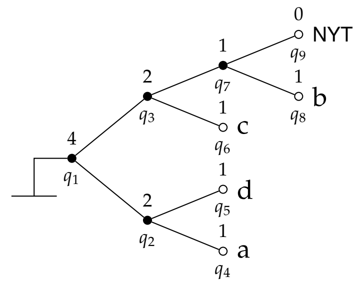


The new NYT-tree is



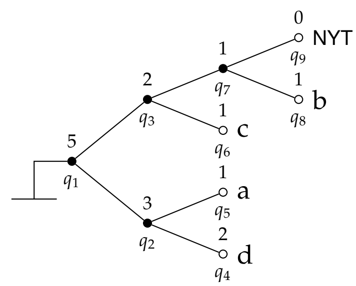
'b': The last symbol to add from the NYT-node. The path back to the root goes through $q_7 \to q_5 \to q_2 \to q_1$. Here $w_5 \not< w_4$ and the tree needs to restructure. By swapping q_3 and q_5 the condition is fulfilled, and the NYT-node can be exapnded with the leaf b.

			a	d	c		b	NYT
q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9
4	2	2	1	1	1	1	1	0



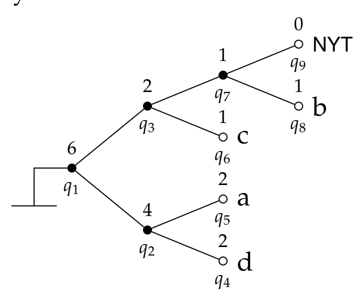
'd': From now on all letters in the alphabet has been added and the NYT-node is empty (and could have been removed). So adding a letter d means incrementing the leaf weight. The path to the root is $q_5 \to q_2 \to q_1$, where $w_5 \not< w_4$ this can be rearanged by swapping q_4 and q_5, and then the weights for the path incremented.

			d	a	c		b	NYT
q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9
5	3	2	2	1	1	1	1	0



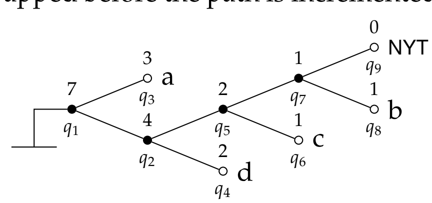
'a': The path from a back to the root is $q_5 \to q_2 \to q_1$. For this path all the nodes fulfil $w_i < w_{i-1}$, and it can be incremeted with preserved sibling property.

			d	a	c		b	NYT
q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9
6	4	2	2	2	1	1	1	0



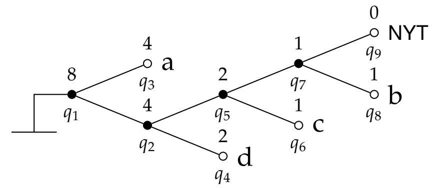
'a': One more a means the sub-trees q_5 and q_3 should be swapped before the path is incremented.

		a	d		c		b	NYT
q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9
7	4	3	2	2	1	1	1	0



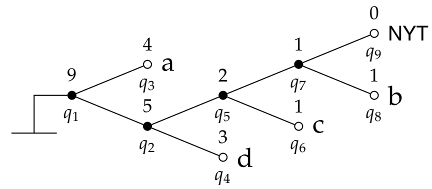
'a': Yet another a means incrementing the path $q3 \to q_1$, which can be done without rearrangements.

		a	d		c		b	NYT
q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9
8	4	4	2	2	1	1	1	0



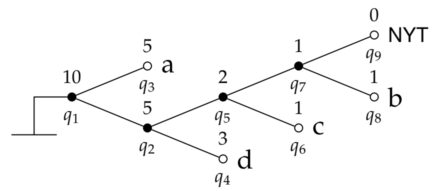
'd': The path from d to the root is $q_4 \to q_2 \to q_1$, which can be incremented directly.

		a	d		c		b	NYT
q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9
9	5	4	3	2	1	1	1	0



'a': The final letter a gives the path $q_3 \to q_1$ which can also be incremented directly.

		a	d		c		b	NYT
q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9
10	5	5	3	2	1	1	1	0



The encoding is done using the tre developed by the previous symbol. If a new letter is encountered, first the path to the NYT- node is given, followed by the path in the NYT-tree. That gives

$\boldsymbol{x} =$ d a c b d a a a d a

$\boldsymbol{y} =$ 11 000 001 000 10 10 0 0 11 0

5.3. The encoding 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 ratio is $R = \frac{360}{266} = 1.3534$.

5.4. Encoding according to

S buffer	B buffer	Codeword
`'I scream, you sc'`	`'ream, we'`	(12,6,'w')
`'m, you scream, w'`	`'e all sc'`	(6,1,' ')
`' you scream, we '`	`'all scre'`	(7,1,'l')
`'ou scream, we al'`	`'l scream'`	(1,1,' ')
`' scream, we all '`	`'scream f'`	(15,6,' ')
`', we all scream '`	`'for ice '`	(0,0,'f')
`' we all scream f'`	`'or ice c'`	(0,0,'o')
`'we all scream fo'`	`'r ice cr'`	(7,1,' ')
`' all scream for '`	`'ice crea'`	(0,0,'i')
`'all scream for i'`	`'ce cream'`	(11,1,'e')
`'l scream for ice'`	`' cream.'`	(4,1,'c')
`'scream for ice c'`	`'ream.'`	(14,4,'.')

There are 11 codewords and an initialisation vector of 16 letters, giving $11(5+4+8)+16\cdot 8=315$ bits. (The codeword length can also be argued to be $4+3+8=15$ bits, but according to the book it should be $\lceil\log(16+1)\rceil+\lceil\log(8+1)\rceil+8=5+4+8=17$). The uncoded length is $49\cdot 8=392$ bits. Then the compression ratio is $R=392/315=1.24$.

5.5. 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.*

5.6. The encoding 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 $R = \frac{360}{281} = 1.28$.

5.7. (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, Compression ratio: 1.1282

(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, Compression ratio: 1.326

(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, Compression ratio: 1.2222

(d)

Index	Codeword	Dictionary	Binary
32		`[ ]`	Init with ASCII table
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, Compression ratio: 1.2816

5.8.

5.9.

step	lexicon	prefix	new symbol	codeword	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.

5.10. Decoding:

Ind	Codeword	Text	Dictionary
256	102	[f]	[fr]
257	114	[r]	[re]
258	101	[e]	[es]
259	115	[s]	[sh]
260	104	[h]	[hl]
261	108	[l]	[ly]
262	121	[y]	[y]
263	32	[ ]	[f]
264	256	[fr]	[fri]
265	105	[i]	[ie]
266	101	[e]	[ed]
267	100	[d]	[d]
268	263	[f]	[fr]
269	257	[re]	[res]
270	259	[sh]	[sh]
271	263	[f]	[fl]
272	108	[l]	[le]
273	258	[es]	[esh]
274	104	[h]	

So the text is: freshly fried fresh flesh

Chapter 6

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

(a) Since nq might not be an integer it can be rounded as $[nq]$. Then there are $\binom{n}{[nq]}$ sequences with $[nq]$ 1s.

(b) To represent the sequences in (a) it is needed $\lfloor \log \binom{n}{[nq]} \rfloor$ bits (the lower integer limit $\lfloor \rfloor$ is used to achieve an integer number). Hence, in total it is needed $\frac{1}{n} \lfloor \log \binom{n}{[nq]} \rfloor$ 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. It is also used that $\frac{1}{n}\log n \to 0,\ n \to \infty$. Notice, that this imply that for large n we get $\binom{n}{k} = \binom{n}{\frac{k}{n}n} \approx 2^{nh(\frac{k}{n})}$

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

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

and

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

Dividing these and using the chain rule concludes the proof.

6.3. 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

$$P(00\dots0) = \Big(\frac{49}{50}\Big)^{100} \approx 0.1326$$

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

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

or, equivalently,

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

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,

$$-\frac{1}{100}\log P(00\dots0)) = -\frac{1}{100}\log\Big(\frac{49}{50}\Big)^{100} = -\log 49 + \log 50 = 1 + 2\log 5 - 2\log 7$$

Thus,

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

and 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

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

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 $2^{100} \approx 1.2677 \cdot 10^{30}$.

6.4. Consider a sequence of n cuts and let $\boldsymbol{x} = x_1 x_2 \dots x_n$ be the the outcome where x_i is the part saved in cut i. If in k cuts the long part is saved and in $n-k$ cuts the short part is saved, 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, the most probable event is the typical, represented by the set $A_\varepsilon(\boldsymbol{X})$. If a typical sequence is considered, the probability is bounded by

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

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 gives

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

Hence, in average, $\frac{2^{3/4}}{3}$ of the length is kept at each cut.

An alternative solution is to notice that, according to the law of large numbers, for large n the short part is saved in $n\frac{1}{4}$ splits and the long part saved in $n\frac{3}{4}$ splits. Then the remaining length is

$$L = \Big(\frac{2}{3}\Big)^{n\frac{3}{4}}\Big(\frac{1}{3}\Big)^{n\frac{1}{4}} = \frac{2^{n\frac{3}{4}}}{3^n} = \Big(\frac{2^{\frac{3}{4}}}{3}\Big)^n$$

6.5. The channels are either symmetric or weakly symmetric, so

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

6.6. 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 $H(\frac{1}{2},\frac{1}{2}) = 1$.

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

6.7. 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

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

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

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

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

$$C=\max_{p(x)} I(X;Y)=\max_{p(x)} H(Y)-\log 3=\log 15-\log 3=\log\frac{15}{3}=\log 5$$

This is achieved for uniform Y which by symmetry is achieved for uniform X, i.e. $p(x_i)=\frac{1}{15}$.

Alternatively the problem can be solved by noting that the channel is a strongly symmetric DMC with 15 symbols and transmission probabilities $\boldsymbol{p}=(\frac{1}{3},\frac{1}{3},\frac{1}{3})$. Hence,

$$C=\log 15-H(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3})=\log 15-\log 3=\log 5$$

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

$$P(Z=0|X=0)=(1-p)^2+p^2$$
$$P(Z=1|X=0)=p(1-p)+(1-p)p=2p(1-p)$$
$$P(Z=0|X=1)=2p(1-p)$$
$$P(Z=1|X=1)=(1-p)^2+p^2$$

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

6.10. (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.

6.11. 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 6.6 with $p = \alpha(1-\beta)$ and $q = \beta$, the capacity follows from the solution.

6.12. (a)

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

where in the third equality the terms $H(Z) - H(Z)$ are added, and it is noted that $H(Z|Y,X) = H(Z|X)$ since the two channels work independently.

(b) Since X is binary with equal probabilities we get directly $I(X;Y) = I(X;Z) = 1 - h(p)$. It also gives that $p(y) = p(z) = 1/2$, and, hence, $I(Y;Z) = H(Y) + H(Z) - H(Y,Z) = 2 - H(Y,Z)$. Then, to get the first part of the problem,

$$
\begin{aligned}
I(X;Y,Z) &= I(X;Y) + I(X;Z) - I(Y;Z) \\
&= 2\big(1-h(p)\big) - \big(2 - H(Y,Z)\big) = H(Y,Z) - 2h(p)
\end{aligned}
$$

To get the distribution for (Y,Z) we follow the hint in the problem and derive $p(y,z|x) = p(y|x)p(z|x)$, which follows from that conditioned on X, Y and Z are independent. Since $p(x) = 1/2$ the unconditional probability is $p(y,z) = \frac{1}{2}\big(p(y,z|x=0) + p(y,z|x=1)\big)$. The probability functions are listed in the following table

X	Y	Z	$p(y,z\|x)$
0	0	0	$(1-p)^2$
0	0	1	$p(1-p)$
0	1	0	$p(1-p)$
0	1	1	p^2
1	0	0	p^2
1	0	1	$p(1-p)$
1	1	0	$p(1-p)$
1	1	1	$(1-p)^2$

Y	Z	$p(y,z)$
0	0	$\frac{1}{2}(p^2+(1-p)^2)$
0	1	$p(1-p)$
1	0	$p(1-p)$
1	1	$\frac{1}{2}(p^2+(1-p)^2)$

Then,

$$
\begin{aligned}
H(Y,Z) &= H\Big(\tfrac{1}{2}\big((1-p)^2+p^2\big), \tfrac{1}{2}\big((1-p)^2+p^2\big), p(1-p), p(1-p)\Big) \\
&= -\big((1-p)^2+p^2\big)\log\frac{(1-p)^2+p^2}{2} - 2(1-p)\log p(1-p)
\end{aligned}
$$

Inserting in the above expression gives

$$
\begin{aligned}
I(X;Y,Z) &= H(Y,Z) - 2h(p) \\
&= p^2 \log \frac{2}{(1-p)^2+p^2} + (1-p)^2 \log \frac{2}{(1-p)^2+p^2} \\
&\quad - 2p(1-p)\log p - 2p(1-p)\log(1-p) + 2p\log p + 2(1-p)\log(1-p) \\
&= p^2 \log \frac{2}{(1-p)^2+p^2} + (1-p)^2 \log \frac{2}{(1-p)^2+p^2} + p^2\log p + (1-p)^2\log(1-p) \\
&\overset{(a)}{=} p^2 \log \frac{2p^2}{(1-p)^2+p^2} + (1-p)^2 \log \frac{2(1-p)^2}{(1-p)^2+p^2} \\
&= \big(p^2+(1-p)^2\big) + p^2 \log \frac{p^2}{(1-p)^2+p^2} + (1-p)^2 \log \frac{(1-p)^2}{(1-p)^2+p^2} \\
&= \big(p^2+(1-p)^2\big)\Big(1 + \frac{p^2}{p^2+(1-p)^2}\log\frac{p^2}{(1-p)^2+p^2} + \frac{(1-p)^2}{p^2+(1-p)^2}\log\frac{(1-p)^2}{(1-p)^2+p^2}\Big) \\
&\overset{(b)}{=} \big(p^2+(1-p)^2\big)\Big(1 - h\Big(\frac{p^2}{(1-p)^2+p^2}\Big)\Big)
\end{aligned}
$$

where (a) and (b) are the results to be shown.

The formula in (b) can be interpreted as follows. Viewed from the receiver $(Y,Z)=(0,1)$ or $(Y,Z)=(1,0)$, which happens with probability $2p(1-p)$, the probability for $X=0$ and $X=1$ are both $1/2$, so there is no information in this event. On the other hand, with probability $p^2+(1-p)^2$ the receiver gets $(0,0)$ or $(1,1)$, which gives the information $1-h\big(\frac{p^2}{(1-p)^2+p^2}\big)$. Here, $\frac{p^2}{(1-p)^2+p^2}$ is $P(Y \neq X, Z \neq X | Y = Z)$, that is, the probability that both Y and Z are wrong if the receiver gets the same result from the two channels.

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

$P(X|Y)$

X	$Y=0$	$Y=1$
0	p	0
1	$(1-p)\alpha$	$((1-p)(1-\alpha))$

and

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

The conditional and unconditional entropies of Y are then given by

$$
\begin{aligned}
H(Y|X) &= H(Y|X=0)p + H(Y|X=1)(1-p) = (1-p)h(\alpha) \\
H(Y) &= h\big((1-p)(1-\alpha)\big)
\end{aligned}
$$

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

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

which gives

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

Inserting to the mutual information gives

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

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 in Figure 23.

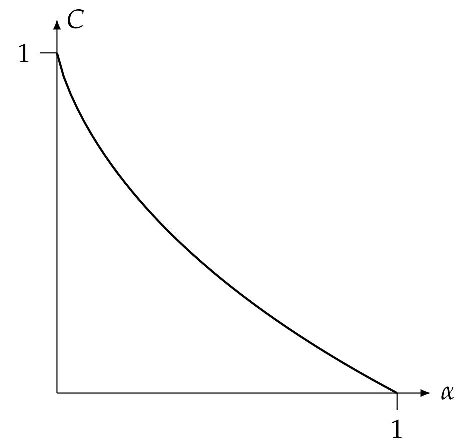


Figure 23: Capacity for the Z channel.

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

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

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

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

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

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

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

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

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

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

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

$P(A, B)$

X \ Y	0	1	2
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

Thus, $d_{\min} = \min_{\boldsymbol{x} \neq \boldsymbol{0}} \{w_H(\boldsymbol{x})\} = 3$

(d) From part (b) it is noted that $G = (I\ P)$. Then, from

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

it is concluded

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

(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 syndrome (111).

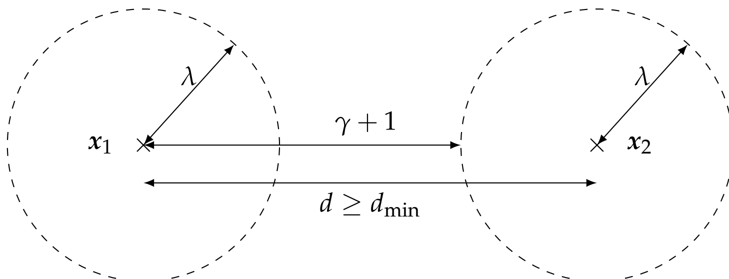
(f) One (correctable) error

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

An uncorrectable error

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

7.2. Consider the graphical interpretation of $\mathbb{F}_2^n$ and the two codewords x_i and 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, it is equivalent to show that the resulting vector has even weight. For this the position-wise AND function is used to get the positions in which both codewords have ones. Then, if $y_1, y_2 \in \mathcal{B}$ the weight of their sum can be written as

$$w_H(y_1+y_2) = w_H(y_1) + w_H(y_2) - 2w_H(y_1 \& y_2)$$

here it should be noted 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 the resulting codeword is also 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 $y = (y_1 \dots y_{n+1})$ is a codeword iff $yH_E^T = \mathbf{0}$. This gives

$$\begin{aligned} yH_E^T &= (y_1 \dots y_n y_{n+1}) \begin{pmatrix} & & & 1 \\ & H^T & & \vdots \\ & & & 1 \\ 0 & \dots & 0 & 1 \end{pmatrix} \\ &= \left((y_1 \dots y_n)H^T \quad \textstyle\sum_{i=1}^{n+1} y_i\right) = \mathbf{0} \end{aligned}$$

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

(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 $d_E = d + 1$.

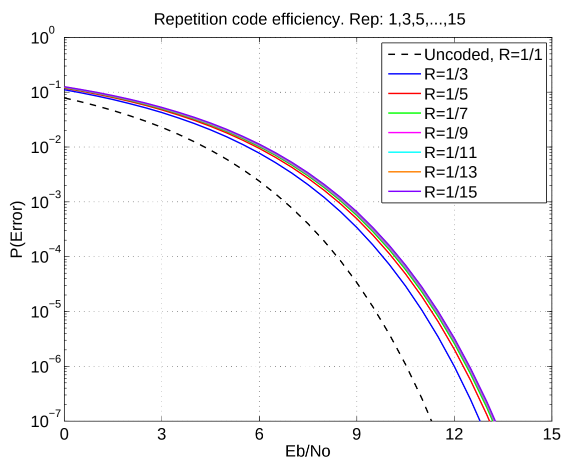
7.4. $H = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$

7.5. 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 $P_{b,N} = Q(\sqrt{2E_b/N_0 N})$. 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

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

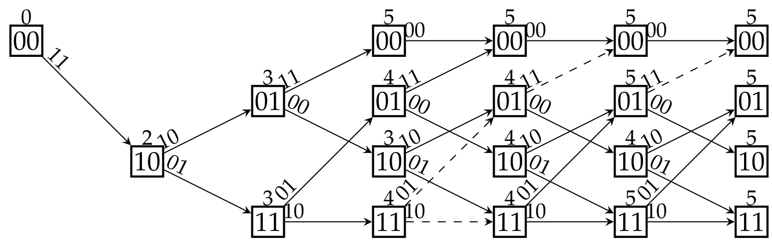
In MATLAB the Q-function as qfunc(x).

A plot of the results for $N = 3, 5, \dots, 15$ 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 introductions. 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 is used as one pulse, the result will be roughly the same for all cases. This can be employed by using a soft decoding algorithm instead of hard (bit by bit) decision.

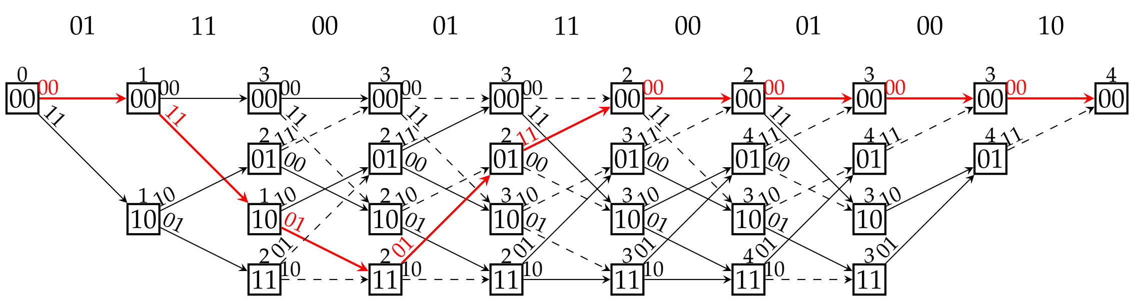
7.6. The free distance is the minimum weight of a non-zero path, starting and ending in the all-zero state. A simple and brute force method to find it is to start in the zero state and give a non-zero input. After this, follow all possible paths, counting the Hamming weight of the paths, until the zero 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.

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.



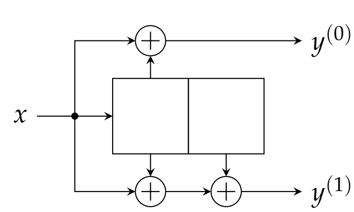
In the figure the red path is found by following the surviving branches from the end node to the start node. This corresponds to the minimum distance path, or the maximum likelihood path. In this case it is

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

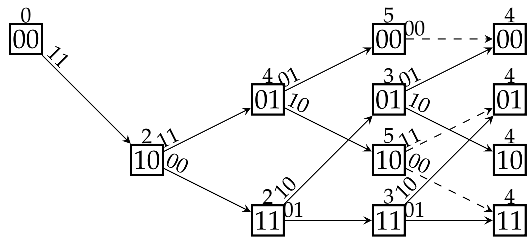
So the answer is that the most likely transmitted information sequence is $\hat{\boldsymbol{u}}_1 = 0110000$.

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.

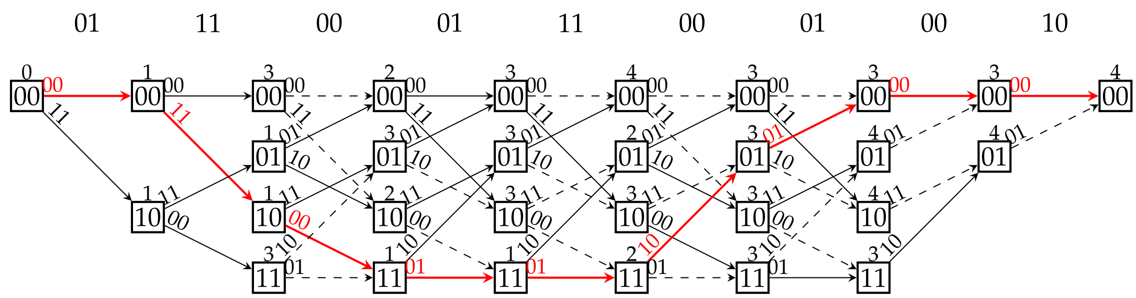
7.7. The encoder circuit for this generator matrix is



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



Decoding is done as follows.

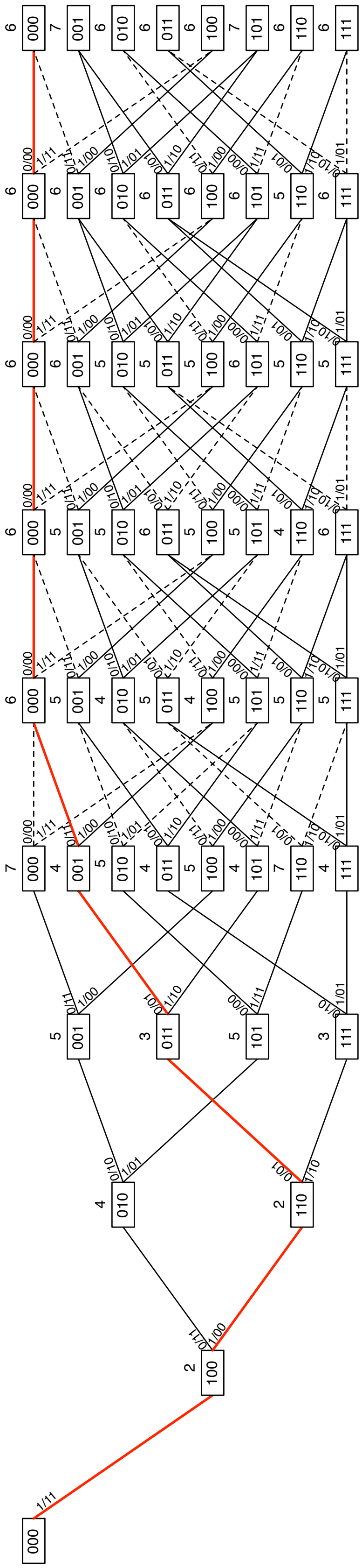


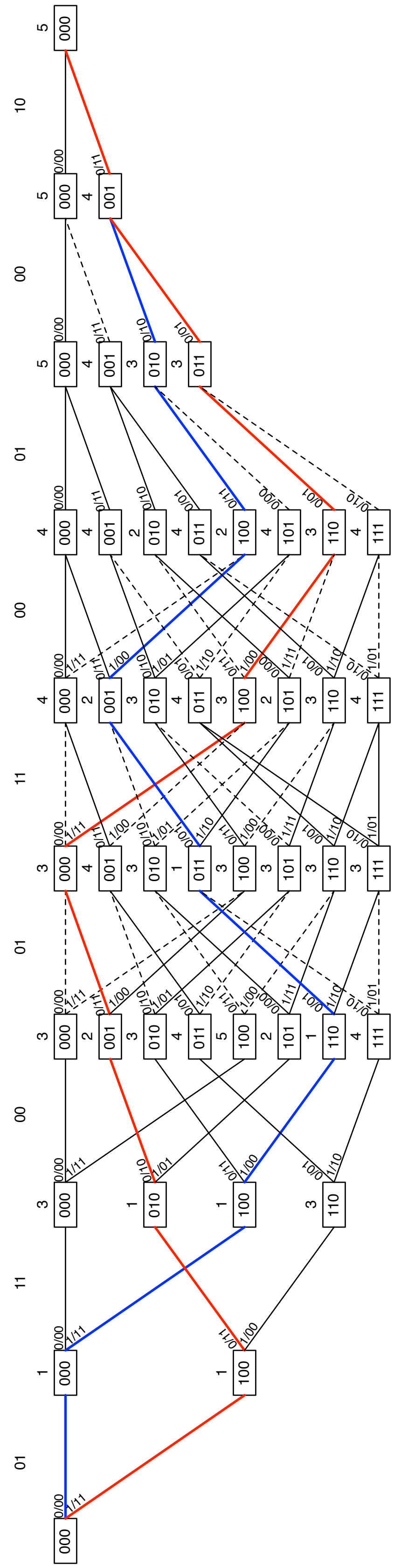
Hence, the most likely code sequence is $\hat{\boldsymbol{v}} = 00\ 11\ 00\ 01\ 01\ 10\ 01\ 00\ 00$ and the corresponding information sequence is $\hat{\boldsymbol{u}} = 0111100$.

7.8. (a) According to the left trellis on the next page, $d_{\text{free}} = 6$. It corresponds to the information path $110000\ldots$.

(b) Decoding is done with the right trellis on the next page. There are two paths through the trellis giving the minimum distance,

$$\hat{u}_1 = 100011\ (000)$$
$$\hat{u}_2 = 011001\ (000)$$





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 ($H(X) = -\int f(x)\log f(x)\,dx$),

(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)

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

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

8.2. (a) Use that $\int f(x,y)dxdy = 1$ to get

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

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

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

(c) 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

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

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

8.3. (a) To get α,

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

(b) The probability is

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

(c) 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

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

(d) 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,

$$\begin{aligned}
P(X < a) &= P\big(e^Y < a\big) = P\big(Y < \ln a\big) \\
&= \int_{-\infty}^{\ln a} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy = \begin{bmatrix} x = e^y \;\Rightarrow\; y = \ln x \\ dy = \frac{1}{x} dx \end{bmatrix} \\
&= \int_0^a \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} dx
\end{aligned}$$

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

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

(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. Since X and Y are independent,

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

where it is first used that when X is known it does not affect the entropy, and second that X and Y are independent.

(a) Since the sum of two Gaussian random variables is again Gaussian, $Y \sim \mathrm{N}(0, \sqrt{2})$,

$$I(X;Z) = H(Z) - H(Y) = \frac{1}{2}\log 2\pi e 2 - \frac{1}{2}\log 2\pi e = \frac{1}{2}\log 2 = \frac{1}{2}$$

(b) Since Y is uniform over an interval of length 1, the entropy is $H(Y) = \log 1 = 0$. The distribution for Z can be derived as

$$f_Z(z) = f_X * f_Y(z) = \begin{cases} 1+z, & -1 \le z \le 0 \\ 1-z, & 0 \le z \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Then,

$$H(Z) = -\int_{-1}^{0}(1+z)\log(1+z)dz - \int_{0}^{1}(1-z)\log(1-z)dz$$

These integrals can be solved as

$$-\frac{1}{\ln 2}\int_{-1}^{0}(1+z)\ln(1+z)dz = -\frac{1}{\ln 2}\int_{0}^{1} t\ln t\, dt = -\frac{1}{\ln 2}\Big[\frac{t^2}{2}\ln t - \frac{t^2}{4}\Big]_0^1 = \frac{1}{4\ln 2}$$
$$-\frac{1}{\ln 2}\int_{0}^{1}(1-z)\ln(1-z)dz = -\frac{1}{\ln 2}\int_{0}^{1} t\ln t\, dt = \frac{1}{4\ln 2}$$

where the variable changes $t = 1+z$ and $t = 1-z$ have been used in the first and second equations, respectively. Also, the limit value $\lim_{t\to 0} t^2 \ln t = 0$ was used. That means the mutual information can be derived as

$$I(X;Z) = H(Z) - H(Y) = \frac{1}{2\ln 2} = \frac{1}{2}\log e$$

8.6. It is shown in Problem 2.8 that

$$E\big[(X-\alpha)^2\big] \ge E\big[(X-E[X])^2\big] = \sigma^2$$

On the other hand, the differential entropy is maximised for the Gaussian distribution,

$$H(X) \le \frac{1}{2}\log 2\pi e\sigma^2$$

Taking both sides as exponents to 2 gives

$$2^{2H(X)} \le 2\pi e\sigma^2$$

or, equivalently,

$$\sigma^2 \ge \frac{1}{2\pi e}2^{2H(X)}$$

Combining the above gives the result.

8.7. For both cases

$$I(X_k;Y) = H(Y) - H(Y|X_k) = H(Y) - H\big(\sum_i X_i|X_k\big)$$
$$= H(Y) - H\big(\sum_{i\neq k} X_i|X_k\big) = H(Y) - H\big(\sum_{i\neq k} X_i|\big) = H(Y) - H(Z)$$

where $Z = \sum_{i\neq k} X_i$

(a) With $X_i \sim \mathrm{N}(0,1)$, the sums become $Y \sim \mathrm{N}\big(0,\sqrt{n}\big)$ and $Z \sim \mathrm{N}\big(0,\sqrt{n-1}\big)$ which gives

$$I(X_k;Y) = H(Y) - H(Z) = \frac{1}{2}\log 2\pi e n - \frac{1}{2}\log 2\pi e(n-1) = \frac{1}{2}\log\frac{n}{n-1}$$

(b) With $X_i \sim \mathrm{N}(m_i,\sigma_i)$,

$$Y \sim \mathrm{N}\Big(\sum_i m_i, \sqrt{\sum_i \sigma_i^2}\Big)$$
$$Z \sim \mathrm{N}\Big(\sum_{i\neq k} m_i, \sqrt{\sum_{i\neq k} \sigma_i^2}\Big)$$

which gives

$$I(X_k;Y) = H(Y) - H(Z) = \frac{1}{2}\log 2\pi e \sum_i \sigma_i^2 - \frac{1}{2}\log 2\pi e \sum_i \sigma_{i\neq k}^2 = \frac{1}{2}\log\frac{\sum_i \sigma_i^2}{\sigma_{i\neq k}^2}$$

8.8. (a) Since $\mathcal{R}$ is finite the number of outcomes is also finite, $\sum_{x\in\mathcal{R}} 1 = k$. A uniform distribution has probabilities $u(\boldsymbol{x}) = \frac{1}{k}$, and the entropy $H_u(\boldsymbol{X}) = -\sum_{\boldsymbol{x}} u(\boldsymbol{x})\log u(\boldsymbol{x}) = \log k = \log\frac{1}{u(\boldsymbol{x})}$. Let $p(\boldsymbol{x})$ be any probability function defined over the same region. Then,

$$H_u(\boldsymbol{X}) - H_p(\boldsymbol{X}) = \sum_{\boldsymbol{x}} p(\boldsymbol{x})\log\frac{1}{u(\boldsymbol{x})} + \sum_{\boldsymbol{x}} p(\boldsymbol{x})\log p(\boldsymbol{x})$$
$$= \sum_{\boldsymbol{x}} p(\boldsymbol{x})\log\frac{p(\boldsymbol{x})}{u(\boldsymbol{x})} = D(p||u) \geq 0, \quad \text{eq. iff } p(\boldsymbol{x}) = u(\boldsymbol{x})$$

In other words, $H_u(\boldsymbol{X}) \geq H_p(\boldsymbol{X})$, with equality if and only if $p(\boldsymbol{x}) = u(\boldsymbol{x})$, $\forall \boldsymbol{x} \in \mathcal{R}$.

(b) Since $\mathcal{R}$ is finite its volume is also finite, $\int_{\mathcal{R}} 1 d\boldsymbol{x} = V$. A uniform distribution has density function $u(\boldsymbol{x}) = \frac{1}{V}$, and the entropy $H_u(\boldsymbol{X}) = -\int_{\mathcal{R}} u(\boldsymbol{x})\log u(\boldsymbol{x}) = \log V = \log\frac{1}{u(\boldsymbol{x})}$. Let $f(\boldsymbol{x})$ be any density function defined over the same region. Then,

$$H_u(\boldsymbol{X}) - H_f(\boldsymbol{X}) = \int_{\mathcal{R}} f(\boldsymbol{x})\log\frac{1}{u(\boldsymbol{x})}d\boldsymbol{x} + \int_{\mathcal{R}} f(\boldsymbol{x})\log f(\boldsymbol{x})d\boldsymbol{x}$$
$$= \int_{\mathcal{R}} f(\boldsymbol{x})\log\frac{f(\boldsymbol{x})}{u(\boldsymbol{x})}d\boldsymbol{x} = D(f||u) \geq 0, \quad \text{eq. iff } f(\boldsymbol{x}) = u(\boldsymbol{x})$$

In other words, $H_u(\boldsymbol{X}) \geq H_p(\boldsymbol{X})$, with equality if and only if $f(\boldsymbol{x}) = u(\boldsymbol{x})$, $\forall \boldsymbol{x} \in \mathcal{R}$.

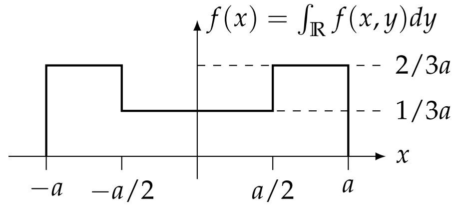
8.9. (a) To simplify notations, let $\mathcal{B}$ denote the shaded region in the figure. Then, since the area of $\mathcal{B}$ is $3ab$, the density function is

$$f(x,y) = \begin{cases} \frac{1}{3ab}, & x,y \in \mathcal{B} \\ 0, & x,y \notin \mathcal{B} \end{cases}$$

The entropy is

$$H(X,Y) = -\int_{\mathcal{B}} \frac{1}{3ab} \log \frac{1}{3ab} dxdy = \log \frac{3}{a} b \int_{\mathcal{B}} \frac{1}{3ab} dxdy = \log 3ab$$

(b) To get $f(x)$, integrate $f(x,y)$ over y, to get



Then the entropy of X can be derived as

$$\begin{aligned} H(X) &= -\int_{-a}^{-a/2} \frac{2}{3a} \log \frac{2}{3a} dx - \int_{-a/2}^{a/2} \frac{1}{3a} \log \frac{1}{3a} dx - \int_{a/2}^{a} \frac{2}{3a} \log \frac{2}{3a} dx \\ &= a\frac{2}{3a} \log \frac{3a}{2} + a\frac{1}{3a} \log 3a = \log 3a - \frac{2}{3} \end{aligned}$$

Similarly, $H(Y) = \log 3b - \frac{2}{3}$.

(c) The mutual information is

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= \log 3a - \frac{2}{3} + \log 3b - \frac{2}{3} - \log 3ab = \log 3 - \frac{4}{3} \end{aligned}$$

(d) Since $I(X;Y) = H(X) - H(X|Y)$,

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

Similarly, $H(Y|X) = \frac{2}{3} - \log b$.

8.10.

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

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

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

8.12. (a)

$$\begin{aligned}
D(f_1||f_2) &= \int_{\mathbb{R}} f_1(x)\log\frac{f_1(x)}{f_2(x)}dx\\
&= \int_{\mathbb{R}} f_1(x)\log\frac{\frac{1}{\sqrt{2\pi\sigma_1^2}}e^{-(x-\mu_1)^2/2\sigma_1^2}}{\frac{1}{\sqrt{2\pi\sigma_2^2}}e^{-(x-\mu_2)^2/2\sigma_2^2}}dx\\
&= \int_{\mathbb{R}} f_1(x)\log\frac{\sigma_2}{\sigma_1}e^{-(x-\mu_1)^2/2\sigma_1^2}e^{(x-\mu_2)^2/2\sigma_2^2}dx\\
&= \log\frac{\sigma_2}{\sigma_1} - \frac{\log e}{2\sigma_1^2}\int_{\mathbb{R}} f_1(x)(x-\mu_1)^2dx + \frac{\log e}{2\sigma_2^2}\int_{\mathbb{R}} f_1(x)(x-\mu_2)^2dx\\
&= \log\frac{\sigma_2}{\sigma_1} - \frac{1}{2\ln 2} + \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2\ln 2}
\end{aligned}$$

where in the last equality it is used that

$$E_1\big[(X-\mu_2)^2\big] = E_1\big[X^2\big] - 2E_1\big[X\big] + \mu_2^2 = \sigma_1^2 + \mu_1^2 - 2\mu_1\mu_2 + \mu_2^2 = \sigma_1^2 + (\mu_1-\mu_2)^2$$

(b) Since $X = X_1 + X_2 \sim \mathrm{N}(\mu_1+\mu_2, \sqrt{\sigma_1^2+\sigma_2^2})$,

$$\begin{aligned}
D(f||f_2) &= \log\frac{\sigma_2}{\sigma} - \frac{1}{2\ln 2} + \frac{\sigma^2+(\mu-\mu_2)^2}{2\sigma_2^2\ln 2}\\
&= \log\frac{\sigma_2}{\sqrt{\sigma_1^2+\sigma_2^2}} - \frac{1}{2\ln 2} + \frac{\sigma_1^2+\sigma_2^2+(\mu_1+\mu_2-\mu_2)^2}{2\sigma_2^2\ln 2}\\
&= \frac{1}{2}\log\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2} - \frac{1}{2\ln 2} + \frac{\sigma_1^2}{2\sigma_2^2\ln 2} + \frac{\sigma_2^2}{2\sigma_2^2\ln 2} + \frac{\mu_1^2}{2\sigma_2^2\ln 2}\\
&= \frac{1}{2\ln 2}\frac{\sigma_1^2+\mu_1^2}{\sigma_2^2} - \frac{1}{2}\log\Big(1+\frac{\sigma_1^2}{\sigma_2^2}\Big)
\end{aligned}$$

8.13. 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) $H(X) = \log(200-100) = \log 100 \approx 6.644$

(c) Without taking the units into account, the length in (b) is scaled with a factor 100 of the distribution in (a). Thus

$$H(X_b) = H(100X_a) = H(X_a) + \log 100 = \log 100$$

8.14. (a) $H(\boldsymbol{X}) = \sum_i H(X_i|X_{i-1}\dots X_1) \le \sum_i H(X_i) = \sum_i \frac{1}{2}\log 2\pi e\sigma_i^2 = \frac{1}{2}\log(2\pi e)^n \prod_i \sigma_i^2$

(b) Since $H(\boldsymbol{X}) = \frac{1}{2}\log(2\pi e)|\Lambda|$ we get

$$\tfrac{1}{2}\log(2\pi e)^n|\Lambda| \le \tfrac{1}{2}\log(2\pi e)^n \prod_i \sigma_i^2$$

The result is obtained by noting that log is an increasing function.

Chapter 9

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

$$C = \max_{f(X):E[Y^2]\leq P} I(X;Y) = \max_{f(X):E[Y^2]\leq P} \big(H(Y) - H(Y|X)\big)$$
$$= \max_{f(X):E[Y^2]\leq P} \big(H(Y) - H(Z)\big)$$

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

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

9.2. With the numbers given in the problem

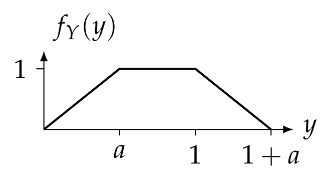
$$C = W\log\Big(1 + \frac{P}{WN_0}\Big) = 1 \text{ kb/s}$$

9.3. From the problem formulation we know that

$$X \sim \mathrm{U}(1)$$
$$Z \sim \mathrm{U}(a), \quad 0 < a \leq 1$$

Then the additiv result $Y = X + Z$ has the density function

$$f_Y(y) = f_X(x) * f_Z(z) = \begin{cases} \frac{y}{a} & 0 \leq y \leq a \\ 1 & a \leq y \leq 1 \\ 1 + \frac{1}{a} - \frac{y}{a} & 1 \leq y \leq 1+a \end{cases}$$



The mutual information can be derived as

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(Z) = H(Y) - \log a$$

To derive the entropy of Y split the derivation in three parts according to the linear slopes in the density function. Use that $\int y\ln y dy = \frac{y^2}{2}\ln y - \int \frac{y^2}{2}\frac{1}{y}dy = \frac{y^2}{2}\ln y - \frac{y^2}{4}$.

$$H_1 = -\int_0^a \frac{y}{a}\log\frac{y}{a}dy = \frac{1}{a}\log a\int_0^a ydy - \frac{1}{a}\int_0^a y\log ydy = \frac{a}{2}\log a - \frac{a}{2}\log a + \frac{a}{4\ln 2} = \frac{a}{4\ln 2}$$
$$H_2 = -\int_a^1 1\log 1 dy = 0$$
$$H_3 = -\int_1^{1+a}\Big(1+\frac{1}{a}-\frac{y}{a}\Big)\log\Big(1+\frac{1}{a}-\frac{y}{a}\Big)dy = \begin{bmatrix} s = a+1-y \\ ds = -dy \\ y = 1 \Rightarrow s = a \\ y = 1+a \Rightarrow s = 0 \end{bmatrix}$$
$$= -\int_0^a \frac{s}{a}\log\frac{s}{a}ds = H_1(Y) = \frac{a}{4\ln 2}$$

The variable substitution is found from $\frac{s}{a} = 1 + \frac{1}{a} - \frac{y}{a}$. Summing up gives the entropy $H(Y) = H_1 + H_2 + H_3 = \frac{a}{2\ln 2}$, and the mutual information becomes

$$I(X;Y) = \frac{a}{2\ln 2} - \log a$$

9.4. (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

$$\frac{N}{2} = \frac{N_1|H_2|^2 + N_2}{2}$$

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

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

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

$$\begin{aligned}
\mathrm{SNR}_1 &= \frac{|H_1|^2 P_X}{W N_1} = \frac{P_Y}{W N_1} \\
\mathrm{SNR}_2 &= \frac{|H_2|^2 P_Y}{W N_2} = \frac{P_Z}{W N_2}
\end{aligned}$$

Then the total SNR can be expressed as

$$\begin{aligned}
\mathrm{SNR} &= \frac{|H_1|^2|H_2|^2 P_X}{W(N_1|H_2|^2 + N_2)} = \frac{\frac{|H_1|^2 P_X |H_2|^2 P_Y}{W N_1 W N_2}}{\frac{P_Y}{W^2 N_1 N_2} W(N_2 + N_1|H_2|^2)} \\
&= \frac{\frac{|H_1|^2 P_X}{W N_1}\frac{|H_2|^2 P_Y}{W N_2}}{\frac{P_Y}{W N_1} + \frac{|H_1|^2 P_Y}{W N_2}} = \frac{\mathrm{SNR}_1 \cdot \mathrm{SNR}_2}{\mathrm{SNR}_1 + \mathrm{SNR}_2}
\end{aligned}$$

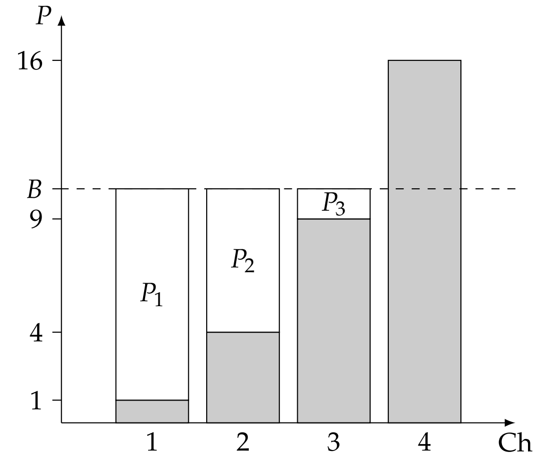
Notice, that by considering the invers of the SNR, the noise to signal ratio, the derivations can be considerably simplified,

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

which is equivalent to the desired result.

9.5. 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

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



9.6. (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

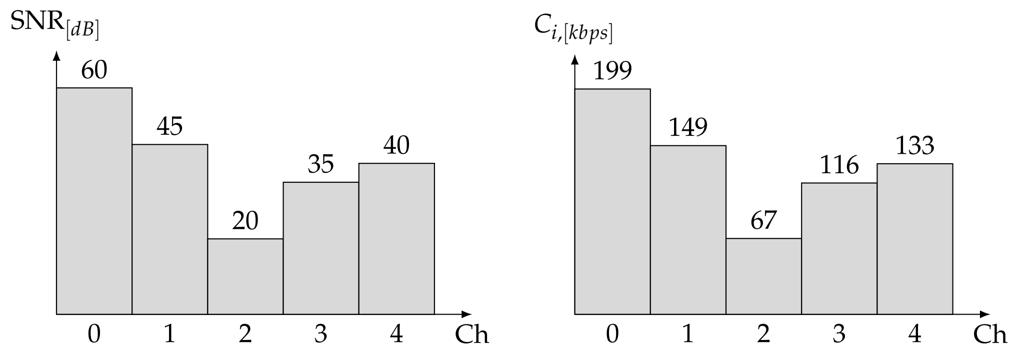
$$\begin{aligned} C &= \sum_i \frac{1}{2}\log\Big(1 + \frac{19/6}{N_i}\Big) \\ &= 0.24 + 0.17 + 0.15 + 0.20 + 0.13 + 0.31 = 1.19 \text{ bit/transmission} \end{aligned}$$

(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.

9.7. The allowed power level $P_\Delta = -60$ dBm/Hz $= 10^{-60/10}$ mW/Hz gives the total power in a sub-channel as $P = P_\Delta W$ mW, where $W = 10$ kHz is the bandwidth. A useful measure of the SNR is (in linear scale, i.e. not dB)

$$\mathrm{SNR}_i = \frac{P|G_i|^2}{N_{0,i}W} = \frac{P_\Delta W|G_i|^2}{N_{0,i}W} = \frac{P_\Delta|G_i|^2}{N_{0,i}}$$

In dB scale this gives $\text{SNR}_i = P_\Delta + |G_i|^2 - N_{0,i}$ and is shown below,



The capacity per sub-channel is derived as $C_i = W \log\bigl(1 + \text{SNR}\bigr)$, shown above. The total capacity is the sum, $C = \sum_i C_i = 199 + 149 + 67 + 116 + 133 = 665$ kbps.

9.8. Use the water-filling procedure and iterate the sub-channel powers.

1. The first iteration gives

$$BW_\Delta = 2.08 \cdot 10^{-5}\ \text{mW}$$
$$P_i = \bigl(-0.42 \quad 0.21 \quad 0.11 \quad 0.21\bigr) \cdot 10^{-4}\ \text{mW}$$

Since the first sub-channel has negative power it should be turned of.

2. Second iteration:

$$BW_\Delta = 6.73 \cdot 10^{-6}\ \text{mW}$$
$$P_i = \bigl(0 \quad 0.66 \quad -0.33 \quad 0.67\bigr) \cdot 10^{-5}\ \text{mW}$$

Again, one sub-channel has negative power and sub-channel 3 is turned off.

3. Third iteration

$$BW_\Delta = 5.09 \cdot 10^{-6}\ \text{mW}$$
$$P_i = \bigl(0 \quad 0.496 \quad 0 \quad 0.504\bigr) \cdot 10^{-5}\ \text{mW}$$

Since all derived powers in the third iteration are positive, they can be used to derive the capacity as,

$$C = 5.34 + 6.67 = 12\ \text{kb/s}$$

9.9. With the parameters in the problem

$$P = -20\ \text{dBm} = 0.01\text{mW}$$
$$N_0 = \bigl(0.005 \quad 0.012 \quad 0.32 \quad 0.0016 \quad 0.0013 \quad 0.0032 \quad 0.031 \quad 0.16 \quad 0.079 \quad 0.025\bigr)\, 10^{-11}\ \text{mW/Hz}$$

(a) Water-filling iterations:

1.

$$BW_\Delta = 0.0796\ \text{mW}$$
$$P_i = \bigl(0.080 \quad 0.080 \quad 0.077 \quad 0.080 \quad 0.080 \quad 0.080 \quad 0.064 \quad -0.12 \quad -0.24 \quad -0.17\bigr)\ \text{mW}$$

2.

$$BW_\Delta = 0.0041 \text{ mW}$$
$$P_i = \begin{pmatrix}0.0041 & 0.0041 & 0.0016 & 0.0041 & 0.0041 & 0.0036 & -0.012 & 0 & 0 & 0\end{pmatrix} \text{ mW}$$

3.

$$BW_\Delta = 0.0022 \text{ mW}$$
$$P_i = \begin{pmatrix}0.0022 & 0.0022 & -0.0003 & 0.0022 & 0.0021 & 0.0017 & 0 & 0 & 0 & 0\end{pmatrix} \text{ mW}$$

4.

$$BW_\Delta = 0.0021 \cdot 10^{-7} \text{ mW}$$
$$P_i = \begin{pmatrix}0.0021 & 0.0021 & 0 & 0.0021 & 0.0021 & 0.16 & 0 & 0 & 0 & 0\end{pmatrix} \text{ mW}$$

Capacity derivation:

$$C_i = \begin{pmatrix}87.3 & 60.7 & 0 & 60.7 & 47.4 & 20.9 & 0 & 0 & 0 & 0\end{pmatrix} \text{ kb/s}$$
$$C = 277 \text{ kb/s}$$

(b) With even distribution of the power, each sub-channel has $P_i = 10^{-3}$ mW. Then

$$C_i = \begin{pmatrix}76.5 & 50.3 & 4.8 & 50.3 & 37.6 & 15.8 & 0.88 & 0.072 & 0.046 & 0.057\end{pmatrix} \text{ kb/s}$$

and $C = 236$ kb/s

9.10. To start, water-filling should be used to distribute the power over the channels with attenuation according to the singular values. The first iteration gives

$$B = \frac{1}{3}\Big(P + \sum_{i=1}^{3} \frac{N}{s_i^2}\Big) = 13.63$$
$$P_i = B - \frac{N}{s_i^2} = \begin{pmatrix}13.20 & 10.86 & -14.06\end{pmatrix}$$

Since channel $i = 3$ has a negative value it should be canceled. The second iteration only contains channels $i = 1, 2$,

$$B = \frac{1}{2}\Big(P + \sum_{i=1}^{2} \frac{N}{s_i^2}\Big) = 6.60$$
$$P_i = B - \frac{N}{s_i^2} = \begin{pmatrix}6.17 & 3.83\end{pmatrix}$$

where the power vales are all positive and the iteration can stop. Then the capacity is

$$C = \sum_{i=1}^{2} \frac{1}{2} \log\Big(1 + \frac{s_i^2 P_i}{N}\Big) = 1.96 + 0.63 = 2.59 \text{ bit/channel use}$$

To achieve this information transmission the distribution on the (transformed) transmitter is $\widetilde{\boldsymbol{X}} \sim \mathrm{N}(\mathbf{0}, \Lambda_{\widetilde{X}})$, where

$$\Lambda_{\widetilde{X}} = \begin{pmatrix}6.17 & 0 & 0\\ 0 & 3.83 & 0\\ 0 & 0 & 0\end{pmatrix}$$

That gives the distribution $\boldsymbol{X} \sim \mathrm{N}(\mathbf{0}, \Lambda_X)$, where

$$\Lambda_X = V \Lambda_{\widetilde{X}} V^T = \begin{pmatrix}5.01 & 1.05 & 0.95\\ 1.05 & 0.41 & 1.12\\ 0.95 & 1.12 & 4.58\end{pmatrix}$$

9.11. Starting the water-filling to find the power distribution over the channels $i \in \{1,2,3,4\}$ gives

$$B_1 = \frac{1}{4}\Big(P + \sum_{i=1}^{4} \frac{N}{s_i^2}\Big) = 9.90$$
$$P_i = B_1 - \frac{N}{s_i^2} = \begin{pmatrix} 9.65 & 7.38 & 0.45 & -12.48 \end{pmatrix}$$

Removing channel $i = 4$ due to its negative power, the next iteration uses the channels $i \in \{1,2,3\}$,

$$B_2 = \frac{1}{3}\Big(P + \sum_{i=1}^{3} \frac{N}{s_i^2}\Big) = 5.74$$
$$P_i = B_2 - \frac{N}{s_i^2} = \begin{pmatrix} 5.49 & 3.22 & -3.71 \end{pmatrix}$$

Again, one of the channels has been assigned negative power so it should be removed. Continuing with the channels $i \in \{1,2\}$ gives

$$B_3 = \frac{1}{2}\Big(P + \sum_{i=1}^{2} \frac{N}{s_i^2}\Big) = 3.88$$
$$P_i = B_3 - \frac{N}{s_i^2} = \begin{pmatrix} 3.63 & 1.37 \end{pmatrix}$$

All channels have positive powers which gives the capacity

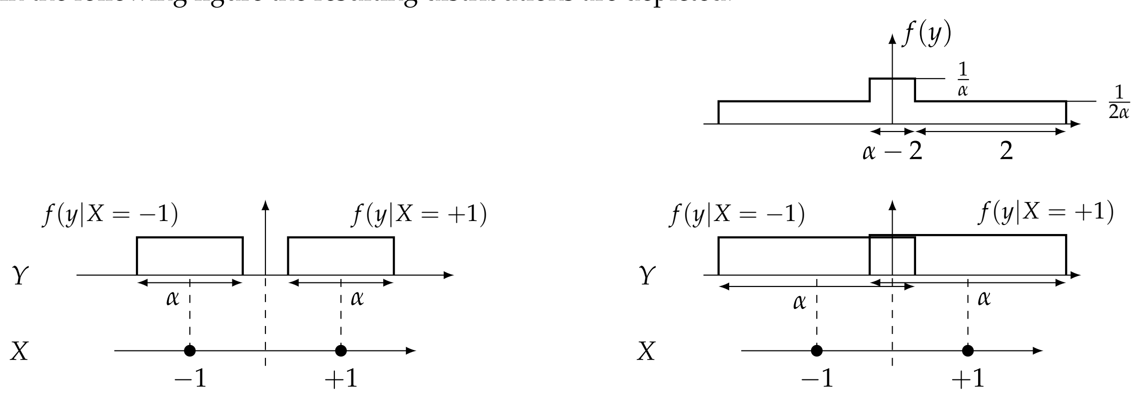
$$C = \sum i = 1^2 \frac{1}{2} \log\Big(1 + \frac{s_i^2 P_i}{N}\Big) = 1.99 + 0.31 = 2.30 \text{ bit/channel use}$$

The distribution on $\widetilde{X}$ is Gaussian with zero mean and covariance matrix

$$\Lambda_{\widetilde{X}} = \begin{pmatrix} 3.63 & 0 & 0 & 0 & 0 \\ 0 & 1.37 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

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:

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

(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,

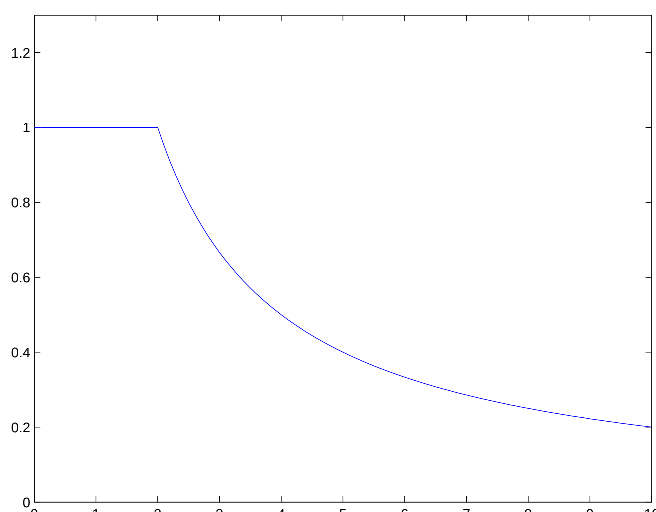
$$\begin{aligned}
H(Y) &= -\int_{-1-\frac{\alpha}{2}}^{1-\frac{\alpha}{2}} \frac{1}{2\alpha} \log \frac{1}{2\alpha} dx - \int_{1-\frac{\alpha}{2}}^{-1+\frac{\alpha}{2}} \frac{1}{\alpha} \log \frac{1}{\alpha} dx - \int_{-1+\frac{\alpha}{2}}^{1+\frac{\alpha}{2}} \frac{1}{2\alpha} \log \frac{1}{2\alpha}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

Thus, $I(X;Y) = \frac{2}{\alpha}$ bit/channel use.

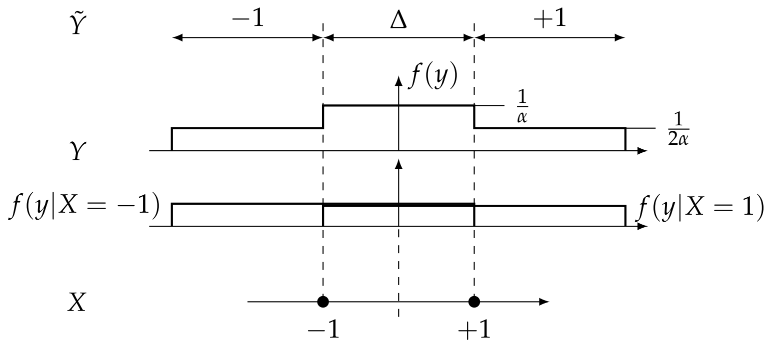
Summarising, the mutual information becomes

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

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 $-3 \le Y \le -2$, $(p_{-2}+p_{-1})/2$ for $-2 \le Y \le -1$ 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 $H(Y) = \log(3-(-3)) = \log 6$. That gives $C = \log 6 - \log 2 = \log 3$.

10.4. (a) The maximum amount of information per transmission is given by the mutual information, $I(X;Y) = H(Y) - H(Y|X)$. Here

$$H(Y|X) = H(Z) = -\int_{-1.5}^{-0.5} \frac{1}{4}\log\frac{1}{4}dz - \int_{-0.5}^{0.5} \frac{1}{2}\log\frac{1}{2}dz - \int_{0.5}^{1.5} \frac{1}{4}\log\frac{1}{4}dz = \frac{3}{2}$$

Since the outcomes for X are equally likely, in this case the density function for Y becomes

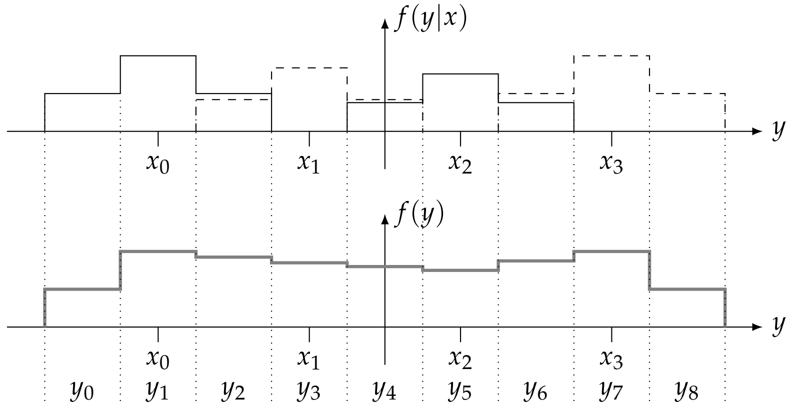
$$f(y) = \begin{cases} \frac{1}{8}, & -3.5 \le y \le 3.5 \\ \frac{1}{16}, & -4.5 \le y < -3.5 \text{ and } 3.5 < y \le 4.5 \\ 0, & \text{o.w.} \end{cases}$$

which gives the entropy

$$H(Y) = -\int_{-4.5}^{-3.5} \frac{1}{16}\log\frac{1}{16}dy - \int_{-3.5}^{3.5} \frac{1}{8}\log\frac{1}{8}dy - \int_{3.5}^{4.5} \frac{1}{16}\log\frac{1}{16}dy = \frac{25}{8}$$

and thus, $I(X;Y) = \frac{25}{8} - \frac{3}{2} = \frac{13}{8} \approx 1.625$ bit/transmission

(b) When assuming that X is not equally distributed, still $H(Y|X) = H(Z) = \frac{3}{2}$. So to maximise $I(X;Y)$ we need to maximise $H(Y)$. In the figure below the density functions for $\{Z|X\}$ are shown. These sum up to give the density function for Y. Since it is composed of flat areas, i.e. intervalls in which $f(y)$ is constant, it is possible to construct an equivalent DMC with symbols $\{y_0, y_1, \dots, y_8\}$ corresponding to intervals.



To maximise over all distributions on X we can by symmetry reasons set $P(x_0) = P(x_3) = p$ and $P(x_1) = P(x_2) = \frac{1}{2} - p$. Then the distributions on the intervals becomes

$$P(y_i) = \begin{cases} \frac{p}{4}, & i = 0, 8 \\ \frac{p}{2}, & i = 1, 7 \\ \frac{1}{8}, & i = 2, 6 \\ \frac{1}{4} - \frac{p}{2}, & i = 3, 4, 5 \end{cases}$$

Then the entropy becomes

$$\begin{aligned} H(Y) &= -2\frac{p}{4}\log\frac{p}{4} - 2\frac{p}{2}\log\frac{p}{2} - 2\frac{1}{8}\log\frac{1}{8} - 3(\frac{1}{4} - \frac{p}{2})\log(\frac{1}{4} - \frac{p}{2}) \\ &= \dots = \frac{3}{4}h(2p) + \frac{p}{2} + \frac{9}{4} \end{aligned}$$

Letting the derivative equal to zero gives

$$\frac{\partial}{\partial p}H(Y) = \frac{3}{4}\Big(2\log(1-2p) - 2\log 2p\Big) + \frac{1}{2} = \frac{3}{2}\log\frac{1-2p}{2p} + \frac{1}{2} = 0$$

or, equivalently,

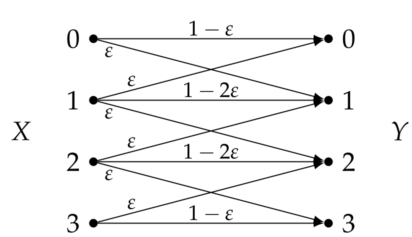
$$p = \frac{1}{2^{2/3} + 2}$$

which gives $H(Y) \approx 3.1322$ and $I(X;Y) \approx 1.6322$ bit/transmission.
The average power is increased from $E[X^2] = 5$ for equally distribution to $E[X^2] = 2 \cdot 3^2 p + 2(\frac{1}{2} - p) \approx 5.46$ for the optimal distribution.

10.5. (a)

$P(Y|X)$

X	$Y=0$	$Y=1$	$Y=2$	$Y=3$
0	$1-\varepsilon$	ε	0	0
1	ε	$1-2\varepsilon$	ε	0
2	0	ε	$1-2\varepsilon$	ε
3	0	0	ε	$1-\varepsilon$



(b) With $p(x) = \frac{1}{4}$ the joint probability is $p(x,y) = \frac{1}{4}p(y|x)$ and, consequently,

$$p(y) = \sum_x p(x,y) = \begin{cases} \frac{1}{4}(1-\varepsilon) + \frac{1}{4}\varepsilon = \frac{1}{4}, & y = 0,3 \\ \frac{1}{4}(1-2\varepsilon) + 2\frac{1}{4}\varepsilon = \frac{1}{4}, & y = 1,2 \end{cases}$$

That is, $H(Y) = \log 4 = 2$. Then,

$$\begin{aligned} H(Y|X) &= \sum_x H(Y|X{=}x)P(X{=}x) = 2\tfrac{1}{4}H(1-\varepsilon),\varepsilon) + 2\tfrac{1}{4}H(1-2\varepsilon,\varepsilon,\varepsilon) \\ &= \tfrac{1}{2}h(\varepsilon) + \tfrac{1}{2}\Big(h(2\varepsilon) + 2\varepsilon\,\underbrace{h(\tfrac{\varepsilon}{2\varepsilon})}_{=1}\Big) = \tfrac{1}{2}h(\varepsilon) + \tfrac{1}{2}h(2\varepsilon) + \varepsilon \end{aligned}$$

where in the third equality it is used that $H(\alpha,\beta,\gamma) = h(\alpha) + (1-\alpha)h(\frac{\beta}{1-\alpha})$, see Problem 3.14. The mutual information then becomes

$$I(X;Y) = H(Y) - H(Y|X) = 2 - \tfrac{1}{2}h(\varepsilon) - \tfrac{1}{2}h(2\varepsilon) - \varepsilon$$

In Figure 24 the mutual information is plotted as a function of ε.

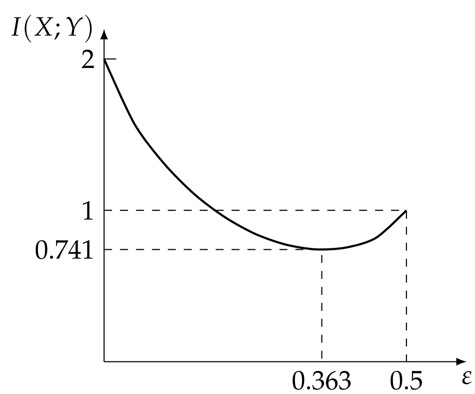
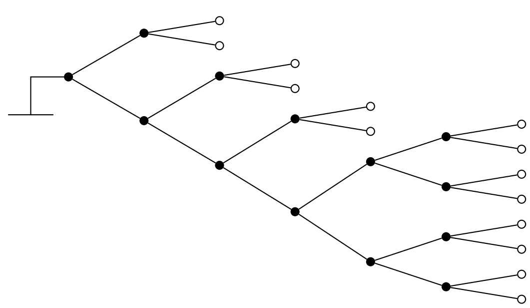
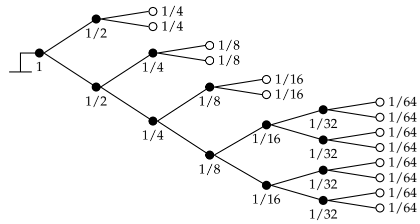


Figure 24: $I(X;Y)$ as a function of ε.

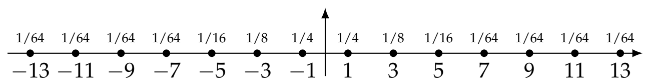
10.6. It is missing a tree in the problem formulation in the book. In sub-problem a the following tree should be attached:



(a) The original system is 8-PAM with equal probabilities, which has a second order moment $E[X^2] = 21$. Assigning equal probability for zero and one in the tree, the probabilities for the nodes in the tree becomes

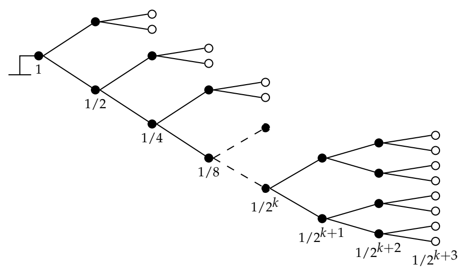


By summing the inner node probabilities, the average length is 3, hence it is comparable in bit rate with the 8-PAM system. If the leaves are mapped to signals in an 14-PAM constellation where the least probable nodes has the highes distance to the center, the following constellation probabilities are obtained



Thus, the second order moment is $E[X^2] = 19$, which gives the shaping gain $\gamma_s = 10\log_{10}\frac{21}{19} = 0.4347$ dB.

(b) Below is a tree with k levels and the probabilities for the levels.



The average lengths for the paths in the tree is, according to the path length lemma,

$$
\begin{aligned}
L &= 1 + 2\frac{1}{2} + 2\frac{1}{4} + 2\frac{1}{8} + \cdots + 2\frac{1}{2^{k+1}} + 4\frac{1}{2^{k+2}} \\
&= 1 + 1 + \frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^k} + \frac{1}{2^k} \\
&= 1 + \sum_{i=0}^{k}(\tfrac{1}{2})^i + \frac{1}{2^k} = 1 + \frac{1-(\frac{1}{2})^{k+1}}{1-\frac{1}{2}} + \frac{1}{2^k} = 1 + 2 - (\tfrac{1}{2})^k + \frac{1}{2^k} = 3
\end{aligned}
$$

(c) In the tree there are $2k+4$ leaves, each mapped to a signal point in a PAM constellation. Mapping high probability leaves to short distance fro origin, means that on the positive signal axis, signal $2i-1$ has probability $1/2^{i+1}$, $i = 1, \dots k$. Then there are also four signals at positions $2k+1$, $2k+3$, $2k+5$ and $2k+7$ with probabilities $1/2^{k+3}$. Deriving the second order moment for the positive half gives

$$
\begin{aligned}
\frac{1}{2}E[X^2] &= \sum_{i=1}^{k}(2i-1)^2\frac{1}{2^{i+1}} + \frac{1}{2^{k+3}}\Big(\underbrace{(2k+1)^2 + (2k+3)^2 + (2k+5)^2 + (2k+7)^2}_{\approx 4\cdot 4k^2 = 2^4k^2,\ k \text{ large}}\Big) \\
&= \underbrace{2\sum_{i=1}^{k} i^2\frac{1}{2^i}}_{\to 12} + \underbrace{2\sum_{i=1}^{k} i\frac{1}{2^i}}_{\to 4} + \frac{1}{2}\underbrace{\sum_{i=1}^{k}\frac{1}{2^i}}_{\to\frac{1}{2}} + \underbrace{\frac{2^2}{2^{k-1}}}_{\to 0} \to \frac{17}{2}, \quad k \to \infty
\end{aligned}
$$

Hence, $E[X^2] \to 17$ as k grows to infinity. The shaping gain is $\gamma_s = 10\log_{10}\frac{21}{17} = 0.9177$ dB.

10.7. (a) The radius is derived in the book as $R = \frac{\Gamma(\frac{N}{2}+1)^{1/N}}{\sqrt{\pi}}$.

(b) The integral specifies that one variable (dimension) is fixed and the integral spans over the remaining dimensions. This gives the projection in one dimension. Using the formula

$$\int_{|\boldsymbol{x}|^2 \le R^2} f(|\boldsymbol{x}|)d\boldsymbol{x} = \frac{2\pi^{N/2}}{\Gamma\left(\frac{N}{2}\right)} \int_0^R x^{N-1} f(x)dx$$

gives

$$\begin{aligned} f(x) = f_X(x) &= \int_{|\tilde{\boldsymbol{x}}|^2 \le R^2 - x^2} d\tilde{\boldsymbol{x}} = \frac{2\pi^{\frac{N-1}{2}}}{\Gamma(\frac{N-1}{2})} \int_0^{\sqrt{R^2-x^2}} x^{N-2} dx \\ &= \frac{2\pi^{\frac{N-1}{2}}}{\Gamma(\frac{N-1}{2})} \frac{(R^2-x^2)^{\frac{N-1}{2}}}{N-1} \\ &= \frac{\pi^{\frac{N-1}{2}}}{\Gamma(\frac{N+1}{2})} \left(\frac{\Gamma\left(\frac{N}{2}+1\right)^{2/N}}{\pi} - x^2 \right)^{\frac{N-1}{2}} \end{aligned}$$

where in the last equality the radius is inserted and it is used that $s\Gamma(s) = \Gamma(s+1)$.

(c) Using Stirling's approximation gives

$$\begin{aligned} f(x) &\approx \frac{\pi^{\frac{N-1}{2}}}{\left(\frac{\frac{N-1}{2}}{e}\right)^{\frac{N-1}{2}}} \left(\frac{\left(\frac{\frac{N}{2}}{e}\right)^{\frac{N}{2}\frac{2}{N}}}{\pi} - x^2 \right)^{\frac{N-1}{2}} \\ &= \left(\frac{\pi 2e}{N-1} \left(\frac{N}{2\pi e} - x^2 \right) \right)^{\frac{N-1}{2}} \\ &= \left(1 + \frac{\frac{1}{2} - \pi e x^2}{\frac{N-1}{2}} \right)^{\frac{N-1}{2}} \end{aligned}$$

(d) Letting $N \to \infty$, and hence $\frac{N-1}{2} \to \infty$, gives

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

where $\sigma^2 = \frac{1}{2\pi e}$. Hence, projecting the infinity-dimensional spherical uniform distribution, to one dimension gives the Normal distribution, $X \sim N(0, \sqrt{\frac{1}{2\pi e}})$.

10.8. (a) The capacity is

$$C = W \log\left(1 + \frac{P}{WN_0}\right) = 67 \text{ kb/s}$$

(b) For $P_e = 10^{-6}$ the SNR gap is $\Gamma \approx 9$ dB ≈ 7.94, which gives

$$R_{b,\max} = W \log\left(1 + \frac{P}{\Gamma W N_0}\right) = 38 \text{ kb/s}$$

10.9. The SNR gaps for $P_e \in \{10^{-3}, 10^{-6}, 10^{-9}\}$ becomes

$$\Gamma = \frac{1}{3}\left(Q^{-1}\left(\frac{P_e}{2}\right)\right)^2 = \{3.61, 7.98, 12.44\}$$

With

$$\frac{R_{b,\max}}{C} = \frac{\log\left(1+\frac{P}{\Gamma W N_0}\right)}{\log\left(1+\frac{P}{W N_0}\right)}$$

the plots can be drawn as in Figure 25. The gray dashed line in plot corresponds to the value derived in Problem 10.8. The plots show the estimated bit rate for an uncoded PAM system compared to the capacity. For good channels (high SNR) the system works well, and the better the channel becomes the closer to capacity it will come. On the other hand, the worse the channel becomes the worse the system works, and there is a need for improvements using coding.

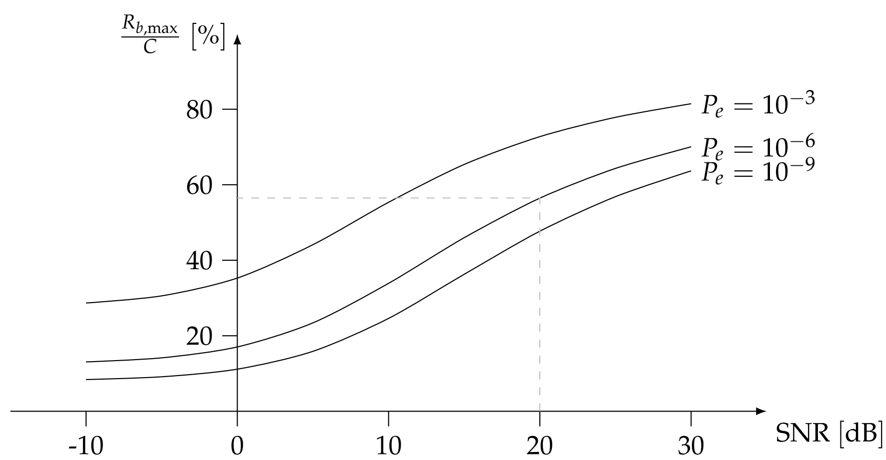


Figure 25: Relation between achievable bit rate and the capacity as a function of the SNR.

10.10. (a) There is a constant power level of $P = -70$ dBm/Hz over the whole bandwith. Similarly, the noise level is $N_0 = -140$ dBm/Hz. However the attenuation of the transmitted signal varies over the channel as $|H_i|^2 = 5i + 10$ dB. (In reality this can resemblance copper cable transmission, where the cable act as a low-pass filter, attenuating higher frequencies stronger than lower. However, the attenuation curve is a bit more complicated than a linearly decreasing function.)

The received signal to noise ratio for each sub-channel becomes

$$\mathrm{SNR}_i = -70 - (5i + 10) + 140 = 60 - 5i \text{ dB}$$

and the derived capacity per sub-channel

$$C_i = \Delta f \log(1 + 10^{\mathrm{SNR}_i/10}) = 10^4 \log\left(1 + 10^{(60-5i)/10}\right)$$

In the following table the attenuation, SNR and capacity is listed for the sub-channels

i :	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\lvert H_i\rvert^2$[dB] :	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85
SNR_i[dB] :	60	55	50	45	40	35	30	25	20	15	10	5	0	-5	-10	-15
C_i[kbps] :	199	183	166	149	133	116	100	83	66	50	35	21	10	4.0	1.4	0.45

Summing over all sub-channels gives the total capacity as

$$C = \sum_i C_i = 1317 \text{ [kbps]}$$

(b) Instead of the capacity, we want to derive an estimate of the established bit rate when the system is working with an error rate of 10^{-6} and an error correcting code with coding gain $\gamma_c = 3$ dB. The bit error rate gives an SNR gap of $\Gamma = 9$ dB, and the efficient SNR becomes

$$\widetilde{\text{SNR}}_i = \text{SNR}_i - \Gamma + \gamma_c = \text{SNR}_i - 6 \text{ dB}$$

The estimated bit rate is

$$R_i = \Delta f \log\big(1 + 10^{\widetilde{\text{SNR}}_i/10}\big)$$

In the following table the effective SNR and the estimated bit rate is shown.

i :	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\widetilde{\text{SNR}}_i$[dB] :	54	49	44	39	34	29	24	19	14	9	4	-1	-6	-11	-16	-21
R_i[kbps] :	179	163	146	130	113	96	80	63	47	32	18	8.4	3.2	1.1	0.36	0.11

The total bit rate is

$$R = \sum_i R_i = 1080 \text{ [kbps]}$$

10.11. The parameters used in the problem are $P = 1000$ mW and $W_\Delta = 1$ MHz. Since $\frac{N_{0,i}}{|G_i|^2}$ is given in dBm/Hz it has to be converted to linear scale as

$$\frac{N_{0,i}}{|G_i|^2} = 10^{\frac{[N_{0,i}/|G_i|^2]_{dB}}{10}} \quad \text{[mW/Hz]}$$

The specification that a PAM system should work at $P_e = 10^{-6}$ means the SNR gap is set to $\Gamma = 9 \text{ dB} = 7.94$.

(a) The estimated bit rate for sub-channel i is

$$R_{b,i} = W_\Delta \log\Big(1 + \frac{|G_i|^2 P_i}{W_\Delta N_{0,i} \Gamma}\Big) = W_\Delta \log\Big(1 + \frac{P_i}{\frac{N_{0,i}}{|G_i|^2} W_\Delta \Gamma}\Big)$$

where P_i is the power used in channel i. Since the sub-channels are supposed to be independent the total bit rate is $R_b = \sum_i R_{b,i}$, which should be maximised when $\sum_i P_i = P$. Assign an optimisation function

$$J = \sum_{i=1}^{10} W_\Delta \log\Big(1 + \frac{P_i}{\frac{N_{0,i}}{|G_i|^2} W_\Delta \Gamma}\Big) + \lambda\Big(\sum_{i=1}^{10} P_i - P\Big)$$

By setting the derivative equal zero, $\frac{\partial}{\partial P_j} J = 0$, gives

$$P_j + \frac{N_{0,j}}{|G_j|^2} W_\Delta \Gamma = -\frac{W_\Delta}{\lambda \ln 2} = W_\Delta B \quad \Rightarrow \quad P_j = W_\Delta\Big(B - \frac{N_{0,j}}{|G_j|^2}\Gamma\Big)$$

With Khun-Tucker optimisation this gives

$$\begin{cases} P_j = W_\Delta\Big(B - \frac{N_{0,j}}{|G_j|^2}\Gamma\Big)^+ \\ \sum_j P_j = P \end{cases}$$

which is the water-filling procedure.

(b) As a first attempt, distribute the power on all sub-channels,

$$\sum_j P_j = \sum_j W_\Delta\Big(B - \frac{N_{0,j}}{|G_j|^2}\Gamma\Big) = W_\Delta\Big(10B - \Gamma\sum_j \frac{N_{0,j}}{|G_j|^2}\Big) = P$$

Rewriting gives an expression for B as

$$B = \frac{P}{10W_\Delta} + \frac{\Gamma}{10}\sum_j \frac{N_{0,j}}{|G_j|^2} = 2.85\cdot 10^{-4}\ \text{mW/Hz} = -35.45\ \text{dBm/Hz}$$

With this the first iteration P_j can be derived as

$$P_j = (245,\ -113,\ 159,\ 185,\ -216,\ 235,\ 34,\ 34,\ 235,\ 185)\quad \text{mW}$$

Here it is seen that sub-channels 2 and 5 have negative power and have to set to zero. This will affect the noise levels on the remaining sub-channels, and the power distribution must be restarted. With eight sub-channels left the corresponding calculations give

$$B = \frac{P}{8W_\Delta} + \frac{\Gamma}{8}\sum_{j\neq 2,5} \frac{N_{0,j}}{|G_j|^2} = 2.44\cdot 10^{-4}\ \text{mW/Hz} = -36.13\ \text{dBm/Hz}$$

and the power per sub-channel

$$P_j = (204,\ 0,\ 118,\ 144,\ 0,\ 212,\ -7,\ -7,\ 194,\ 144)\quad \text{mW}$$

Again two sub-channels have negative powers, 7 and 8, and have to be disconnected.
The third iteration gives the water filling level

$$B = \frac{P}{6W_\Delta} + \frac{\Gamma}{6}\sum_{j\neq 2,5,7,8} \frac{N_{0,j}}{|G_j|^2} = 2.41\cdot 10^{-4}\ \text{mW/Hz} = -36.18\ \text{dBm/Hz}$$

which gives

$$P_j = (201,\ 0,\ 115,\ 141,\ 0,\ 210,\ 0,\ 0,\ 191,\ 141)\quad \text{mW}$$

Since all used sub-channels have positive powers, we can continue to derive the corresponding bit rates,

$$R_{b,j} = (2599,\ 0,\ 938,\ 1270,\ 0,\ 2931,\ 0,\ 0,\ 2267,\ 1270)\quad \text{kbps}$$

and the total bit rate is $\sum_j R_{b,j} = 11.28$ Mbps This can be compared with the case when the power is distributed equally in all the sub-channels, $P_j = 100$ mW, which gives a total of 9.85 Mbps.

(c) With the coding gain $\gamma_c = 3$ dB the bit rate for the jth sub-channel can be expressed as

$$R_{b,j} = W_\Delta \log\Big(1 + \frac{P_j\gamma_c}{\frac{N_{0,j}}{|G_j|^2}W_\Delta\Gamma}\Big) = W_\Delta \log\Big(1 + \frac{P_j}{\frac{N_{0,j}}{|G_j|^2}W_\Delta\Gamma_{eff}}\Big)$$

where $\Gamma_{eff} = \frac{\Gamma}{\gamma_c}$, or equivalently in dB scale $\Gamma_{eff} = \Gamma - \gamma_c = 6$ dB. That means the same optimisation as before can be used, but with $\Gamma = 6$ dB. In the first iteration

$$B = 1.93 \cdot 10^{-4}\ \text{mW/Hz} = -37.15\ \text{dBm/Hz}$$
$$P_j = (173,\ -7,\ 130,\ 143,\ -59,\ 177,\ 67,\ 67,\ 168,\ 143)$$

which means that sub-channels 2 and 5 should not be used. Then, the second iteration gives

$$B = 1.85 \cdot 10^{-4}\ \text{mW/Hz} = -37.34\ \text{dBm/Hz}$$
$$P_j = (165,\ 0,\ 121,\ 134,\ 0,\ 169,\ 59,\ 59,\ 159,\ 134)$$

which gives

$$R_{b,j} = (3209,\ 0,\ 1548,\ 1880,\ 0,\ 3541,\ 551,\ 551,\ 2877,\ 1880) \quad [\text{kbps}]$$
$$R_b = 16.04\ \text{Mbps}$$

Chapter 11

11.1. From the problem we have $P(X = j) = \frac{1}{k}$, $j = 0, 1, \ldots, k-1$, and that te Hamming distortion is used. Assign the probability of distortion as $P(X \neq \hat{X})$. Then the average distortion is $E[d(X, \hat{X})] = \delta$ which is within the minimisation criteria. The mutual information between X and $\hat{X}$ is bounded by

$$I(X; \hat{X}) = H(X) - H(X|\hat{X}) = \log k - H(X|\hat{X}) \geq \log k - \delta \log(k-1) - h(\delta)$$

where the inequality follows from Fano's lemma as

$$H(X|\hat{X}) \leq h\big(P(X \neq \hat{X})\big) + P(X \neq \hat{X}) \log(k-1) = \delta \log(k-1) + h(\delta)$$

To show the rate distortion function we need to find a distribution on $P(X|\hat{X})$ that achieves equality in the bound above. From our assumptions we get $P(X = \hat{X}) = 1 - \delta$. A reasonable attempt is to set uniform distribution for the case when $X \neq \hat{X}$, i.e.

$$P(X = j | \hat{X} = i) = \begin{cases} 1 - \delta, & i = j \\ \frac{\delta}{k-1}, & i \neq j \end{cases}$$

The conditional entropy can be derived as

$$H(X|\hat{X}) = \sum_i P(\hat{X} = i) \sum_i H(X|\hat{X} = i)$$

where

$$\begin{aligned} H(X|\hat{X} = i) &= -\sum_{j \neq j} \frac{\delta}{k-1} \log \frac{\delta}{k-1} - (1-\delta)\log(1-\delta) \\ &= -\delta \log \delta + \delta \log(k-1) - (1-\delta)\log(1-\delta) = \delta \log(k-1) + h(\delta) \end{aligned}$$

Since $iH(X|\hat{X} = i)$ is independent of i, the assumed distribution achieves equality in the bound for the mutual information, and $R(\delta) = \log k - \delta \log(k-1) - h(\delta)$. Finally, we need to find the limits on δ, where the rate distortion function reaches zero. Then, observing that when $\delta = \frac{k-1}{k}$ the conditional distribution $P(X = j|\hat{X} = i) = \frac{1}{k}$, independent of i and j, this gives a point where $H(X|\hat{X}) = \log k$. Hence, at this point $R(\delta) = 0$. Since the rate distortion function is non-increasing and non-negative,

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

11.2. (a) The Lagrange optimisation function:

$$J(f) = -\int_0^\infty f \ln f dx + \lambda_0 \Big(\int_0^\infty f dx - 1 \Big) + \lambda_1 \Big(\int_0^\infty x f dx - \frac{1}{\lambda} \Big)$$

When taking the derivative of the function above one can think of the integrals as sums over infinite vectors.

$$\frac{\partial}{\partial f} J(f) = -\ln f - 1 + \lambda_0 + \lambda_1 x = 0$$

or, equivalently,

$$f = e^{-1+\lambda_0+\lambda_1 x} = e^{\alpha+\beta x}$$

where $\alpha = -1 + \lambda_0$ and $\beta = \lambda_1$. The requirements on the distribution gives for $\beta < 0$

$$1 = \int_0^\infty e^{\alpha+\beta x} dx = -\frac{1}{\beta} e^\alpha$$
$$\frac{1}{\lambda} = \int_0^\infty x e^{\alpha+\beta x} dx = -\frac{1}{\beta^2} e^\alpha$$

which is solved by $\beta = -\lambda$ and $\alpha = \ln \lambda$, and the density function is

$$f = e^{\ln \lambda - \lambda x} = \lambda e^{-\lambda x}$$

which is the exponential distribution.

(b) The entropy of the exponential distribution, $f(x) = \lambda e^{-\lambda x}$, is

$$H_f(X) = -\int_0^\infty f(x) \ln \lambda e^{-\lambda x} dx = \lambda \int_0^\infty x f(x) dx - \ln \lambda \int_0^\infty f(x) dx = 1 - \ln \lambda$$

Let $g(x)$ be an arbitrary density function where $\int_0^\infty g(x)dx = 1$ and $\int_0^\infty x g(x) dx = 1/\lambda$. Then,

$$\begin{aligned}
H_g(X) &= -\int_0^\infty g(x) \ln g(x) dx \\
&= -\int_0^\infty g(x) \ln \frac{g(x)}{f(x)} f(x) dx \\
&= -\int_0^\infty g(x) \ln f(x) dx - D(g||f) \\
&\leq -\int_0^\infty g(x) \ln f(x) dx \\
&= \lambda \int_0^\infty x g(x) dx - \ln \lambda \int_0^\infty g(x) dx \\
&= 1 - \ln \lambda = H_f(X)
\end{aligned}$$

11.3. (a) The mutual information is bounded as

$$\begin{aligned}
I(X;\hat{X}) &= H(X) - H(X|\hat{X}) = 1 - \ln \lambda - H(X - \hat{X}|\hat{X}) \\
&\geq 1 - \ln \lambda - H(X - \hat{X}) \geq 1 - \ln \lambda - (1 - \ln \frac{1}{\delta}) = -\ln \lambda \delta
\end{aligned}$$

where the first inequality comes from dropping the condition in the entropy and the second from the exponential distribution maximising the entropy. Thus, the bound is fulfilled with equality if and only if $\{X - \hat{X}|\hat{X}\} \sim \text{Exp}(\frac{1}{\delta})$.

(b) Constructing a backward test channel from $\hat{X}$ to X can be done using an additive channel, i.e. $X = \hat{X} + Z$ where $X \sim \text{Exp}(\lambda)$ and $Z =\sim \text{Exp}(\frac{1}{\delta})$. Then the distortion requirement is fulfilled since $E[d(X,\hat{X})] = E[X - \hat{X}] = E[Z] = \delta$. To find the distribution on $\hat{X}$ consider that the density function of $X = \hat{X} + Z$ is the convolution $f_X = f_{\hat{X}} * f_Z$. The convolution is best solved in a transform plane, and thus we need the Laplace transform of an exponential distribution (or rather the density function). The transform for X is

$$E\big[e^{-sX}\big] = \int_0^\infty e^{-st}\lambda e^{-\lambda t}dt = \lambda\int_0^\infty e^{-(s+\lambda)t}dt = \frac{\lambda}{s+\lambda} = \frac{1}{1+\frac{s}{\lambda}}$$

Similarly, the transform of Z is

$$E\big[e^{-sZ}\big] = \frac{1}{1+s\delta}$$

The convolution above gives $E\big[e^{-sX}\big] = E\big[e^{-s\hat{X}}\big]E\big[e^{-sZ}\big]$ and, thus,

$$E\big[e^{-s\hat{X}}\big] = \frac{1+s\delta}{1+\frac{s}{\lambda}} = \delta\lambda + (1-\delta\lambda)\frac{1}{1+\frac{s}{\lambda}}$$

which gives the inverse transform

$$f_{\hat{X}}(\hat{x}) = \delta\lambda\delta(\hat{x}) + (1-\delta\lambda)\lambda e^{-\lambda\hat{x}}$$

where $\delta(\hat{x})$ is the Dirac function. This means that $P(\hat{X}=0) = \delta\lambda$, and for $\hat{x} > 0$ it follows an exponential distribution with $f_{\hat{X}|\hat{X}>0}(\hat{x}) = \lambda e^{-\lambda\hat{x}}$. This distribution has a meaning as lons as $P(\hat{X} = X)$ is less than one, i.e. when $0 \le \delta\lambda \le 1$, or, equivalently, $0 \le \delta \le 1/\lambda$. For $\delta > 1/\lambda$ choose $\hat{X} = 0$ to get $E[d(X,\hat{X})] = E[X-\hat{X}] = E[X] = \frac{1}{\lambda} < \delta$ and thus the requirement is fulfilled. Since $\hat{X}$ is deterministic and independent of X there is no information and $R(\delta) = 0$. Summarising, we have

$$R(\delta) = \begin{cases} -\ln(\lambda\delta), & 0 \le \delta \le 1/\lambda \\ 0, & \delta > 1/\lambda \end{cases}$$

Note: This solution is based on [Verdu96] from 1996.

11.4. Follows directly from Problem 11.3.

11.5. (a) The density function is $f(x) = \frac{1}{2\sqrt{\pi}}e^{x^2/4}$. Since everything is symmetric around $x = 0$, the derivations will be made only for the positive half. The numerical integrations that follows can be performed in different ways, here a trapezoid method was used. Assuming a set of x-values, $\boldsymbol{x} = x_1, \ldots, x_n$, with constant separation $x_i - x_{i-1} = \Delta$. Let $\boldsymbol{y} = y_1, \ldots, y_n$ be the corresponding set of function values. Then the area can be approximated by

$$\int_{x_1}^{x_n} y(x)dx \approx \Delta\Big(\sum_{i=1}^n y_i - \frac{y_1+y_n}{2}\Big)$$

To derive the distortion the intervals $\{[0,1],[1,2],[2,3],[3,\infty]\}$ is used. In the numerical derivations setting ∞ to 10 seems good enough. Then, assign $\delta_i = E[(X - X_{q,i})^2]$ to get

$$\delta_1 = \int_0^1 (x-0.5)^2 f(x)dx \approx 0.0214$$

$$\delta_2 = \int_1^2 (x-1.5)^2 f(x)dx \approx 0.0134$$

$$\delta_3 = \int_2^3 (x-2.5)^2 f(x)dx \approx 0.0053$$

$$\delta_4 = \int_3^\infty (x-3.5)^2 f(x)dx \approx 0.0036$$

To derive the total average distortion we can use $E[(X - X_q)] = \sum_i E[(X - X_{q,i})]$ over both the positive and negative side, which gives the average distortion

$$E[(X - X_q)^2] = 2(\delta_1 + \delta_2 + \delta_3 + \delta_4) \approx 0.0874$$

(b) In general,the distortion in the interval $[a, b]$ when reconstructing to x_q is $\delta = \int_a^b (x - x_q)^2 f(x)dx$. Optimising with respect to the reconstruction value gives

$$\frac{\partial}{\partial x_q}\delta = -\int_a^b 2(x - x_q)f(x)dx = 2x_q \int_a^b f(x)dx - 2\int_a^b xf(x)dx = 0$$

hence, the optimal reconstruction value is given by

$$x_q^{(opt)} = \frac{\int_a^b xf(x)dx}{\int_a^b f(x)dx} \approx \begin{cases} 0.48, & i = 1 \\ 1.44, & i = 2 \\ 2.40, & i = 3 \\ 3.51, & i = 4 \end{cases}$$

The corresponding distortion measures are given by

$$\delta_i^{(opt)} \approx \begin{cases} 0.0213, & i = 1 \\ 0.0129, & i = 2 \\ 0.0047, & i = 3 \\ 0.0036, & i = 4 \end{cases}$$

and the total distortion $E[(X - x_q^{(opt)})^2] \approx 0.0850$.

(c) If the quantiser is followed by a compression algorithm, and the samples can be viewed as independent, a limit on the number of bits per symbol is given by the entropy,

$$L \geq H(P) = 2.55 \text{ bit/sample}$$

Note: If the minimum length is estimated by a Huffman code instead, it becomes 2.6 bit/sample.

11.6. The distortion is

$$\begin{aligned}
E\Big[\Big(X - X_Q\Big)^2\Big] &= \int_{-\infty}^{0} \Big(x + \sqrt{\frac{2}{\pi}}\sigma\Big)^2 f(x) + \int_0^{\infty} \Big(x - \sqrt{\frac{2}{\pi}}\sigma\Big)^2 f(x) \\
&= \int_{-\infty}^{\infty} x^2 f(x)dx + 2\sqrt{\frac{2}{\pi}}\sigma\Big(\int_{-\infty}^{0} xf(x)dx - \int_0^{\infty} xf(x)dx\Big) + \frac{2}{\pi}\sigma^2 \int_{-\infty}^{\infty} f(x)dx \\
&= \sigma^2 - 4\sqrt{\frac{2}{\pi}}\sigma \int_0^{\infty} xf(x)dx + \frac{2}{\pi}\sigma^2 \\
&= \sigma^2 - 4\sqrt{\frac{2}{\pi}}\sigma\frac{\sigma}{\sqrt{2\pi}} + \frac{2}{\pi}\sigma^2 = \sigma^2\Big(1 - \frac{2}{\pi}\Big) = \frac{\sigma^2}{\pi}(\pi - 2)
\end{aligned}$$

where it is used that $\int_{-\infty}^{0} xf(x)dx = -\int_0^{\infty} xf(x)dx$ and that

$$\int_0^{\infty} xf(x)dx = \frac{1}{\sqrt{2\pi}\sigma}\int_0^{\infty} xe^{-x^2/2\sigma^2}dx = \frac{1}{\sqrt{2\pi}\sigma}\Big[-\sigma^2 e^{-x^2/2\sigma^2}\Big]_0^{\infty} = \frac{\sigma}{\sqrt{2\pi}}$$

11.7. In Figure 26 the quantisation intervals and the reconstruction values are shown for positive x. Clearly, since the Gaussian distribution is symmetric the negative side will be a direct mirroring of this. The quantisation error can be derives as

$$E\big[(X-x_Q)^2\big] = 2\Big(\int_0^{\Delta}\big(x-\tfrac{\Delta}{2}\big)^2 f(x)dx + \int_{\Delta}^{2\Delta}\big(x-\tfrac{3\Delta}{2}\big)^2 f(x)dx + \int_{2\Delta}^{3\Delta}\big(x-\tfrac{5\Delta}{2}\big)^2 f(x)dx + \int_{3\Delta}^{\infty}\big(x-\tfrac{7\Delta}{2}\big)^2 f(x)dx\Big)$$

where $f(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. Deriving this function using numerical integrations for different values of Δ gives Figure 27. In the figure the minimum error 0.0374 at $\Delta = 0.586$ is also depicted. The minimum value is found by narrowing down the interval of the plot.

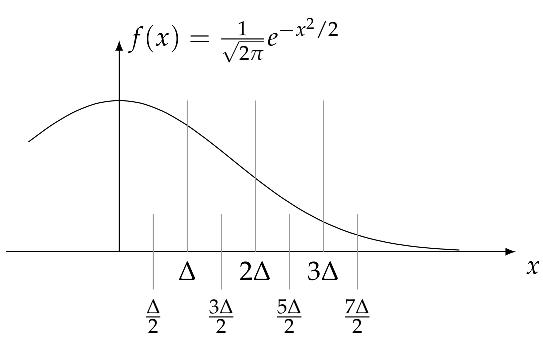


Figure 26: Limits and reconstruction levels of Problem 11.7.

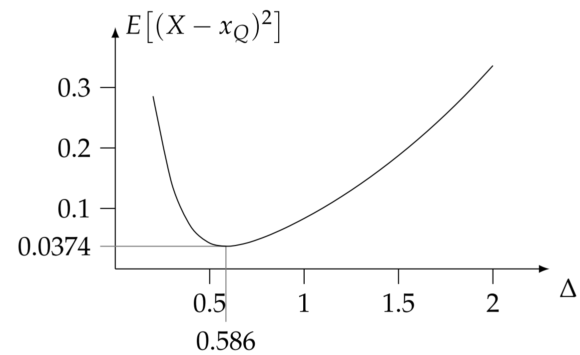


Figure 27: The quantisation error as a function of the quantisation interval Δ.

11.8. Since the logarithm is a concave function, $\ln\sum_i \frac{1}{n}a_i \geq \sum_i \frac{1}{n}\ln a_i$. This can be seen by letting a_i be an outcome of the random variable A, and $\frac{1}{n}$ the probability. Then it is equivalent to $\ln E\big[A\big] \geq E\big[\ln A\big]$ which follows directly from Jensen's inequality. Then, since the exponential function is strictly increasing for positive argument,

$$\frac{1}{n}\sum_i a_i = e^{\ln\frac{1}{n}\sum_i a_i} = e^{\ln\sum_i \frac{1}{n}a_i} \geq e^{\sum_i \frac{1}{n}\ln a_i} = e^{\sum_i \ln a_i^{1/n}} = \prod_i e^{\ln a_i^{1/n}} = \prod_i a_i^{1/n} = \Big(\prod_i a_i\Big)^{1/n}$$

There is equality in the bound if and only if $a_i = a_j$, $\forall i,j$.