LUND UNIVERSITY
Electrical and Information Technology

Information Theory

Solutions to problems

February 17, 2017

Solutions

Chapter 2

2.1. (a) From the problem we have

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

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

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

where we see that Z is independent of p.

(b) Similar to (a) let

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

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

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

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

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

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

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

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

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

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

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

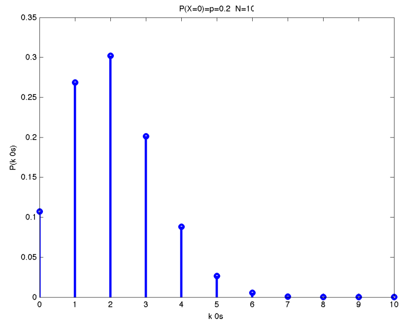
In the second alternative, consider the total number of favourable cases related to the total number of cases. There are 2^n different binary vectors (results of n flips) resulting in a total of $2^n 2^n = 2^{2n}$ different outcomes of $2n$ tosses. Among those we need to find the total number of favourable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

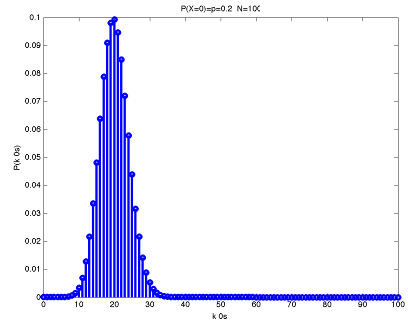
$$P(X=Y) = \frac{\text{nbr favourable cases}}{\text{nbr cases}} = \frac{\sum_{k=0}^{n}\binom{n}{k}^2}{2^{2n}}$$

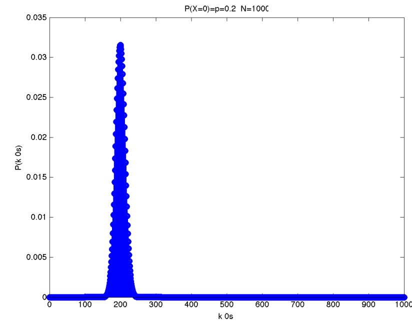
2.4. Let k be the number of 0s in the vector. Then we get the following table.

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

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







2.5. (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	0	1	2	3	4	5	6
$P(k\,\text{black})$	0.1176	0.3025	0.3241	0.1852	0.0595	0.0102	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	0	1	2	3
$\binom{6}{k}\binom{4}{3-k}$	4	36	60	20
$P(k\,\text{black})$	0.0333	0.3	0.5	0.1667

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

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

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,

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

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

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

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 achievesd equality in thelimit as $n \to \infty$. 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 $9.4 \cdot 10^{157}$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.5. (a)

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

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

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

	$P(Y\|X)$	Y: a	b	c
X	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:

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

	$P(Y)$	
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	$\frac{1}{5} + \frac{2}{15} = \frac{1}{3}$

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

	$P(Y\|X)$	Y: a	b	c
X	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 1.

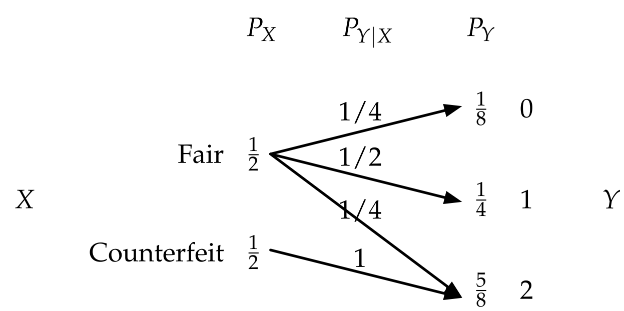


Figure 1: Probabilities or two flips with unknown coin.

Hence,

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

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

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

	$p(x,y)$	Y: b	r	g
	b	$\frac{10}{18}\frac{9}{17}$	$\frac{10}{18}\frac{5}{17}$	$\frac{10}{18}\frac{3}{17}$
X	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 horisontally we get $P(X)$. Both variants give the same results,

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

In other words,

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

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

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

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

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

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

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

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

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

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

(c) To start with derive the marginals as

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

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

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

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

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

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.

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

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

3.14. (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)

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

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

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

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

(b) The mutual information can be derived as

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

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

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

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

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

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

3.16.

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

3.17. (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.

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

3.19. Let C be a random variable specifying the cost for the icecream.

(a) It is well known that the uniform distribution, $p_i = \frac{1}{3}$, maximises the entropy to $H(C) = \log 3$. The average cost is $E[C] = \frac{1}{3}(2+3+4) = 3$ €.

(b) With the knowledge of the average cost as $E[C] = 2.5$ €and that p_i are probabilities, we can set up the Lagrangian maximisation function

$$J = -\sum_i p_i \log p_i + \lambda_0\Big(\sum_i p_i c_i - 2.5\Big) + \lambda_0\Big(\sum_i p_i - 1\Big)$$

Setting the derivative equal to zero gives

$$\frac{\partial}{\partial p_j} J = -\log p_j - \frac{1}{\ln 2} + \lambda_0 c_j + \lambda_1 = 0$$

or,

$$p_j = 2^{\lambda_0 c_j + \lambda_1 - \frac{1}{\ln 2}} = 2^{\lambda_0 c_j + \mu}$$

where $\mu = \lambda_1 - \frac{1}{\ln 2}$. The condition that p_i are probabilities gives

$$\sum_i p_i = 2^{\mu} \sum_i 2^{\lambda_0 c_i} = 1 \quad \Rightarrow \quad 2^{\mu} = \frac{1}{\sum_i 2^{\lambda_0 c_i}}$$

Hence, the probability can be written as

$$p_j = \frac{2^{\lambda_0 c_j}}{\sum_i 2^{\lambda_0 c_i}}$$

The condition on the average price gives then

$$\sum_j c_j p_j = \sum_j c_j \frac{2^{\lambda_0 c_j}}{\sum_i 2^{\lambda_0 c_i}} = \frac{\sum_j c_j 2^{\lambda_0 c_j}}{\sum_i 2^{\lambda_0 c_i}} = 2.5$$

or,

$$\sum_j c_j 2^{\lambda_0 c_j} = 2.5 \sum_i 2^{\lambda_0 c_i}$$

Written out, the equation becomes

$$(2-2.5)\big(2^{\lambda_0}\big)^2 + (3-2.5)\big(2^{\lambda_0}\big)^3 + (4-2.5)\big(2^{\lambda_0}\big)^4 = 0$$

Rewritten it gives

$$1.5\big(2^{\lambda_0}\big)^2\Big(\big(2^{\lambda_0}\big)^2 + \tfrac{0.5}{1.5}2^{\lambda_0} - \tfrac{0.5}{1.5}\Big)$$

which is solved by $2^{\lambda_0} = -\frac{1}{6} + \frac{1}{6}\sqrt{13}$. Henve the probabilities are

$$p_1 = \frac{2^{\lambda_0 2}}{2^{\lambda_0 2} + 2^{\lambda_0 3} + 2^{\lambda_0 4}} \approx 0.6162$$

$$p_2 = \frac{2^{\lambda_0 3}}{2^{\lambda_0 2} + 2^{\lambda_0 3} + 2^{\lambda_0 4}} \approx 0.2676$$

$$p_3 = \frac{2^{\lambda_0 4}}{2^{\lambda_0 2} + 2^{\lambda_0 3} + 2^{\lambda_0 4}} \approx 0.1162$$

The resulting entropy is $H(C) = 1.3002$.

3.20. (a) The transition matrix is

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

The stationary distribution is found from

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

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

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

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

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

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

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

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

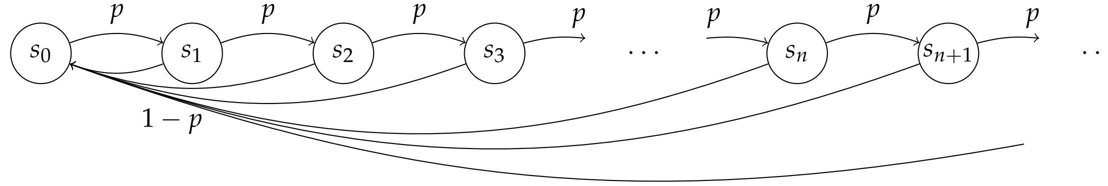
$$\boldsymbol{\mu} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{2}{9} & \frac{1}{9} \end{pmatrix}$$

which is the distribution of the islands.

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

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

3.22. (a) 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 staeady 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

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

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

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

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

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

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

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

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

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

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

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

Then, we know that the steady state distribution is

$$\boldsymbol{\mu} = \Big[\frac{W_i}{2W}\Big] = \begin{pmatrix} \frac{2}{12} & \frac{2}{12} & \frac{4}{12} & \frac{2}{12} & \frac{2}{12} \end{pmatrix} = \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$$

and the entropy rate is

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

3.24. (a) Let the probability for stat ei be $\pi_i = \frac{N_i}{N}, \forall i$. Then the probability for the next state to be j is

$$\begin{aligned} P(S^+ = j) &= \begin{pmatrix} \pi_1 & \dots & \pi_W \end{pmatrix} \begin{pmatrix} p_{1j} \\ \vdots \\ p_{Wj} \end{pmatrix} = \sum_i \pi_i p_{ij} \\ &= \sum_i \frac{k_{ij}}{N_i}\frac{N_i}{N} = \sum_i \frac{k_{ij}}{N} = \sum_i \frac{k_{ji}}{N} = \frac{N_j}{N} = \pi_j \end{aligned}$$

which shows that π_i is the stationary distribution. In the third last equality it is used that the branches are undirected and that $k_{ij} = k_{ji}$.

(b) The entropy rate is

$$\begin{aligned} H_\infty(S) &= \sum_i \pi_i H(S^+|S = i) = \sum_i \pi_i \log N_i = \sum_i \frac{N_i}{N}\log N_i \\ &= \sum_i \frac{N_i}{N}\log\frac{N_i}{N} + \sum_i \frac{N_i}{N}\log N = \log N + \sum_i \pi_i \log \pi_i = \log N - H(\boldsymbol{\pi}) \end{aligned}$$

3.25. (a) –

(b) –

(c) –

(d) –

Chapter 4

4.1. (a) Yes

(b) Yes

(c) No

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

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

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

(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!

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

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

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

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

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

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

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

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

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

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

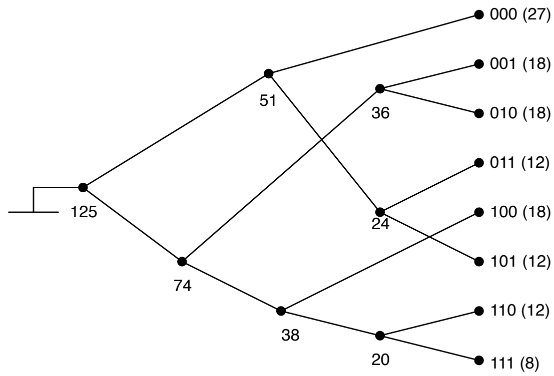
where, in the second last equality, we used that $p(x|y) = 1$ if $x = y$ and $p(x|y) = 0$ if $x \neq y$. The above derivation tells that we can gain the same amount of bits that is needed to encode sequence $\boldsymbol{y}$.

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

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

$\boldsymbol{x}$	$p(\boldsymbol{x})$	$L(\boldsymbol{x})$	$\boldsymbol{x}$	$p(\boldsymbol{x})$	$L(\boldsymbol{x})$
000	27/125	1	100	18/125	3
001	18/125	3	101	12/125	5
010	18/125	3	110	12/125	5
011	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):

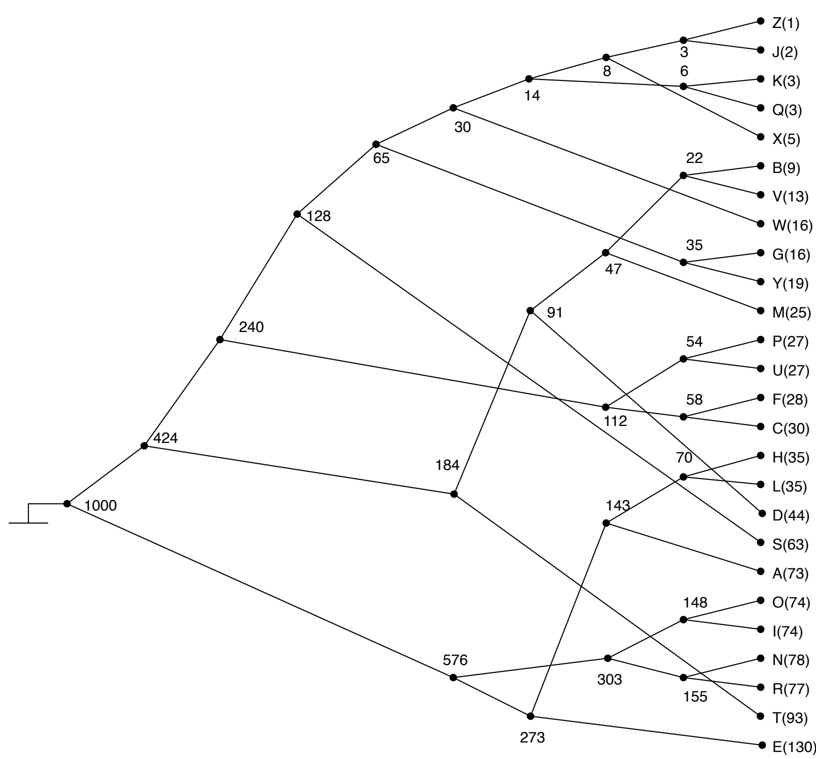


The code table becomes

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).



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

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

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

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

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

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.

X	0	1
Y	0	1

which gives average length $L_1 = 1$.

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

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

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

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

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

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

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

(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

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

Hence, it is a probability function.

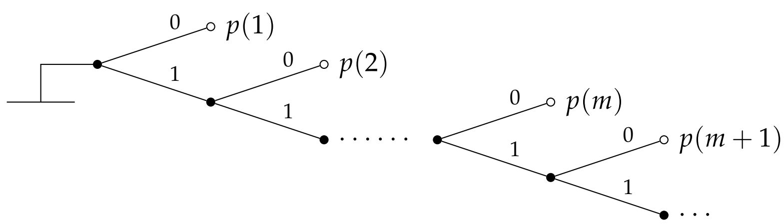
(b) With $k = 2$ we get $P(n) = \big(\frac{1}{2}\big)^n$. By considering the optimal codeword lengths

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

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

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

(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 attemmpt 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

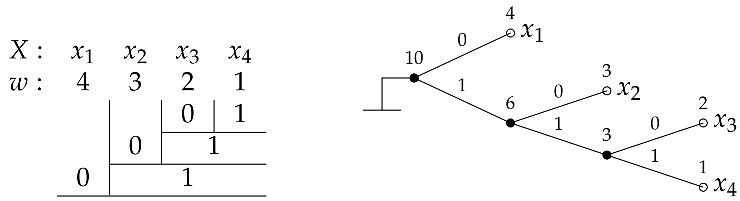
$$\begin{aligned} r(m) &= \sum_{n=m+1}^{\infty}(k-1)\big(\tfrac{1}{k}\big)^n = (k-1)\big(\tfrac{1}{k}\big)^{m+1}\sum_{l=0}^{\infty}\big(\tfrac{1}{k}\big)^l \\ &= (k-1)\big(\tfrac{1}{k}\big)^{m+1}\frac{1}{1-\frac{1}{k}} = \big(\tfrac{1}{k}\big)^m \leq (k-1)\big(\tfrac{1}{k}\big)^m = p(m) \end{aligned}$$

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

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

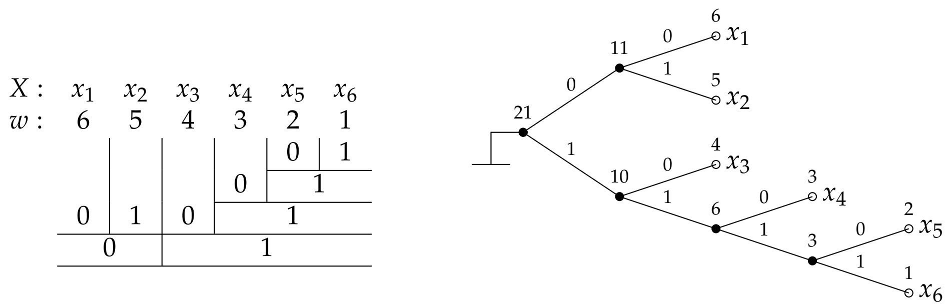
4.11. For simpicity, the common denominator in the probabilities, for each sub-problem, is dropped and the numerator is used as weight in the algorithm.

(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 sepoarates $\{x_1\}$ 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

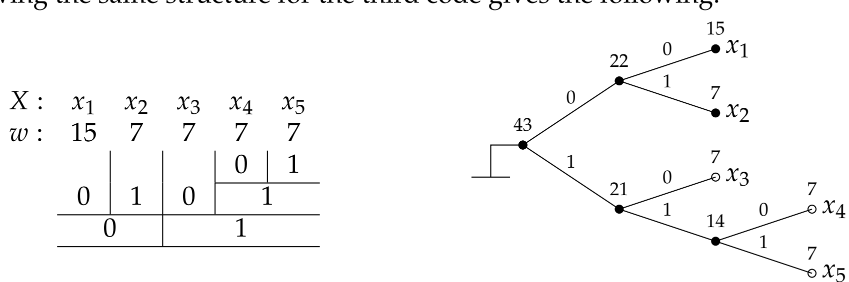


When constructing a Huffman code, first thre leaves x_5 and x_6 are merged, then $\{x_5x_6\}$ and x_4 are merged. After this the nodes in the algorithm are $(x_1, x_2, x_3, \{x_4x_5x_6\})$ with weights $(6, 5, 4, 6)$. So in the next step in the Huffman procedure the nodes x_2 and x_3 are merged. this is not the case in the tree above, and hence the code is not a Huffman code. Continuing the Huffman procedude 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. –

Chapter 5

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

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

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

5.2. Encoding according to

S buffer	B buffer	Codeword
'I screem, you sc'	'reem, we'	(12,6,'w')
'm, you screem, w'	'e all sc'	(6,1,' ')
' you screem, we '	'all scre'	(7,1,'l')
'ou screem, we al'	'l scream'	(1,1,' ')
' screem, we all '	'scream f'	(15,6,' ')
', we all scream '	'for icec'	(0,0,'f')
' we all scream f'	'or icecr'	(0,0,'o')
'we all scream fo'	'r icecre'	(7,1,' ')
' all scream for '	'icecream'	(0,0,'i')
'all scream for i'	'cecream.'	(11,1,'e')
'l scream for ice'	'cream.'	(13,5,'.')

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 course book it should be $\lceil 16+1 \rceil + \lceil 8+1 \rceil + 8 = 5+4+8 = 17$). The uncoded length is $49 \cdot 8 = 392$ bits. Then the compression ratio is $R = 392/315 = 1.24$.

5.3. 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 :	$(0, \sqcup)$	␣
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 :	$(9, \sqcup)$	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.".

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

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

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

5.5. (a)

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

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

(b)

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

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

(c)

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

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

(d)

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

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

5.6.

5.7. **Encoding**

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

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

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

Chapter 6

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

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

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

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

and np as integers),

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

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

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

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

and

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

Dividing these and using the chain rule concludes the proof.

6.3. 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

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

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

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

or, equivalently,

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

Here,

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

and, for the all-zero sequence,

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

Thus, we get

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

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

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

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

Hence, for ϵ-typical sequences

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

So, the number of ϵ-typical sequences is

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

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

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

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

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

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

or, equivalently,

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

Going back to the remaining length we get

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

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

6.5. 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,

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

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

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

Then

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

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

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

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

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

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

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

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

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

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

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

$$H(Y)=H\left((1-A)+\frac{A}{2},\frac{A}{2}\right)=H\left(1-\frac{A}{2},\frac{A}{2}\right)=h(\frac{A}{2})$$
$$H(Y|X)=P(X=0)H(Y|X=0)+P(X=1)H(Y|X=1)=Ah(\frac{1}{2})=A$$

and we conclude

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

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

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

6.10. 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 $C=1-h(\epsilon)=1-h(2p(1-p))$.

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

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

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

Y	Z
A	000
B	001
C	01
D	1

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

6.12. The overall channel has the probabilities

$$P(Z=0|X=0)=(1-\alpha)(1-\beta) \qquad P(Z=1|X=1)=(1-\alpha)(1-\beta)$$
$$P(Z=\Delta|X=0)=(1-\alpha)\beta+\alpha\beta=\beta \qquad P(Z=\Delta|X=0)=\beta$$
$$P(Z=1|X=0)=\alpha(1-\beta) \qquad P(Z=0|X=1)=\alpha(1-\beta)$$

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

6.13. (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 distrubiution 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 $p(y,z) = \frac{1}{2}(p(y,z|x=0) + p(y,z|x=1))$. 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 $1 - h\left(\frac{p^2}{(1-p)^2+p^2}\right)$. 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.

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

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

The conditional and unconditional entropies of Y are then given by

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

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

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

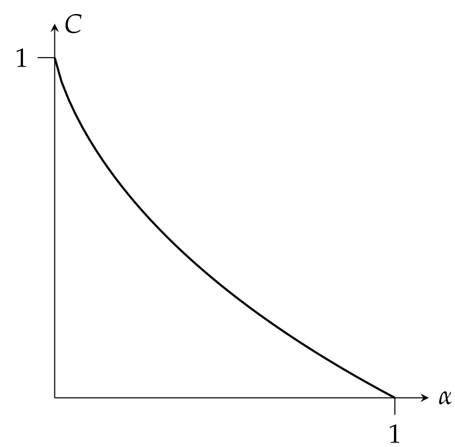
which gives

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

Inserting to the mutual information gives

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

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 here to the right.



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

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

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

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

The corresponding derivative with respect to p is

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

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

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

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

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

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

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

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

Then, for $p = \frac{1}{2}$, we get

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

Hence, the capacity is

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

(b) The right hand inequality is straight forward since

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

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

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

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

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

Then we can identify in the capacity formula above

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

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

u	x
000	000000
001	001110
010	010101
011	011011

u	x
100	100011
101	101101
110	110110
111	111000

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

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

$$(I \quad P)\begin{pmatrix} P \\ I \end{pmatrix} = P \oplus P = 0$$

we get

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

(e) List the most probable error patterns

e	$s = eH^T$
000000	000
100000	011
010000	101
001000	110
000100	100
000010	010
000001	001
100100	111

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

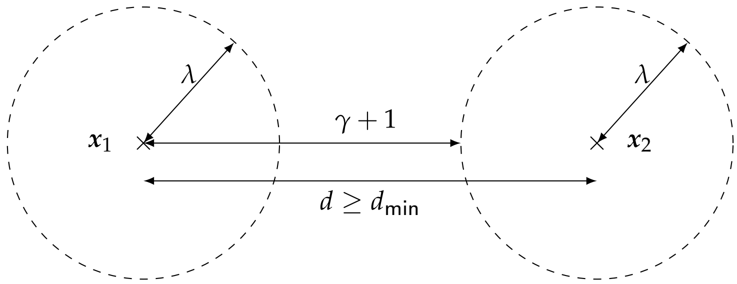
(f) One (correctable) error

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

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 we need to show that the resulting vector has even weight. For this we use the position-wise AND function to get the positions in which both codewords have ones. Then if $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathcal{B}$ the weight of their sum can be written as

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

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

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

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

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

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 $d_E = d+1$.

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

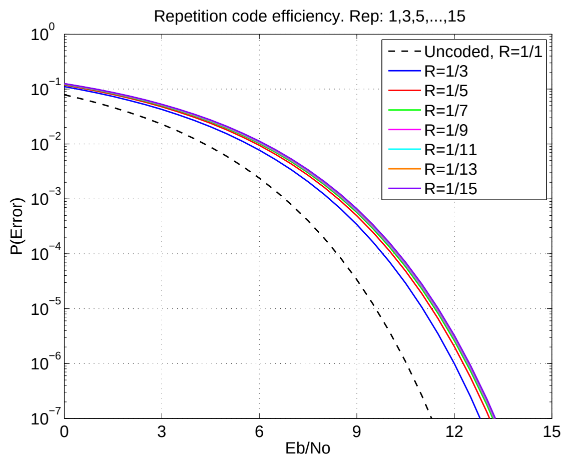
7.4. The error probability when transmitting one bit with energy E_b over a channel with Gaussian noise of level $N_0/2$ is $P_b = Q(\sqrt{2E_b/N_0})$. A repetition code with N repetitions gives the energy E_b/N per transmitted bit, and thus the error probability $P_{b,N} = Q(\sqrt{2E_b/N_0N})$. On the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

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

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

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

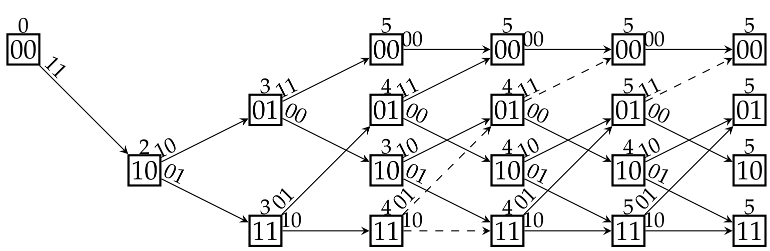
A plot of the results for $N = 3, 5, \ldots, 15$ is shown in the figure below.



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

7.5. —

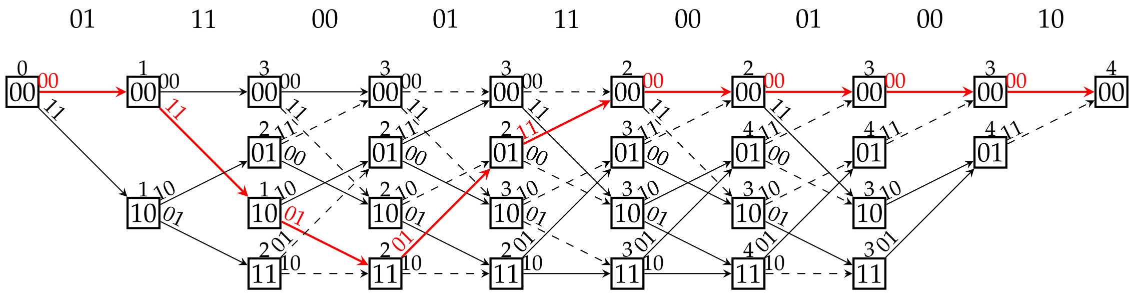
7.6. The free distance is the minimum weight of a non-zero path, starting and ending in the all-zero state. A simple and brute force method to find it is to start in the zero state and give a non-zero input. After this, follow all possible paths, counting the Hamming weight of the paths, until the zero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



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

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

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



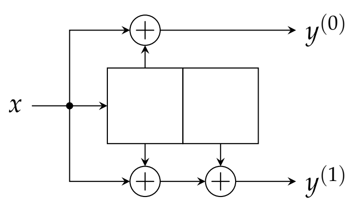
In the figure the red path is found by following the surviving branches from the end node to the start node. This corresponds to the minimum distance path, or the maximum likelihood path. In this case it is

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

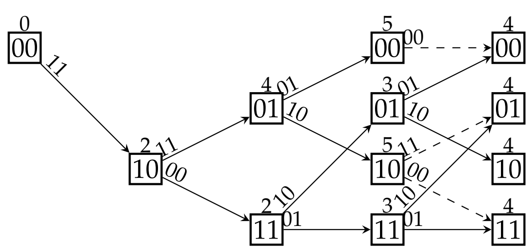
So the answer is that the most likely transmitted information sequence is $\hat{\boldsymbol{u}}_1 = 0110000$.

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

7.7. The encoder circuit for this generator matrix is

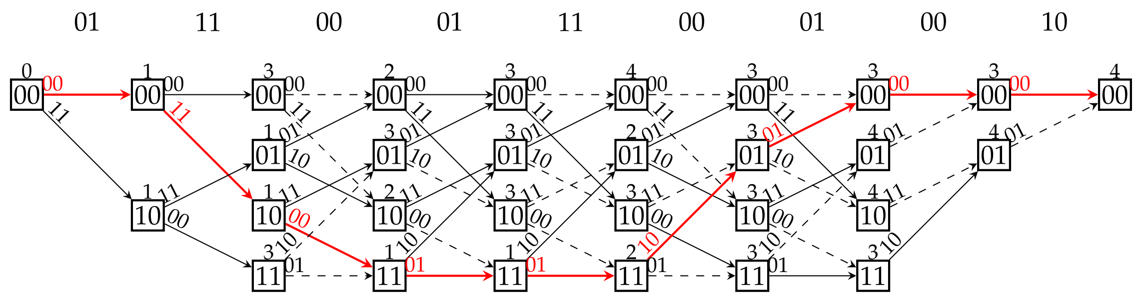


Following the same structure and methods as in Problem 7.6, the free distance is derived from the following trellis.



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

Decoding is done as follows.

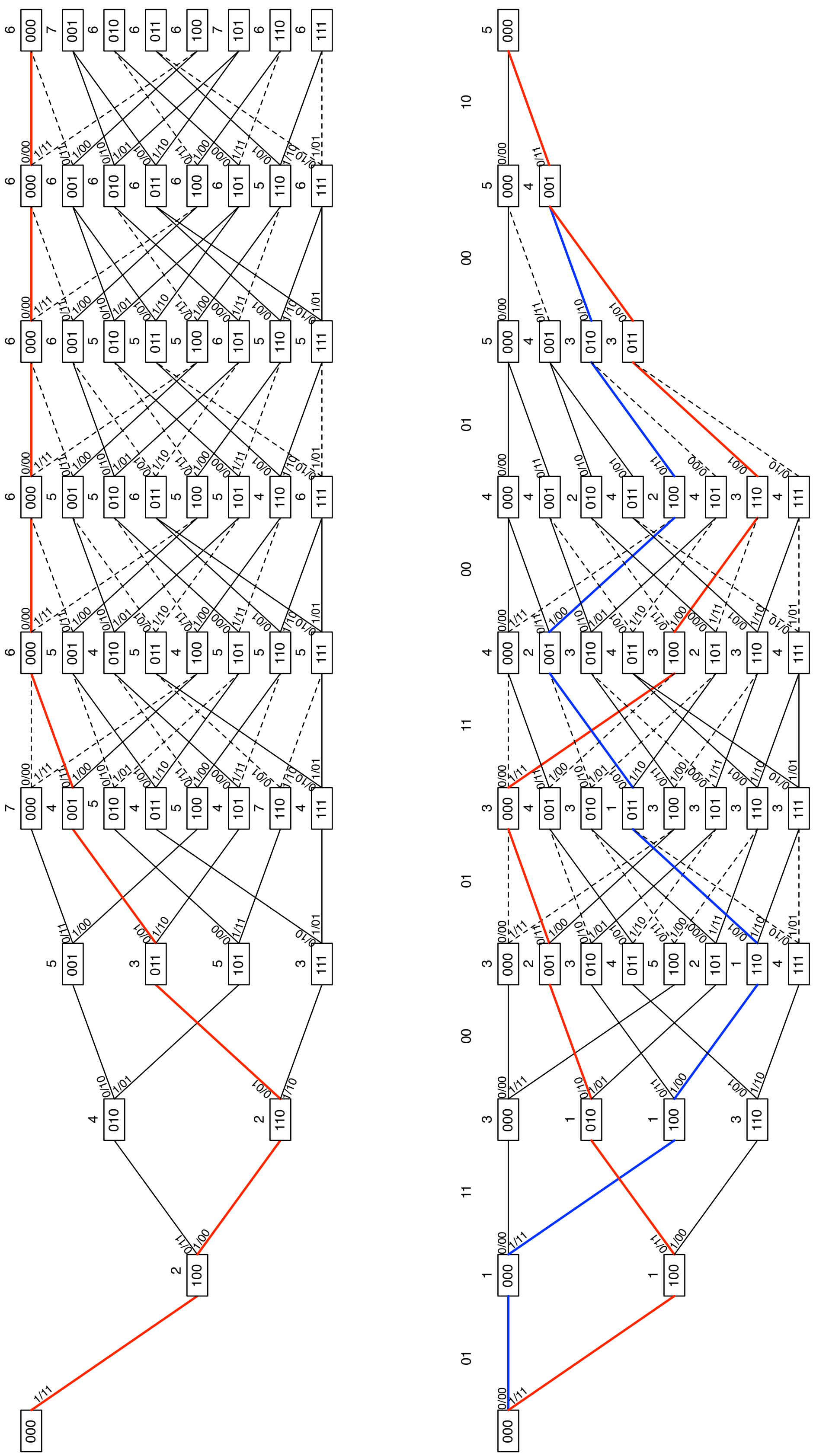


Hence, the most likely code sequence is $\hat{\boldsymbol{v}} = 00\ 11\ 00\ 01\ 01\ 10\ 01\ 00\ 00$ and the corresponding information sequence is $\hat{\boldsymbol{u}} = 0111100$.

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

(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 ($H(X) = -\int f(x) \log f(x)\, dx$) we get that:

(a)

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

(b)

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

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

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

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

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

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

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

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

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

The joint entropy is

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

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

8.3. First get α from

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

(a) The probability is

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

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

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

where

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

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

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

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

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

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

(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. —

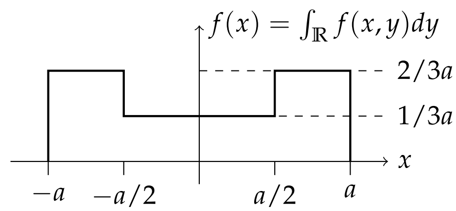
8.6. (a) To simplify notations, let $\mathcal{B}$ denote the shaded region in the figure. Then, since the area of $\mathcal{B}$ is $3ab$, the density function is

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

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

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

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

8.7. Consider

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

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

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

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

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

Chapter 9

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

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

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

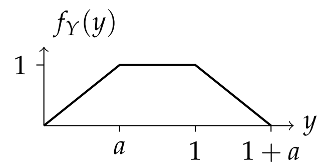
$$C = \frac{1}{2}\log(2\pi e(P-\sigma+\sigma)) - \frac{1}{2}\log(2\pi e(\sigma)) = \frac{1}{2}\log(2\pi eP) - \frac{1}{2}\log(2\pi e\sigma) = \frac{1}{2}\log(\frac{P}{\sigma})$$

9.2. From the problem formulation we know that

$$X \sim U(1)$$
$$Z \sim U(a), \quad 0 < a \le 1$$

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

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



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 $\int y \ln y dy = \frac{y^2}{2}\ln y - \int \frac{y^2}{2}\frac{1}{y}dy = \frac{y^2}{2}\ln y - \frac{y^2}{4}$.

$$H_1 = -\int_0^a \frac{y}{a}\log\frac{y}{a}dy = \frac{1}{a}\log a\int_0^a ydy - \frac{1}{a}\int_0^a y\log ydy = \frac{a}{2}\log a - \frac{a}{2}\log a + \frac{a}{4\ln 2} = \frac{a}{4\ln 2}$$

$$H_2 = -\int_a^1 1\log 1dy = 0$$

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

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

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

9.3. (a) The received power is

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

and the received noise is Gaussian with variance

$$\frac{N}{2} = \frac{N_1|H_1|^2 + N_2}{2}$$

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

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

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

$$\text{SNR}_1 = \frac{|H_1|^2 P_X}{W N_1} = \frac{P_Y}{W N_1}$$
$$\text{SNR}_2 = \frac{|H_2|^2 P_Y}{W N_2} = \frac{P_Z}{W N_2}$$

Then the total SNR can be expressed as

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

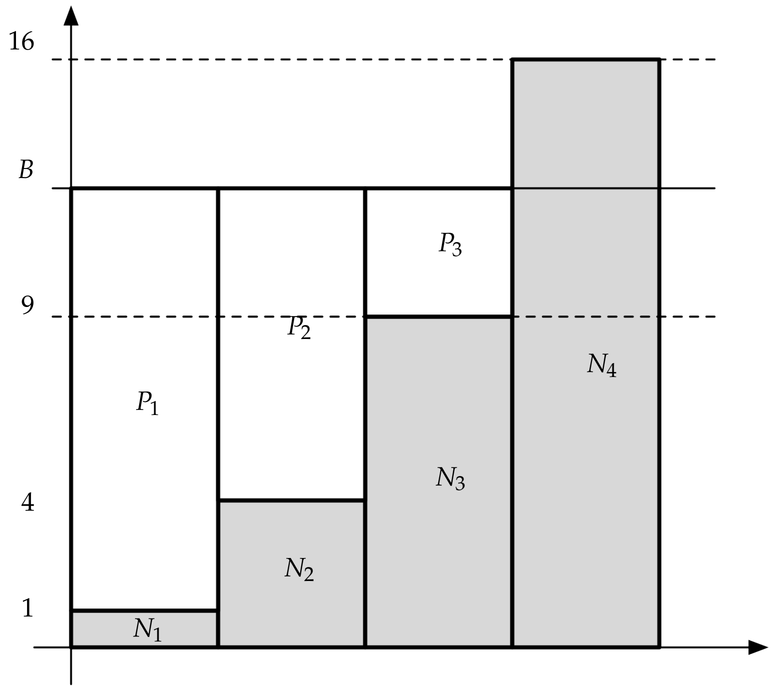
Notice, that by considering the invers of the SNR, the noise to signal ratio, the derivations can be considerably simplified,

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

which is equivalent to the desired result.

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

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



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

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

The capacities in the sub-channels are

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

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

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

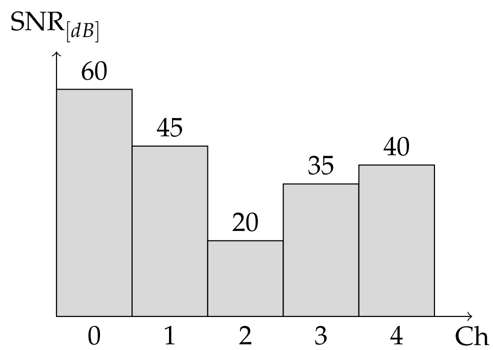
$$\begin{aligned} C &= \sum_i \frac{1}{2} \log\Big(1 + \frac{19/6}{N_i}\Big) \\ &= 0.24 + 0.17 + 0.15 + 0.20 + 0.13 + 0.31 = 1.19 \text{ bit/transmission} \end{aligned}$$

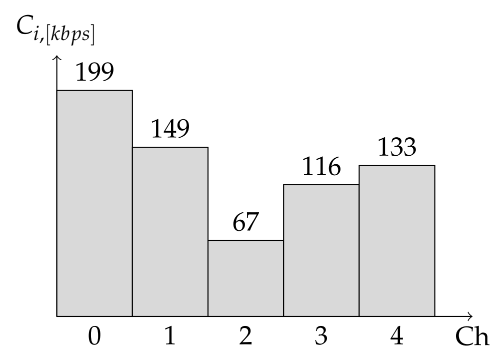
(c) When using only one sub-channel the capacity is maximised if we take the one with least noise, $N = 6$. This gives the capacity $C = \frac{1}{2}\log 2(1 + 19/6) = 1.03$ bit/transmission.

9.6. 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 bandwidthm. 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 $\mathrm{SNR}_i = P_\Delta + |G_i|^2 - N_{0,i}$ and is shown below,

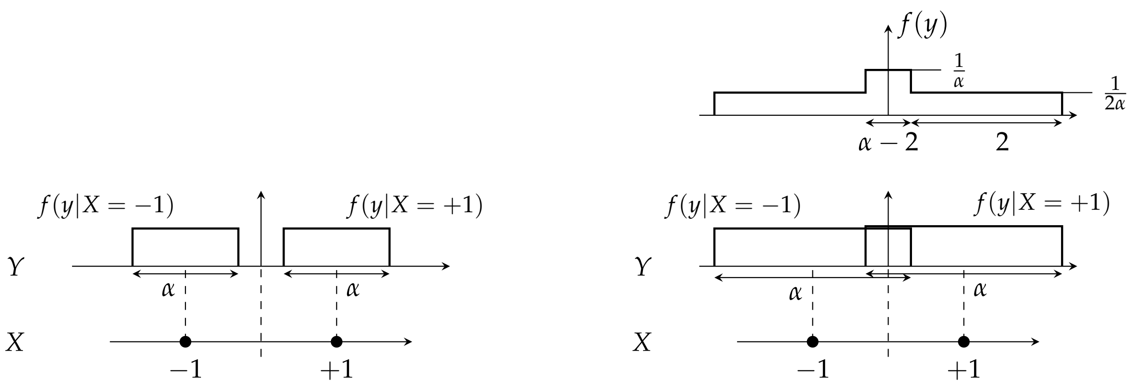




The capacity per sub-channel is derived as $C_i = W\log\big(1 + \mathrm{SNR}\big)$, shown above. The total capaicty is the sum, $C = \sum_i C_i = 199 + 149 + 67 + 116 + 133 = 665$ kbps.

Chapter 10

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



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

The result can also be found from the following derivations:

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

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

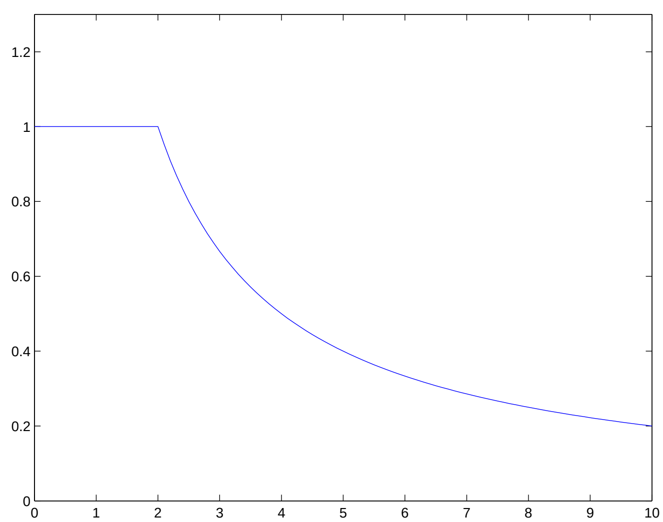
$$
\begin{aligned}
H(Y) &= -\int_{-1-\frac{\alpha}{2}}^{1-\frac{\alpha}{2}} \frac{1}{2\alpha} \log \frac{1}{2\alpha} dx - \int_{1-\frac{\alpha}{2}}^{-1+\frac{\alpha}{2}} \frac{1}{\alpha} \log \frac{1}{\alpha} dx - \int_{-1+\frac{\alpha}{2}}^{1+\frac{\alpha}{2}} \frac{1}{2\alpha} \log \frac{1}{2\alpha}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

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

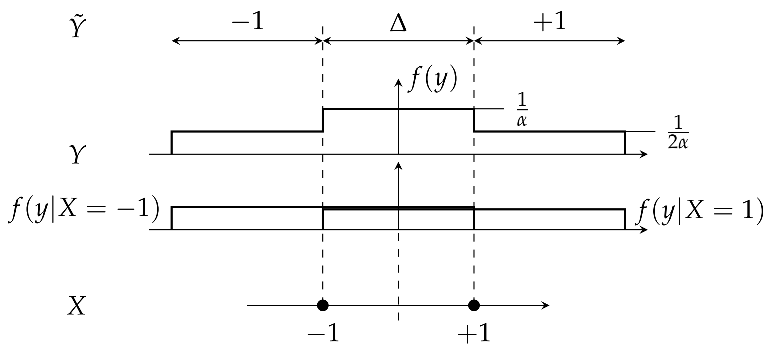
Summarising, the mutual information becomes

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

which is plotted below.



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



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

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

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

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

10.3. The mutual information is $I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(X+Z|X) = H(Y) - H(Z)$, where $H(Z) = \log(1-(-1)) = \log 2$. Since Y ranges from -3 to 3 with uniform weights $p_{-2}/2$ for $-3 \le Y \le -2$, $(p_{-2}+p_{-1})/2$ 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 $H(Y) = \log(3-(-3)) = \log 6$.

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

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

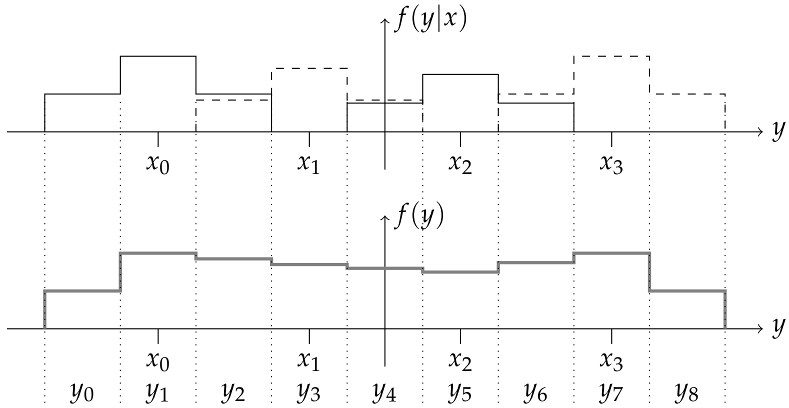
$$f(y) = \begin{cases} \frac{1}{8}, & -3.5 \le y \le 3.5 \\ \frac{1}{16}, & -4.5 \le y < -3.5 \text{ and } 3.5 < y \le 4.5 \\ 0, & \text{o.w.} \end{cases}$$

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 $\{y_0, y_1, \ldots, y_8\}$ corrsponding 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

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

Takinh the derivative equal to zero gives

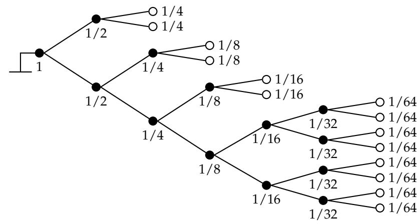
$$\frac{\partial}{\partial p}H(Y) = \frac{3}{4}\Big(2\log(1-2p) - 2\log 2p\Big) + \frac{1}{2} = \frac{3}{2}\log\frac{1-2p}{2p} + \frac{1}{2} = 0$$

or, equivalently,

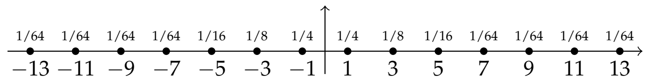
$$p = \frac{1}{2^{2/3} + 2}$$

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 equaly distribution to $E[X^2] = 2 \cdot 3^2 p + 2(\frac{1}{2} - p) \approx 5.46$ for the optiomal distribution.

10.5. (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 becoms

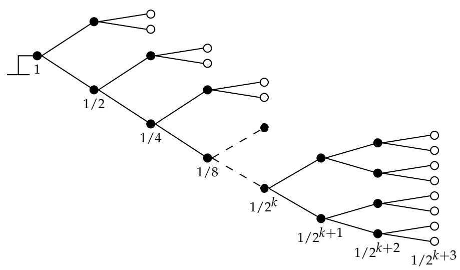


By summing the inner node probabilities, the average length is 3, hence it is comapable in bit rate with the 8-PAM system. If the leaves are mapped to signals in an 14-PAM constellation where the least probable nodes has the highes distance to the center, the following constellation probabilities are 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 therer 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

$$\begin{aligned}
\frac{1}{2}E[X^2] &= \sum_{i=1}^{k}(2i-1)^2\frac{1}{2^{i+1}} + \frac{1}{2^{k+3}}\Big(\underbrace{(2k+1)^2 + (2k+3)^2 + (2k+5)^2 + (2k+7)^2}_{\approx 4\cdot 4k^2 = 2^4k^2,\ k \text{ large}}\Big) \\
&= \underbrace{2\sum_{i=1}^{k} i^2\frac{1}{2^i}}_{\to 12} + \underbrace{2\sum_{i=1}^{k} i\frac{1}{2^i}}_{\to 4} + \underbrace{\frac{1}{2}\sum_{i=1}^{k}\frac{1}{2^i}}_{\to\frac{1}{2}} + \underbrace{\frac{2^2}{2^{k-1}}}_{\to 0} \to \frac{17}{2}, \quad k\to\infty
\end{aligned}$$

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.

10.6. (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

$$\int_{|\boldsymbol{x}|^2 \leq R^2} f(|\boldsymbol{x}|) d\boldsymbol{x} = \frac{2\pi^{N/2}}{\Gamma\left(\frac{N}{2}\right)} \int_0^R x^{N-1} f(x) dx$$

gives

$$\begin{aligned} f(x) = f_X(x) &= \int_{|\tilde{\boldsymbol{x}}|^2 \leq R^2 - x^2} d\tilde{\boldsymbol{x}} = \frac{2\pi^{\frac{N-1}{2}}}{\Gamma(\frac{N-1}{2})} \int_0^{\sqrt{R^2-x^2}} x^{N-2} dx \\ &= \frac{2\pi^{\frac{N-1}{2}}}{\Gamma(\frac{N-1}{2})} \frac{(R^2 - x^2)^{\frac{N-1}{2}}}{N-1} \\ &= \frac{\pi^{\frac{N-1}{2}}}{\Gamma(\frac{N+1}{2})} \left(\frac{\Gamma\left(\frac{N}{2}+1\right)^{2/N}}{\pi} - x^2 \right)^{\frac{N-1}{2}} \end{aligned}$$

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

10.7. (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 reallity 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 linerly decreasing function.)

The received signal to noise ratio for each sub-channel becomes

$$\text{SNR}_i = -70 - (5i + 10) + 140 = 60 - 5i \text{ dB}$$

and the derived capacity per sub-chanel

$$C_i = \Delta f \log\big(1 + 10^{\mathrm{SNR}_i/10}\big) = 10^4 \log\Big(1 + 10^{(60-5i)/10}\Big)$$

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
$\|H_i\|^2$[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-chanels 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

$$\widetilde{\mathrm{SNR}}_i = \mathrm{SNR}_i - \Gamma + \gamma_c = \mathrm{SNR}_i - 6 \text{ dB}$$

The estimated bit rate is

$$R_i = \Delta f \log\big(1 + 10^{\widetilde{\mathrm{SNR}}_i/10}\big)$$

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
$\widetilde{\mathrm{SNR}}_i$[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

$$R = \sum_i R_i = 1080 \text{ [kbps]}$$

Chapter 11

11.1. From the problem we have $P(X = j) = \frac{1}{k}$, $j = 0, 1, \dots, k-1$, 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 betweeen 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 asumptions 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 ther rate distortion function is non-increasing and non-negative, we conclude that

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

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 $\beta < 0$

$$\begin{aligned} 1 &= \int_0^\infty e^{\alpha+\beta x} dx = -\frac{1}{\beta} e^\alpha \\ \frac{1}{\lambda} &= \int_0^\infty x e^{\alpha+\beta x} dx = -\frac{1}{\beta^2} e^\alpha \end{aligned}$$

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 exponetial 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 $\int_0^\infty xg(x)dx = 1/\lambda$. Then,

$$\begin{aligned}
H_g(X) &= -\int_0^\infty g(x)\ln g(x)dx \\
&= -\int_0^\infty g(x)\ln \frac{g(x)}{f(x)} f(x)dx \\
&= -\int_0^\infty g(x)\ln f(x)dx - D(g||f) \\
&\leq -\int_0^\infty g(x)\ln f(x)dx \\
&= \lambda \int_0^\infty xg(x)dx - \ln\lambda \int_0^\infty g(x)dx \\
&= 1 - \ln\lambda = H_f(X)
\end{aligned}$$

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 exponetial distribution maximising the entropy. THus, the bound is fiÃžlfilled 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 dne 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 $0 \leq \delta\lambda \leq 1$, or, equivalently, $0 \leq \delta \leq 1/\lambda$. 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 erformed in different ways, here a trapetsoid method was used. Assuming a set of x-values, $\boldsymbol{x} = x_1, \dots, x_n$, with constant separation $x_i - x_{i-1} = \Delta$. Let $\boldsymbol{y} = y_1, \dots, y_n$ 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 reqonstruction 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 estimetaed by a Huffamncode 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.

11.8.