Exercises with solutions (1)

1. Investigate the relationship between independence and correlation.

   (a) Two random variables X and Y are said to be *correlated* if and only if their covariance C_{XY} is not equal to 0.

   Can two independent random variables X and Y be correlated?

   Solution:

   Without loss of generality, we assume that the statistical properties of the random variables X and Y are given by the joint probability density function $f_{XY}(x,y)$ and marginal probability density functions $f_X(x)$ and $f_Y(y)$. Note that for a discrete random variable X with alphabet $\mathcal{A}$, the pdf $f_X(x)$ can be written using the probability mass function $p_X(a)$ and the Dirac delta function $\delta(x)$,

$$f_X(x) = \sum_{a \in \mathcal{A}} p_X(a) \cdot \delta(x-a).$$

   Similarly, a joint pdf $f_{XY}(x,y)$ can be constructed using the Dirac delta function if either or both random variables X and Y are discrete random variables.

   Two random variables X and Y are independent if and only if the joint pdf is equal to the product of the marginal pdfs, $\forall x, y \in \mathbb{R}$, $f_{XY}(x,y) = f_X(x) f_Y(y)$. For the covariance C_{XY} of two independent random variables X and Y, we then obtain

$$\begin{aligned}
C_{XY} &= E\{(X - E\{X\})\,(Y - E\{Y\})\} \\
&= E\{XY - XE\{Y\} - E\{X\}Y + E\{X\}E\{Y\}\} \\
&= E\{XY\} - E\{X\}E\{Y\} - E\{X\}E\{Y\} + E\{X\}E\{Y\} \\
&= E\{XY\} - E\{X\}E\{Y\} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\,y\,f_{XY}(x,y)\,\mathrm{d}x\,\mathrm{d}y - E\{X\}E\{Y\} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\,y\,f_X(x)\,f_Y(y)\,\mathrm{d}x\,\mathrm{d}y - E\{X\}E\{Y\} \\
&= \int_{-\infty}^{\infty} x\,f_X(x) \left(\int_{-\infty}^{\infty} y\,f_Y(x)\,\mathrm{d}y \right) \mathrm{d}x - E\{X\}E\{Y\} \\
&= \left(\int_{-\infty}^{\infty} x\,f_X(x)\,\mathrm{d}x \right) \left(\int_{-\infty}^{\infty} y\,f_Y(x)\,\mathrm{d}y \right) - E\{X\}E\{Y\} \\
&= E\{X\}E\{Y\} - E\{X\}E\{Y\} \\
&= 0.
\end{aligned}$$

   Two independent random variables are always uncorrelated.

(b) Let X be a continuous random variable with a variance $\sigma_X^2 > 0$ and a pdf $f_X(x)$. The pdf shall be non-zero for all real numbers, $f_X(x) > 0$, $\forall x \in \mathbb{R}$. Furthermore, the pdf $f_X(x)$ shall be symmetric around zero, $f_X(x) = f_X(-x)$, $\forall x \in \mathbb{R}$. Let Y be a random variable given by $Y = a\,X^2 + b\,X + c$ with $a, b, c \in \mathbb{R}$.

For which values of a, b, and c are X and Y uncorrelated?

For which values of a, b, and c are X and Y independent?

Solution:

First we investigate the correlation of the random variables X and Y. Due to the symmetry of the pdf around zero, $f_X(x) = f_X(-x)$, $\forall x \in \mathbb{R}$, the expectation values of the odd integer powers of the random variable X are equal to 0. With the integer variable $n \geq 0$, we have

$$\begin{aligned} E\{X^{2n+1}\} &= \int_{-\infty}^{\infty} x^{2n+1}\, f_X(x)\, \mathrm{d}x \\ &= \int_{0}^{\infty} x^{2n+1}\, f_X(x)\, \mathrm{d}x + \int_{-\infty}^{0} x^{2n+1}\, f_X(x)\, \mathrm{d}x. \end{aligned}$$

Using the substitution $t = -x$ for the second integral, we obtain

$$\begin{aligned} E\{X^{2n+1}\} &= \int_{0}^{\infty} x^{2n+1}\, f_X(x)\, \mathrm{d}x + \int_{\infty}^{0} (-t)^{2n+1}\, f_X(-t)\, (-\mathrm{d}t) \\ &= \int_{0}^{\infty} x^{2n+1}\, f_X(x)\, \mathrm{d}x + \int_{\infty}^{0} t^{2n+1}\, f_X(t)\, \mathrm{d}t \\ &= \int_{0}^{\infty} x^{2n+1}\, f_X(x)\, \mathrm{d}x - \int_{0}^{\infty} t^{2n+1}\, f_X(t)\, \mathrm{d}t \\ &= 0. \end{aligned}$$

In particular, we have $E\{X\} = 0$ and $E\{X^3\} = 0$. For the covariance C_{XY}, we then obtain

$$\begin{aligned} C_{XY} &= E\{XY\} - E\{X\}E\{Y\} \\ &= E\{XY\} \\ &= E\{aX^3 + bX^2 + cX\} \\ &= aE\{X^3\} + bE\{X^2\} + cE\{X\} \\ &= b \cdot \sigma_X^2. \end{aligned}$$

The random variables X and Y are uncorrelated if and only if b is equal to 0.

Now, we investigate the dependence of the random variables X and Y. The random variables X and Y are independent if and only if

$f_{XY}(x,y) = f_X(x)f_Y(y)$. Since $f_{XY}(x,y) = f_{Y|X}(y|x)f_X(x)$, we can also say that X and Y are independent if and only if the marginal pdf for $f_Y(y)$ is equal to the conditional pdf $f_{Y|X}(y|x)$.

The value of the random variable Y is completely determined by the value of the random variable X. Hence, the conditional pdf $f_{Y|X}(y|x)$ is given by the Dirac delta function

$$f_{Y|X}(y|x) = \delta(y - ax^2 - bx - c).$$

If the conditional pdf $f_{Y|X}(y|x)$ depends on the value x of the random variable X, the random variables X and Y are not independent, since $f_Y(y)$ cannot be equal to $f_{Y|X}(y|x)$ in this case. The conditional pdf $f_{Y|X}(y|x)$ does not depend on x if one of the following conditions is fulfilled:

- $a \neq 0$ and $f_X(x) = \frac{w}{2}\delta(x - x_1) + \frac{1-w}{2}\delta(x - x_2)$, where x_1 and x_2 are the roots of the quadratic equation $ax^2 + bx = d$ for any value of $d > -b^2/(4a)$, and $0 \leq w \leq 1$;
- $a = 0$, $b \neq 0$, and $f_X(x) = \delta(x - x_0)$ with x_0 being any constant real value;
- $a = 0$ and $b = 0$.

Since it is given that $f_X(x) > 0$, $\forall x \in \mathbb{R}$, we do not need to consider the first two cases. Hence, for all parameters a, b, and c with $a \neq 0$ or $b \neq 0$, the random variables X and Y are dependent.

For the case $a = 0$ and $b = 0$, the conditional pdf is given by

$$f_{Y|X}(y|x) = \delta(y - c),$$

and the random variable Y is given by $Y = c$. The random variable Y is always equal to c. Consequently, its marginal pdf is given by

$$f_Y(y) = \delta(y - c)$$

and is equal to the conditional pdf $f_{Y|X}(y|x)$.

The random variables X and Y are independent if and only if $a = 0$ and $b = 0$.

(c) Which of the following statements for two random variables X and Y are true?

i. If X and Y are uncorrelated, they are also independent.
ii. If X and Y are independent, $E\{XY\} = 0$.
iii. If X and Y are correlated, they are also dependent.

Solution:

i. The statement "if X and Y are uncorrelated, they are also independent" is wrong. As a counterexample consider the random variables X and Y in problem (1b) for $a \neq 0$ and $b = 0$. In this case, the random variables are uncorrelated, but are dependent.

ii. The statement “if X and Y are independent, $E\{XY\} = 0$” is wrong. As shown in problem (1a), the independence of X and Y implies $C_{XY} = E\{XY\} - E\{X\}E\{Y\} = 0$. If, however, both mean values $E\{X\}$ and $E\{Y\}$ are not equal to zero, then $E\{XY\}$ is also not equal to zero.

iii. The statement “if X and Y are correlated, they are also dependent” is true. This statement is the contraposition of the statement “if X and Y are independent, they are also uncorrelated”, which has been proved in problem (1a).

2. A fair coin is tossed an infinite number of times. Let Y_n be a random variable, with $n \in \mathbb{Z}$, that describes the outcome of the n-th coin toss. If the outcome of the n-th coin toss is head, Y_n is equal to 1; if it is tail, Y_n is equal to 0. Now consider the random process $\mathbf{X} = \{X_n\}$. The random variables X_n are determined by $X_n = Y_n + Y_{n-1}$, and thus describe the total number of heads in the n-th and $(n-1)$-th coin tosses.

   (a) Determine the marginal pmf $p_{X_n}(x_n)$ and the marginal entropy $H(X_n)$. Is it possible to design a uniquely decodable code with one codeword per possible outcome of X_n that has an average codeword length equal to the marginal entropy?

   Solution:

   Since we consider a fair coin, both possible outcomes (head and tail) of a single coin toss are equally likely. Hence, the pmf for the random variables Y_n is given by $p_{Y_n}(0) = P(Y_n = 0) = \frac{1}{2}$ and $p_{Y_n}(1) = P(Y_n = 1) = \frac{1}{2}$. The random variables Y_n and Y_m with $n \neq m$ are independent. Furthermore, two different k-symbol sequences of heads and tails "$Y_n Y_{n-1} \cdots Y_{n-k+1}$" are mutually exclusive events. The alphabet $\mathcal{A}$ for the random variables X_n consists of three possible outcomes $\mathcal{A} = \{0, 1, 2\}$. Hence, the marginal pmf can be obtained as follows:

$$\begin{aligned}
p_{X_n}(0) &= P(X_n = 0) \\
&= P(\text{“}Y_n Y_{n-1}\text{”} = \text{“00”}) \\
&= p_{Y_n}(0) \cdot p_{Y_n}(0) \\
&= \frac{1}{2} \cdot \frac{1}{2} \\
&= \frac{1}{4},
\end{aligned}$$

$$\begin{aligned}
p_{X_n}(1) &= P(X_n = 1) \\
&= P(\{\text{“}Y_n Y_{n-1}\text{”} = \text{“01”}\} \cup \{\text{“}Y_n Y_{n-1}\text{”} = \text{“10”}\}) \\
&= P(\text{“}Y_n Y_{n-1}\text{”} = \text{“01”}) + P(\text{“}Y_n Y_{n-1}\text{”} = \text{“10”}) \\
&= p_{Y_n}(0) \cdot p_{Y_n}(1) + p_{Y_n}(1) \cdot p_{Y_n}(0) \\
&= \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \\
&= \frac{1}{2},
\end{aligned}$$

$$\begin{aligned}
p_{X_n}(2) &= P(X_n = 2) \\
&= P(\text{“}Y_n Y_{n-1}\text{”} = \text{“11”}) \\
&= p_{Y_n}(1) \cdot p_{Y_n}(1) \\
&= \frac{1}{2} \cdot \frac{1}{2} \\
&= \frac{1}{4}.
\end{aligned}$$

The marginal entropy $H(X_n)$ is given by

$$\begin{aligned} H(X_n) &= -\sum_{x_n \in \mathcal{A}} p_{X_n}(x_n)\ \log_2 p_{X_n}(x_n) \\ &= -p_{X_n}(0) \log_2 p_{X_n}(0) - p_{X_n}(1) \log_2 p_{X_n}(1) \\ &\quad -p_{X_n}(2) \log_2 p_{X_n}(2) \\ &= -\frac{1}{4} \log_2 \left(\frac{1}{4}\right) - \frac{1}{2} \log_2 \left(\frac{1}{2}\right) - \frac{1}{4} \log_2 \left(\frac{1}{4}\right) \\ &= 2 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} \\ &= \frac{3}{2}. \end{aligned}$$

Since all marginal probabilities are integer powers of 2, it is possible to develop a Huffman code for which the average codeword length is equal to the marginal entropy.

An example for such a code is given in the table below.

x_n	$p_{X_n}(x_n)$	codeword	$\ell(x_n)$
0	0.25	10	2
1	0.50	0	1
2	0.25	11	2

The average codeword length is

$$\bar{\ell} = \sum_{x_n \in \mathcal{A}} p_{X_n}(x_n) \cdot \ell(x_n) = \frac{1}{4} \cdot 2 + \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 = \frac{3}{2},$$

and is thus equal to the marginal entropy $H(X_n)$.

(b) Determine the conditional pmf $p_{X_n|X_{n-1}}(x_n|x_{n-1})$ and the conditional entropy $H(X_n|X_{n-1})$.

Design a conditional Huffman code.

What is the average codeword length of the conditional Huffman code?

Solution:

The conditional pmf $p_{X_n|X_{n-1}}(x_n|x_{n-1})$ can be calculated using the relationship

$$p_{X_n|X_{n-1}}(x_n|x_{n-1}) = \frac{p_{X_n X_{n-1}}(x_n, x_{n-1})}{p_{X_n}(x_{n-1})}.$$

The probability masses of the marginal pmf $p_{X_n}(x_n)$ have been calculated in (2a). The joint probability masses $p_{X_n X_{n-1}}(x_n, x_{n-1})$ can be calculated in a similar way.

As an example, consider the joint probability mass

$$\begin{aligned}
& p_{X_n X_{n-1}}(1,1) \\
&\quad = P(X_n = 1, X_{n-1} = 1) \\
&\quad = P(\text{“}Y_n Y_{n-1} Y_{n-2}\text{”} = \text{“010”}) + P(\text{“}Y_n Y_{n-1} Y_{n-2}\text{”} = \text{“101”}) \\
&\quad = \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^3 \\
&\quad = \frac{1}{4}.
\end{aligned}$$

Note that for some combinations of x_n and x_{n-1}, the joint probability masses $p_{X_n X_{n-1}}(x_n, x_{n-1})$ are equal to zero, since the corresponding event $\{X_n = x_n \cap X_{n-1} = x_{n-1}\}$ cannot occur. If $X_{n-1} = 0$, i.e., if the result of the $(n-1)$-th and the $(n-2)$-th coin toss is tail, the random variable X_n can only take the values 0 or 1. Similarly, if $X_{n-1} = 2$, X_n can only take the values 1 or 2. Consequently, the joint probability masses $p_{X_n X_{n-1}}(2,0)$ and $p_{X_n X_{n-1}}(0,2)$ are equal to 0. The following table shows that probability masses of the joint pmf $p_{X_n X_{n-1}}(x_n, x_{n-1})$ and the conditional pmf $p_{X_n|X_{n-1}}(x_n|x_{n-1})$.

x_{n-1}	x_n	$p_{X_n}(x_{n-1})$	$p_{X_n X_{n-1}}(x_n, x_{n-1})$	$p_{X_n\|X_{n-1}}(x_n\|x_{n-1})$
0	0	0.25	0.125	0.50
	1		0.125	0.50
	2		0.000	0.00
1	0	0.50	0.125	0.25
	1		0.250	0.50
	2		0.125	0.25
2	0	0.25	0.000	0.00
	1		0.125	0.50
	2		0.125	0.50

The conditional entropy $H(X_n|X_{n-1})$ is given by

$$H(X_n|X_{n-1}) = -\sum_{\substack{x_n \in \mathcal{A} \\ x_{n-1} \in \mathcal{A}}} p_{X_n X_{n-1}}(x_n, x_{n-1}) \, \log_2 p_{X_n|X_{n-1}}(x_n|x_{n-1}).$$

Some of the joint probability masses are equal to 0. These terms can be simply excluded from the summation, as can be shown by considering the following limit, where p denotes the joint probability $p_{X_n X_{n-1}}(x_n, x_{n-1})$ and q denotes the marginal probability $p_{X_n}(x_{n-1})$, which is always greater than 0,

$$\lim_{p \to 0} -p \, \log_2 \left(\frac{p}{q}\right) \quad \text{with} \quad q > 0.$$

By applying L'Hôpital's rule, we obtain

$$\lim_{p \to 0} -p \, \log_2 \left(\frac{p}{q}\right) \; = \; \lim_{p \to 0} \, -\frac{\log_2 \left(\frac{p}{q}\right)'}{\left(\frac{1}{p}\right)'} = \lim_{p \to 0} \, -\frac{\frac{1}{\ln 2} \cdot \frac{q}{p} \cdot \frac{1}{q}}{-\frac{1}{p^2}}$$

$$
\begin{aligned}
&= \lim_{p \to 0} \frac{1}{\ln 2} \cdot p = \frac{1}{\ln 2} \lim_{p \to 0} p \\
&= 0.
\end{aligned}
$$

Inserting the values of the joint and conditional pmf, which are given in the table above, into the expression for the conditional entropy yields

$$
\begin{aligned}
H(X_n|X_{n-1}) &= -4 \cdot \frac{1}{8} \cdot \log_2\left(\frac{1}{2}\right) - 2 \cdot \frac{1}{8} \cdot \log_2\left(\frac{1}{4}\right) \\
&\quad -1 \cdot \frac{1}{4} \cdot \log_2\left(\frac{1}{2}\right) \\
&= 4 \cdot \frac{1}{8} \cdot 1 + 2 \cdot \frac{1}{8} \cdot 2 + 1 \cdot \frac{1}{4} \cdot 1 \\
&= \frac{1}{2} + \frac{1}{2} + \frac{1}{4} \\
&= \frac{5}{4}.
\end{aligned}
$$

An example for a conditional Huffman code is shown in the following table. Note that we do not assign a codeword to the impossible events $\{X_n = 0 \cap X_{n-1} = 2\}$ and $\{X_n = 2 \cap X_{n-1} = 0\}$.

x_{n-1}	x_n	$p_{X_n\|X_{n-1}}(x_n\|x_{n-1})$	codeword	$\ell(x_n\|x_{n-1})$
	0	0.50	0	1
0	1	0.50	1	1
	2	0.00	-	0
	0	0.25	10	2
1	1	0.50	0	1
	2	0.25	11	2
	0	0.00	-	0
2	1	0.50	0	1
	2	0.50	1	1

The average codeword length of the conditional Huffman code is given by

$$
\begin{aligned}
\bar{\ell} &= -\sum_{\substack{x_n \in \mathcal{A} \\ x_{n-1} \in \mathcal{A}}} p_{X_n X_{n-1}}(x_n, x_{n-1})\, \ell(x_n|x_{n-1}) \\
&= 4 \cdot \frac{1}{8} \cdot 1 + 2 \cdot \frac{1}{8} \cdot 0 + 2 \cdot \frac{1}{8} \cdot 2 + 1 \cdot \frac{1}{4} \cdot 1 \\
&= \frac{5}{4}.
\end{aligned}
$$

The average codeword length of the conditional Huffman code is equal to the conditional entropy $H(X_n|X_{n-1})$.

(c) Is the random process $\mathbf{X}$ a Markov process?

Solution:

The characteristic property of a Markov process is that the future states of the process depend only on the present state, not on the sequence of events that precede it. Using the conditional pmfs, this property can be written as

$$p_{X_n|X_{n-1}X_{n-2}\cdots}(x_n|x_{n-1}, x_{n-2}, \cdots) = p_{X_n|X_{n-1}}(x_n|x_{n-1}).$$

Now, let us investigate the given process $\mathbf{X}$. If $X_{n-1} = 0$, we know that the result of the $(n-1)$-th and $(n-2)$-th coin tosses was tail. Hence, the random variable X_n can only take the values 0 (for $Y_n = 0$) or 1 (for $Y_n = 1$); both with the probability of $\frac{1}{2}$. By considering additional random variables X_{n-k} with $k > 1$, we cannot improve the knowledge about X_n. We have

$$\begin{aligned} & p_{X_n|X_{n-1}X_{n-2}\cdots}(x_n|0, x_{n-2}, \cdots) \\ &= p_{X_n|X_{n-1}}(x_n|0) = \begin{cases} 0.5 & : \; x_n = 0 \\ 0.5 & : \; x_n = 1 \\ 0.0 & : \; x_n = 2 \end{cases}. \end{aligned}$$

Similarly, if $X_{n-1} = 2$, we know that the result of the $(n-1)$-th and $(n-2)$-th coin tosses was head. Hence, the random variable X_n can only take the values 1 (for $Y_n = 0$) or 2 (for $Y_n = 1$); both with the probability of $\frac{1}{2}$. By considering additional random variables X_{n-k} with $k > 1$, we cannot improve the knowledge about X_n. We have

$$\begin{aligned} & p_{X_n|X_{n-1}X_{n-2}\cdots}(x_n|2, x_{n-2}, \cdots) \\ &= p_{X_n|X_{n-1}}(x_n|2) = \begin{cases} 0.0 & : \; x_n = 0 \\ 0.5 & : \; x_n = 1 \\ 0.5 & : \; x_n = 2 \end{cases}. \end{aligned}$$

However, for $X_{n-1} = 1$, the situation is different. Here, we do not know the exact sequence "$Y_{n-1}Y_{n-2}$", we only know that it was either "01" or "10". By considering an additional random variable X_{n-2}, we can improve our knowledge about X_n. If, for example, $X_{n-2} = 0$, we know that the sequence "$Y_{n-1}Y_{n-2}Y_{n-3}$" was equal to "100", and then the random variable X_n can only take the values 1 or 2, both with a probability of $\frac{1}{2}$.

For an analytic proof that $\mathbf{X}$ is not a Markov process, we consider the conditional probabilities $p_{X_n|X_{n-1}}(0|1)$ and $p_{X_n|X_{n-1}X_{n-2}}(0|1,2)$. In problem (2b), we calculated the conditional pmf $p_{X_n|X_{n-1}}(x_n|x_{n-1})$ and obtained

$$p_{X_n|X_{n-1}}(0|1) = \frac{1}{4}.$$

The probability mass $p_{X_n|X_{n-1}X_{n-2}}(0|1,2)$ is given by

$$\begin{aligned} p_{X_n|X_{n-1}X_{n-2}}(0|1,2) &= \frac{p_{X_nX_{n-1}X_{n-2}}(0,1,2)}{p_{X_nX_{n-1}}(1,2)} \\ &= \frac{P(\text{“}Y_nY_{n-1}Y_{n-2}Y_{n-3}\text{”} = \text{“0011”})}{P(\text{“}Y_nY_{n-1}Y_{n-2}\text{”} = \text{“011”})} \end{aligned}$$

$$
\begin{aligned}
&= \frac{\left(\frac{1}{2}\right)^4}{\left(\frac{1}{2}\right)^3} \\
&= \frac{1}{2}.
\end{aligned}
$$

Hence, we have

$$p_{X_n|X_{n-1}X_{n-2}}(0|1,2) \neq p_{X_n|X_{n-1}}(0|1).$$

The process $\mathbf{X}$ is not a Markov process.

(d) Derive a general formula for the N-th order block entropy $H_N = H(X_n, \cdots, X_{n-N+1})$.
How many symbols have to be coded jointly at minimum for obtaining a code that is more efficient than the conditional Huffman code developed in (2b)?

Solution:
For the following derivation, let $p_N(x_0, \cdots, x_{N-1})$ denote the N-th order joint pmf $p_{X_n \cdots X_{n-N+1}}(x_n, \cdots, x_{n-N+1})$. The N-th order block entropy is given by

$$
\begin{aligned}
H_N &= H(X_n, \cdots, X_{n-N+1}) \\
&= -\sum_{x_0 \in \mathcal{A}} \cdots \sum_{x_{N-1} \in \mathcal{A}} p_N(x_0, \cdots, x_{N-1}) \log_2 p_N(x_0, \cdots, x_{N-1}).
\end{aligned}
$$

The summation in the above equation is done over 3^N terms. Each N-symbol sequence "$x_0 \cdots x_{N-1}$" can be represented by a number of $(N+1)$-symbol sequences "$y_0 \cdots y_N$", where y_n represents a possible value of the random variable Y_n.
There are 2^{N+1} possible $(N+1)$-symbol sequences "$y_0 \cdots y_N$". We have to differentiate the following three cases:

- All symbols of the symbol sequence "$x_0 \cdots x_{N-1}$" are equal to 1, $x_n = 1$, $\forall n \in [0, N-1]$. In this case, the N-symbol sequence "$x_0 \cdots x_{N-1}$" can be obtained by exactly two $(N+1)$-symbol sequences "$y_0 \cdots y_N$", namely "$0101\cdots$" and "$1010\cdots$". Consequently, the joint probability mass $p_N(x_0, \cdots, x_{N-1})$ is equal to
$$p_{N2} = \left(\frac{1}{2}\right)^{N+1} + \left(\frac{1}{2}\right)^{N+1} = 2^{-N}.$$
- The symbol sequence "$x_0 \cdots x_{N-1}$" is possible and contains at least one "0" or one "2". In this case, the N-symbol sequence "$x_0 \cdots x_{N-1}$" is obtained by exactly one $(N+1)$-symbol sequence "$y_0 \cdots y_N$" and the joint probability mass $p_N(x_0, \cdots, x_{N-1})$ is equal to
$$p_{N1} = \left(\frac{1}{2}\right)^{N+1} = 2^{-(N+1)}.$$
Since there are 2^{N+1} outcomes of tossing a coin $N+1$ times, exactly $2^{N+1} - 2$ probability masses (number of possible outcomes minus the two outcomes considered above) are equal to p_{N1}.

- The symbol sequence “$x_0 \cdots x_{N-1}$” is impossible. This is, for example, the case if the symbol sequence contains the sub-sequences “02”, “20”, “010”, or “212”, which cannot be represented as an outcome of the coin tossing experiment. The joint probability mass $p_N(x_0, \cdots, x_{N-1})$ for the impossible symbol sequences is, of course, equal to
$$p_{N0} = 0.$$
The number of impossible N-symbol sequences “$x_0 \cdots x_{N-1}$” is equal to the number of total symbol sequences (which is 3^N) minus the number of symbol sequences for which all symbols are equal to 1 (which is 1) minus the number of symbol sequences that correspond to exactly one outcome of $N+1$ coin tosses (which is $2^{N+1}-2$). Hence, there are $3^N - 2^{N+1} + 1$ impossible N-symbol sequences “$x_0 \cdots x_{N-1}$”.

For problem (2b), we have shown that

$$\lim_{p \to 0} -p \, \log_2 p = 0.$$

Hence, we do not need to consider the impossible N-symbol sequences, with the probability masses equal to 0, for calculating the N-th order block entropy. Consequently, we obtain

$$\begin{aligned}
H_N &= -1 \cdot p_{N2} \, \log_2 p_{N2} - (2^{N+1}-2) \cdot p_{N1} \, \log_2 p_{N1} \\
&= -1 \cdot 2^{-N} \, \log_2(2^{-N}) - (2^{N+1}-2) \cdot 2^{-(N+1)} \, \log_2(2^{-(N+1)}) \\
&= N \cdot 2^{-N} + (N+1)(1-2^{-N}) \\
&= N \cdot 2^{-N} + (N+1) - N \cdot 2^{-N} - 2^{-N} \\
&= (N+1) - 2^{-N}.
\end{aligned}$$

Since all joint probability masses are either equal to 0 or negative integer powers of 2, we can always construct a Huffman code with an average codeword length per N-symbol sequence equal to the N-th order block entropy.

Such an N-th order block Huffman code is more efficient than the conditional Huffman code, if its average codeword length per symbol $\bar{\ell}_N$ is less than the average codeword length per symbol $\bar{\ell}_C$ for the conditional Huffman code. Hence, we want to find the number of symbols N so that

$$\bar{\ell}_N = \frac{H_N}{N} < \bar{\ell}_C = \frac{5}{4}.$$

By inserting the expression for H_N, we obtain

$$\begin{aligned}
\frac{N+1-2^{-N}}{N} &< \frac{5}{4} \\
4N + 4 - 4 \cdot 2^{-N} &< 5N \\
N &> 4 - 2^{-(N-2)}.
\end{aligned}$$

We can manually check that the above inequality is not fulfilled for the case $N = 1$ (1 is not greater than 2). For $N > 1$, the term

$2^{-(N-2)}$ is always greater than 0 and less then or equal to 1. Since N is an integer number, we can then write

$$N \geq 4.$$

At minimum, we have to code 4 symbols jointly for obtaining a code that is more efficient than the conditional Huffman code developed in (2b).

The following table lists the N-th order block entropy H_N and the average codeword length per symbol (assuming a redundancy of zero) for the block codes with N equal to 1, 2, 3, 4, and 5.

N	H_N	H_N/N
1	$3/2$	$3/2 = 1.5$
2	$11/4$	$11/8 = 1.375$
3	$31/8$	$31/24 = 1.291\bar{6}$
4	$79/16$	$79/64 = 1.234375$
5	$191/32$	$191/160 = 1.19375$

The data in the table additionally show that a joint coding of 4 or more symbols yields an average codeword length per symbol (assuming a redundancy of zero, which can be achieved with a block Huffman code, since all probability masses are integer powers of 2) that is less than the average codeword length of 1.25 for the conditional Huffman code developed in (2b).

(e) Calculate the entropy rate $\bar{H}(\mathbf{X})$ of the random process $\mathbf{X}$.
Is it possible to design a variable length code with finite complexity and an average codeword length equal to the entropy rate? If yes, what requirement has to be fulfilled?

<u>Solution:</u>

The entropy rate $\bar{H}(\mathbf{X})$ is defined by

$$\bar{H}(\mathbf{X}) = \lim_{N\to\infty} \frac{H(X_n, \cdots, X_{n-N+1})}{N} = \lim_{N\to\infty} \frac{H_N}{N}.$$

By inserting the expression for the N-th order block entropy, which we have derived in (2d), we obtain

$$\begin{aligned} \bar{H}(\mathbf{X}) &= \lim_{N\to\infty} \frac{H_N}{N} \\ &= \lim_{N\to\infty} \frac{N+1-2^{-N}}{N} \\ &= \lim_{N\to\infty} \frac{N}{N} + \lim_{N\to\infty} \frac{1}{N} + \lim_{N\to\infty} \frac{1}{N \cdot 2^N} \\ &= 1+0+0 \\ &= 1. \end{aligned}$$

The entropy rate $\bar{H}(\mathbf{X})$ for the random process $\mathbf{X} = \{X_n\}$ is equal to 1 bit per symbol. It should be noted that the entropy rate $\bar{H}(\mathbf{X})$

for the random process $\mathbf{X} = \{X_n\}$ is equal to the entropy rate $\bar{H}(\mathbf{Y})$ and the marginal entropy $H(Y_n)$ of the iid process $\mathbf{Y} = \{Y_n\}$.

We first consider the joint coding of N symbols. The average codeword length per symbol is given by

$$\bar{\ell}_N = \frac{H_N}{N}.$$

By using the expression for H_N that we derived in (2d), we obtain

$$\begin{aligned} \bar{\ell}_N &= \frac{H_N}{N} = \frac{N + 1 - 2^{-N}}{N} \\ &= 1 + \frac{1 - 2^{-N}}{N} \\ &> 1. \end{aligned}$$

By coding a finite number N of symbols jointly, we cannot develop a code with an average codeword length per symbol that is equal to the entropy rate.

Similarly, we cannot achieve the entropy rate by considering a finite number N of previously coded symbols for a conditional code. If we consider the N previously coded symbols x_{n-1} to x_{n-N}, inclusive, we always have to consider the case that all these symbols are equal to 1. If all considered previously coded symbols are equal to 1, there are always two possibilities for the sequence of the corresponding random variables "$Y_{n-1} \cdots Y_{n-N-1}$", namely "$1010\cdots$" and "$0101\cdots$". For this condition, the pmf is equal to $\{\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\}$ and, thus, the average codeword length is equal to $\frac{3}{2}$. For all other possible conditions, the pmf is equal to $\{\frac{1}{2}, \frac{1}{2}, 0\}$ or $\{0, \frac{1}{2}, \frac{1}{2}\}$, and the average codeword length is equal to 1. But since the probability for the condition that all N previously coded symbols are equal to 1 is greater than 0, the average codeword length for the entire conditional code is always greater than 1.

By only observing the random variables X_n, it is not possible to construct a code that achieves the entropy rate. The general problem is that when considering a finite number N of symbols, all symbols can be equal to 1, and in this case we cannot know whether the outcome of the corresponding sequence of coin tosses is "head, tail, head, tail, $\cdots$" or "tail, head, tail, head, $\cdots$".

If, however, we do not only know the values of the random variables X_n at the encoder side, but also the values of the random variables Y_n, we can construct a simple code that achieves the entropy rate. We do not transmit the values of X_n, but the values of Y_n using the simple code in the table below.

y_n	$p_{Y_n}(y_n)$	codeword
0	1/2	0
1	1/2	1

At the decoder side, the values x_n of the random variables X_n are obtained based on the transmitted values y_n of the random variables Y_n by $x_n = y_n + y_{n-1}$. The average codeword length for this code is

$$\bar{\ell} = \sum_{y_n \in \{0,1\}} p_{Y_n}(y_n) \cdot \ell(y_n) = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 = 1.$$

It is equal to the entropy rate $\bar{H}(\mathbf{X})$ of the random process $\mathbf{X} = \{X_n\}$ and the entropy rate $\bar{H}(\mathbf{Y})$ of the random process $\mathbf{Y} = \{Y_n\}$.

3. Given is a discrete iid process $\mathbf{X}$ with the alphabet $\mathcal{A} = \{a, b, c, d, e, f, g\}$. The pmf $p_X(x)$ and 6 example codes are listed in the following table.

x	$p_X(x)$	A	B	C	D	E	F
a	1/3	1	0	00	01	000	1
b	1/9	0001	10	010	101	001	100
c	1/27	000000	110	0110	111	010	100000
d	1/27	00001	1110	0111	010	100	10000
e	1/27	000001	11110	100	110	111	000000
f	1/9	001	111110	101	100	011	1000
g	1/3	01	111111	11	00	001	10

(a) Develop a Huffman code for the given pmf $p_X(x)$, calculate its average codeword length and its absolute and relative redundancy.

<u>*Solution:*</u>

The Huffman algorithm can be described as follows: First, we create a symbol group for each alphabet letter. Then, in each iteration, the symbol groups are sorted according to their associated probabilities. Two symbol groups with the smallest probabilities are selected, and each of the two symbol groups is characterized by a single bit. Then, the two selected symbol groups are summarized to a new symbol group. This process is repeated until a single symbol group is obtained. Finally, the constructed binary code tree is converted into a prefix code using the assigned bits.

The construction of the binary code tree for the given pmf is illustrated in the following table.

	sorted probabilities associated symbol groups assigned bits						
step 1	1/3 a	1/3 g	1/9 b	1/9 f	1/27 c	1/27 d **0**	1/27 e **1**
step 2	1/3 a	1/3 g	1/9 b	1/9 f	2/27 de **0**	1/27 c **1**	
step 3	1/3 a	1/3 g	1/9 b	1/9 f **0**	1/9 cde **1**		
step 4	1/3 a	1/3 g	2/9 $cdef$ **0**	1/9 b **1**			
step 5	1/3 a	1/3 g **0**	1/3 $bcdef$ **1**				
step 6	2/3 $bcdefg$ **0**	1/3 a **1**					

Given the developed binary code tree, the codeword for each particular symbol $x \in \mathcal{A}$ is constructed by concatenating the bits that are assigned to the symbol groups containing the particular symbol, starting with the last iteration (i.e., the largest symbol groups) of the above described algorithm. The resulting code for the above illustrated code construction is shown in the table below.

x	codeword
a	1
b	011
c	01011
d	010100
e	010101
f	0100
g	00

Note that there are multiple codes for a given pmf that can be constructed with the Huffman algorithm. We could sort the probabilities with the same values in a different order, and we could switch the assignment of 0 and 1 bits in some or all of the iteration steps.

The average codeword length per symbol is given by

$$\begin{aligned}\bar{\ell} &= \sum_{x\in\mathcal{A}} p_X(x)\cdot\ell(x)\\ &= \frac{1}{3}+\frac{3}{9}+\frac{5}{27}+\frac{6}{27}+\frac{6}{27}+\frac{4}{9}+\frac{2}{3}\\ &= \frac{3}{3}+\frac{7}{9}+\frac{17}{27}=\frac{27+21+17}{27}\\ &= \frac{65}{27}\approx 2.407.\end{aligned}$$

The entropy of the random variables $X_n = X$ is given by

$$\begin{aligned}H(X) &= -\sum_{x\in\mathcal{A}} p_X(x)\,\log_2 p_X(x)\\ &= -2\cdot\frac{1}{3}\cdot\log_2\left(\frac{1}{3}\right)-2\cdot\frac{1}{9}\cdot\log_2\left(\frac{1}{9}\right)-3\cdot\frac{1}{27}\cdot\log_2\left(\frac{1}{27}\right)\\ &= \frac{2}{3}\cdot\log_2(3)+\frac{2}{9}\cdot\log_2\left(3^2\right)+\frac{3}{27}\cdot\log_2\left(3^3\right)\\ &= \left(\frac{2}{3}+\frac{4}{9}+\frac{9}{27}\right)\cdot\log_2 3=\frac{18+12+9}{27}\cdot\log_2 3\\ &= \frac{13}{9}\cdot\log_2 3\approx 2.289.\end{aligned}$$

The absolute redundancy of the Huffman code is

$$\begin{aligned}\rho &= \bar{\ell}-H(X)\\ &= \frac{65}{27}-\frac{13}{9}\cdot\log_2 3\\ &= \frac{13}{27}\left(5-3\log_2 3\right)\approx 0.118.\end{aligned}$$

The absolute redundancy of the Huffman code is approximately 0.118 bit per symbol.

The relative redundancy of the Huffman code is

$$\begin{aligned}\frac{\rho}{H(X)} &= \frac{\bar{\ell} - H(X)}{H(X)} = \frac{\bar{\ell}}{H(X)} - 1 \\ &= \frac{5}{3\log_2 3} - 1 \approx 0.0515.\end{aligned}$$

The relative redundancy of the Huffman code is approximately 5.15%.

(b) For all codes A, B, C, D, E, and F, do the following:

- Calculate the average codeword length per symbol;
- Determine whether the code is a singular code;
- Determine whether the code is uniquely decodable;
- Determine whether the code is a prefix code;
- Determine whether the code is an optimal prefix code.

Solution:

The average codeword length per symbol is given by

$$\bar{\ell} = \sum_{x \in \mathcal{A}} p_X(x) \cdot \ell(x),$$

where $\ell(x)$ denote the length of the codeword for the alphabet letter x. As an example, the average codeword length for code C is

$$\bar{\ell}_C = \frac{2}{3} + \frac{2}{3} + \frac{3}{9} + \frac{3}{9} + \frac{3}{27} + \frac{4}{27} + \frac{4}{27} = \frac{4}{3} + \frac{6}{9} + \frac{11}{27} = \frac{36 + 18 + 11}{27} = \frac{65}{27}.$$

The average codeword length for all given codes are summarized in the following table, which also includes a summary of the answers for the other questions.

x	$p_X(x)$	A	B	C	D	E	F
a	1/3	1	0	00	01	000	1
b	1/9	0001	10	010	101	001	100
c	1/27	000000	110	0110	111	010	100000
d	1/27	00001	1110	0111	010	100	10000
e	1/27	000001	11110	100	110	111	000000
f	1/9	001	111110	101	100	011	1000
g	1/3	01	111111	11	00	001	10
$\bar{\ell}$		65/27	99/27	65/27	63/27	81/27	65/27
singular		no	no	no	no	yes	no
uniq. dec.		yes	yes	yes	no	no	yes
prefix		yes	yes	yes	no	no	no
opt. prefix		yes	no	yes	no	no	no

In the following, the properties of the given codes are briefly analyzed:

- Code A:
  - The code is not singular, since a different codeword is assigned to each alphabet letter.
  - The code is a prefix code, since no codeword represents a prefix (or the complete bit string) of another codeword.
  - Since the code is a prefix code, it is uniquely decodable.
  - The code is an optimal prefix code, since it is a prefix code and has the same average codeword length as a Huffman code for the given pmf (see 3a).
- Code B:
  - The code is not singular, since a different codeword is assigned to each alphabet letter.
  - The code is a prefix code, since no codeword represents a prefix (or the complete bit string) of another codeword.
  - Since the code is a prefix code, it is uniquely decodable.
  - The code is not an optimal prefix code, since the average codeword length is greater than that of the Huffman code for the given pmf (see 3a).
- Code C:
  - The code is not singular, since a different codeword is assigned to each alphabet letter.
  - The code is a prefix code, since no codeword represents a prefix (or the complete bit string) of another codeword.
  - Since the code is a prefix code, it is uniquely decodable.
  - The code is an optimal prefix code, since it is a prefix code and has the same average codeword length as a Huffman code for the given pmf (see 3a).
- Code D:
  - The code is not singular, since a different codeword is assigned to each alphabet letter.
  - The code is a not a prefix code, since the codeword "01" for the letter a represents a prefix of the codeword "010" for the letter d.
  - The code is not uniquely decodable, since the letter sequences "*aaa*" and "*db*" give the same bit string "010101".
  - The code is not an optimal prefix code, since it is no prefix code.
- Code E:
  - The code is singular, since the same codeword ("001") is assigned to the alphabet letters b and g.
  - Since the code is singular, it is not uniquely decodable, it is no prefix code, and it is not an optimal prefix code.
- Code F:
  - The code is not singular, since a different codeword is assigned to each alphabet letter.

- The code is a not a prefix code, since, for example, the codeword “1” for the letter a represents a prefix of the codeword “100” for the letter b.
- The code is uniquely decodable, since based on the number of successive bits equal to 0, the symbol sequence can be unambiguously determined given a bit sequence. This will be further explained in (3c).
- The code is not an optimal prefix code, since it is no prefix code.

(c) Briefly describe a process for decoding a symbol sequence given a finite sequence of K bits that is coded with code F.

Solution:

The decoding process can be described as follows:

(1) Set $n = 0$.

(2) Read the next bit b_n.

(3) Read all bits b_{n+i} until the next bit b_{n+m} equal to 1, excluding the bit b_{n+m} equal to 1, or, if the remaining bit sequence does not contain a bit equal to 1, the end of the bit sequence.

(4) Determine the number N_0 of read bits equal to 0, excluding the bit b_n and all previously read bits.

(5) Depending on the value of b_n, do the following:

- If b_n is equal to 0, output $(N_0 + 1)/6$ times the symbol e.
- If b_n is equal to 1, do the following:
  - If $N_0 \bmod 6 == 0$, output the symbol a.
  - If $N_0 \bmod 6 == 1$, output the symbol g.
  - If $N_0 \bmod 6 == 2$, output the symbol b.
  - If $N_0 \bmod 6 == 3$, output the symbol f.
  - If $N_0 \bmod 6 == 4$, output the symbol d.
  - If $N_0 \bmod 6 == 5$, output the symbol c.
  - If $N_0 \geq 6$, output $\lfloor N_0/6 \rfloor$ times the symbol e.

(6) Set $n = n + N_0 + 1$.

(7) If $n < K$, go to step (2).

Note that although the considered code K is uniquely decodable, it is not instantaneously decodable. In general, the next symbol is not known before the next bit equal to 1 (or the end of the message) has been detected and the number of successive zero bits can be arbitrarily large.

4. Given is a Bernoulli process $\mathbf{X}$ with the alphabet $\mathcal{A} = \{a, b\}$ and the pmf $p_X(a) = p$, $p_X(b) = 1 - p$. Consider the three codes in the following table.

Code A		Code B		Code C	
symbols	codeword	symbols	codeword	symbol	codeword
aa	1	aa	0001	a	0
ab	01	ab	001	b	1
b	00	ba	01		
		bb	1		

(a) Calculate the average codeword length per symbol for the three codes.

Solution:

The code A is a code that assigns variable-length codewords to variable-length symbol sequences. Let $\mathbf{s}_k$ be the symbol sequences to which the codewords are assigned. The average codeword length per symbol is the average codeword length per symbol sequence $\mathbf{s}_k$ divided by the average number of symbols per symbol sequence $\mathbf{s}_k$. With $\ell(\mathbf{s}_k)$ denoting the length of the codeword that is assigned to $\mathbf{s}_k$, $n(\mathbf{s}_k)$ denoting the number of symbols in the symbol sequence $\mathbf{s}_k$, and $p_S(\mathbf{s}_k)$ denoting the probability of the symbol sequence $\mathbf{s}_k$, we have

$$\bar{\ell}_A = \frac{\sum_{\forall \mathbf{s}_k} p_S(\mathbf{s}_k)\, \ell(\mathbf{s}_k)}{\sum_{\forall \mathbf{s}_k} p_S(\mathbf{s}_k)\, n(\mathbf{s}_k)}.$$

Note that the probability $p(\mathbf{s}_k)$ is given by

$$p_S(\mathbf{s}_k) = p^{n_a(\mathbf{s}_k)} \cdot (1-p)^{n_b(\mathbf{s}_k)},$$

where $n_a(\mathbf{s}_k)$ and $n_b(\mathbf{s}_k)$ represent the number of symbols equal to a and b, respectively, in the symbol sequence $\mathbf{s}_k$. Hence, the average codeword length per symbol for the code A is

$$\begin{aligned}
\bar{\ell}_A &= \frac{p_S(aa)\cdot\ell(aa) + p_S(ab)\cdot\ell(ab) + p_S(b)\cdot\ell(b)}{p_S(aa)\cdot n(aa) + p_S(ab)\cdot n(ab) + p_S(b)\cdot n(b)} \\
&= \frac{p^2\cdot 1 + p(1-p)\cdot 2 + (1-p)\cdot 2}{p^2\cdot 2 + p(1-p)\cdot 2 + (1-p)\cdot 1} \\
&= \frac{p^2 + 2p - 2p^2 + 2 - 2p}{2p^2 + 2p - 2p^2 + 1 - p} \\
&= \frac{2-p^2}{1+p}.
\end{aligned}$$

The code B is a code that assigns a codeword to each possible sequence of two symbols. Hence, the average codeword length per symbol is equal to the average codeword length per symbol sequence divided by 2,

$$\bar{\ell}_B = \frac{1}{2}\sum_{\forall \mathbf{s}_k} p_S(\mathbf{s}_k)\, \ell(\mathbf{s}_k)$$

$$
\begin{aligned}
&= \frac{1}{2}\left(p_S(aa)\ell(aa) + p_S(ab)\ell(ab) + p_S(ba)\ell(ba) + p_S(bb)\ell(bb)\right) \\
&= \frac{1}{2}\left(p^2 \cdot 4 + p(1-p)\cdot 3 + p(1-p)\cdot 2 + (1-p)^2 \cdot 1\right) \\
&= \frac{1}{2}\left(4p^2 + 3p - 3p^2 + 2p - 2p^2 + 1 - 2p + p^2\right) \\
&= \frac{1+3p}{2}.
\end{aligned}
$$

The code C assign a codeword of length 1 to both alphabet letters. Hence, its average codeword length per symbol is

$$\bar{\ell}_C = 1.$$

(b) For which probabilities p is the code A more efficient than code B?

Solution:
The code A is more efficient than code B if its average codeword length is less than the average codeword length of code B,

$$
\begin{aligned}
\bar{\ell}_A &< \bar{\ell}_B \\
\frac{2-p^2}{1+p} &< \frac{1+3p}{2} \\
4-2p^2 &< 1+3p+p+3p^2 \\
-5p^2-4p+3 &< 0 \\
p^2+\frac{4}{5}p-\frac{3}{5} &> 0.
\end{aligned}
$$

The quadratic function $y = p + \frac{4}{5}p - \frac{3}{5}$ is parabola, which opens upward (since the term p^2 is multiplied by a positive number). Hence, $y > 0$ if $p < p_1$ or $p > p_2$, where p_1 and p_2 are the roots of $0 = p^2 + \frac{4}{5}p - \frac{3}{5}$ with $p_1 \leq p_2$. The roots p_1 and p_2 are given by

$$
\begin{aligned}
p_{1/2} &= -\frac{2}{5} \mp \sqrt{\left(\frac{2}{5}\right) + \frac{3}{5}} \\
&= -\frac{2}{5} \mp \sqrt{\frac{4+15}{25}} \\
&= \frac{1}{5}\left(\mp\sqrt{19} - 2\right).
\end{aligned}
$$

Hence, we have

$$
\begin{aligned}
p_1 &= \frac{1}{5}\left(-\sqrt{19} - 2\right) \approx -1.2718, \\
p_2 &= \frac{1}{5}\left(\sqrt{19} - 2\right) \approx 0.4718.
\end{aligned}
$$

Consequently, the code A is more efficient than code B if

$$\frac{1}{5}\left(\sqrt{19} - 2\right) < p \leq 1,$$

or, approximately, if

$$0.4718 < p \leq 1.$$

(c) For which probabilities p is the simple code C more efficient than both code A and code B?

Solution:

The first condition is

$$\begin{aligned} \bar{\ell}_C &< \bar{\ell}_A \\ 1 &< \frac{2-p^2}{1+p} \\ 1+p &< 2-p^2 \\ p^2+p-1 &< 0. \end{aligned}$$

The quadratic function $y = p+p-1$ is parabola, which opens upward. Hence, $y < 0$ if $p_1 < p < p_2$, where p_1 and p_2 are the roots of $0 = p^2+p-1$ with $p_1 \leq p_2$. The roots p_1 and p_2 are given by

$$\begin{aligned} p_{1/2} &= -\frac{1}{2} \mp \sqrt{\left(\frac{1}{2}\right) + 1} \\ &= -\frac{1}{2} \mp \sqrt{\frac{1+4}{4}} \\ &= \frac{1}{2}\left(\mp\sqrt{5}-1\right). \end{aligned}$$

Hence, we have

$$\begin{aligned} p_1 &= \frac{1}{2}\left(-\sqrt{5}-1\right) \approx -1.6180, \\ p_2 &= \frac{1}{2}\left(\sqrt{5}-1\right) \approx 0.6180. \end{aligned}$$

Consequently, the code C is more efficient than code A if

$$0 \leq p < \frac{1}{2}\left(\sqrt{5}-1\right).$$

The second condition is

$$\begin{aligned} \bar{\ell}_C &< \bar{\ell}_B \\ 1 &< \frac{1+3p}{2} \\ 2 &< 1+3p \\ 1 &< 3p \\ p &> \frac{1}{3}. \end{aligned}$$

Hence, the code C is more efficient than code B if

$$\frac{1}{3} < p \leq 1.$$

By combining both derived conditions, we obtain that the simple code C is more efficient than both code A and code B if

$$\frac{1}{3} < p < \frac{1}{2}\left(\sqrt{5} - 1\right),$$

or, approximately, if

$$0.3333 < p < 0.6180.$$

For $0 \leq p < \frac{1}{3}$, code B is more efficient than code A and code C, and for $\frac{1}{2}\left(\sqrt{5} - 1\right) < p \leq 1$, code A is more efficient than code B and code C.

5. Given is a Bernoulli process $\mathbf{B} = \{B_n\}$ with the alphabet $\mathcal{A}_B = \{0, 1\}$, the pmf $p_B(0) = p$, $p_B(1) = 1 - p$, and $0 \le p < 1$. Consider the random variable X that specifies the number of random variables B_n that have to be observed to get exactly one "1".

   Calculate the entropies $H(B_n)$ and $H(X)$.

   For which value of p, with $0 < p < 1$, is $H(X)$ four times as large as $H(B_n)$?

   $$\textit{Hint:} \qquad \forall_{|a|<1}, \; \sum_{k=0}^{\infty} a^k = \frac{1}{1-a}, \qquad \forall_{|a|<1}, \; \sum_{k=0}^{\infty} k\, a^k = \frac{a}{(1-a)^2}.$$

   Solution:

   The entropy for the Bernoulli process $\mathbf{B}$ is

   $$\begin{aligned} H_B(p) &= H(B_n) = -p_B(0) \log_2 p_B(0) - p_B(1) \log_2 p_B(1) \\ &= -p \log_2 p - (1-p) \log_2 (1-p). \end{aligned}$$

   For calculating the entropy of the random variable X, we first determine its pmf $p_X(x)$. The alphabet for the random variable X is $\mathcal{A}_X = \{1, 2, \cdots\}$. We may see a "1" in the first observation of B_n (for the special case $p = 0$, we always see a "1" in the first observation), or in the second observation of B_n, etc. It is, however, also possible that we have to look at an arbitrarily large number of random variables B_n before we see a "1".

   The probability mass $p_X(k)$ is the probability that $X = k$ and, hence, the probability that we see a symbol string "$B_n B_{n+1} \cdots B_{n+k-2} B_{n+k-1}$" equal to "$00 \cdots 01$",

   $$\begin{aligned} p_X(k) = P(X = k) = P(\{B_n = 0\} \cap \{B_{n+1} = 0\} \cap \cdots \\ \cap \{B_{n+k-2} = 0\} \cap \{B_{n+k-1} = 1\}). \end{aligned}$$

   Since the random variables B_n are independent (a Bernoulli process is a binary iid process), we have

   $$\begin{aligned} p_X(k) &= P(B_{n+k-1} = 1) \cdot \prod_{i=0}^{k-2} P(B_{n+i} = 0) = p_B(1) \cdot p_B(0)^{k-1} \\ &= (1-p)\, p^{k-1}. \end{aligned}$$

   The pmf for the random variable X is a geometric pmf. Its entropy is given by

   $$\begin{aligned} H_X(p) &= H(X) = -\sum_{i=1}^{\infty} p_X(i) \log_2 p_X(i) \\ &= -\sum_{i=1}^{\infty} (1-p) p^{i+1} \log_2 \left((1-p) p^{i+1} \right) \\ &= -\sum_{k=0}^{\infty} (1-p) p^{k} \log_2 \left((1-p) p^{k} \right) \end{aligned}$$

$$
\begin{aligned}
&= -\sum_{k=0}^{\infty}(1-p)p^k\left(\log_2(1-p)+k\log_2 p\right)\\
&= -\sum_{k=0}^{\infty}\big((1-p)\log_2(1-p)\big)\,p^k-\big((1-p)\log_2(1-p)\big)\,k\,p^k\\
&= -(1-p)\log_2(1-p)\left(\sum_{k=0}^{\infty}p^k\right)-(1-p)\log_2 p\left(\sum_{k=0}^{\infty}kp^k\right).
\end{aligned}
$$

For the case we are considering, $0 \le p < 1$, the series in the above equation converge and we can write

$$
\begin{aligned}
H_X(p) &= -\big((1-p)\log_2(1-p)\big)\frac{1}{1-p}-\big((1-p)\log_2 p\big)\frac{p}{(1-p)^2}\\
&= -\log_2(1-p)-\frac{p}{1-p}\log_2 p.
\end{aligned}
$$

By reformulating the above expression, we obtain

$$
\begin{aligned}
H_X(p) &= \frac{1}{1-p}\Big(-(1-p)\log_2(1-p)-p\log_2 p\Big)\\
&= \frac{1}{1-p}\,H_B(p).
\end{aligned}
$$

We now determine the value of p, with $0 < p < 1$, for which $H_X(p)$ is four times as large as $H_B(p)$,

$$
\begin{aligned}
H_X(p) &= 4\,H_B(p)\\
\frac{1}{1-p}\,H_B(p) &= 4\,H_B(p)
\end{aligned}
$$

For $0 < p < 1$, $H_B(p)$ is greater than 0. Hence, we can divide the above equation by $H_B(p)$ and obtain

$$
\begin{aligned}
\frac{1}{4} &= 1-p\\
p &= \frac{3}{4} = 0.75.
\end{aligned}
$$

For $p = 0.75$, the entropy of the random variable X is four times as large as the entropy of the Bernoulli process.

6. Proof the chain rule for the joint entropy,

$$H(X,Y) = H(X) + H(Y|X).$$

<u>*Solution:*</u>

With $\mathcal{A}_X$ and $\mathcal{A}_Y$ being the alphabets of the random variables X and Y, respectively, the joint entropy $H(X,Y)$ is defined as

$$H(X,Y) = -\sum_{x\in\mathcal{A}_X}\sum_{y\in\mathcal{A}_Y} p_{XY}(x,y)\,\log_2 p_{XY}(x,y).$$

Using the chain rule $p_{XY}(x,y) = p_X(x)p_{Y|X}(y|x)$ for the joint probability masses, we obtain

$$\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{A}_X}\sum_{y\in\mathcal{A}_Y} p_{XY}(x,y)\,\log_2 p_X(x)\\
&\quad -\sum_{x\in\mathcal{A}_X}\sum_{y\in\mathcal{A}_Y} p_{XY}(x,y)\,\log_2 p_{Y|X}(y|x)\\
&= -\sum_{y\in\mathcal{A}_X}\left(\sum_{x\in\mathcal{A}_Y} p_{XY}(x,y)\right)\log_2 p_X(x) + H(Y|X)\\
&= -\sum_{y\in\mathcal{A}_X} p_X(x)\,\log_2 p_X(x) + H(Y|X)\\
&= H(X) + H(Y|X).
\end{aligned}$$

7. Investigate the entropy of a function of a random variable X. Let X be a discrete random variable with the alphabet $\mathcal{A}_X = \{0, 1, 2, 3, 4\}$ and the binomial pmf

$$p_X(x) = \begin{cases} 1/16 & : \quad x = 0 \vee x = 4 \\ 1/4 & : \quad x = 1 \vee x = 3 \\ 3/8 & : \quad x = 2 \end{cases} .$$

(a) Calculate the entropy $H(X)$.

Solution:

Inserting the given probability masses into the definition of entropy yields

$$\begin{aligned} H(X) &= -2 \cdot \frac{1}{16} \cdot \log_2 \left(\frac{1}{16}\right) - 2 \cdot \frac{1}{4} \cdot \log_2 \left(\frac{1}{4}\right) - 1 \cdot \frac{3}{8} \cdot \log_2 \left(\frac{3}{8}\right) \\ &= 2 \cdot \frac{1}{16} \cdot 4 + 2 \cdot \frac{1}{4} \cdot 2 + 1 \cdot \frac{3}{8} \cdot (3 - \log_2 3) \\ &= \frac{1}{2} + 1 + \frac{9}{8} - \frac{3}{8} \log_2 3 = \frac{21}{8} - \frac{3}{8} \log_2 3 \\ &= \frac{3}{8} (7 - \log_2 3) \approx 2.0306. \end{aligned}$$

(b) Consider the functions $g_1(x) = x^2$ and $g_2(x) = (x - 2)^2$.
Calculate the entropies $H(g_1(X))$ and $H(g_2(X))$.

Solution:

Let Y be the random variable $Y = g(X)$. The alphabet of Y is given by the alphabet of the random variable X and the function $g_1(x)$,

$$\mathcal{A}_Y = \bigcup_{x \in \mathcal{A}_X} g_1(x) = \{0, 1, 4, 9, 16\}.$$

Similarly, let Z be the random variable $Z = g_2(X)$. The alphabet of Z is given by

$$\mathcal{A}_Z = \bigcup_{x \in \mathcal{A}_X} g_2(x) = \{0, 1, 4\}.$$

The pmf for Y is given by

$$p_Y(y) = \sum_{x \in \mathcal{A}_X \colon y = g_1(x)} p_X(x) = \begin{cases} 1/16 & : \quad y = 0 \vee y = 16 \\ 1/4 & : \quad y = 1 \vee y = 9 \\ 3/8 & : \quad y = 4 \end{cases} .$$

Similarly, the pmf for Z is given by

$$p_Z(z) = \sum_{x \in \mathcal{A}_X \colon z = g_2(x)} p_X(x) = \begin{cases} 3/8 & : \quad z = 0 \\ 1/2 & : \quad z = 1 \\ 1/8 & : \quad z = 2 \end{cases} .$$

Using the determined pmfs for calculating the entropies, we obtain

$$\begin{aligned}
H(g_1(X)) &= H(Y) = -\sum_{y\in\mathcal{A}_\mathcal{Y}} p_Y(y)\,\log_2 p_Y(y) \\
&= -2\cdot\frac{1}{16}\cdot\log_2\left(\frac{1}{16}\right) - 2\cdot\frac{1}{4}\cdot\log_2\left(\frac{1}{4}\right) - 1\cdot\frac{3}{8}\cdot\log_2\left(\frac{3}{8}\right) \\
&= 2\cdot\frac{1}{16}\cdot 4 + 2\cdot\frac{1}{4}\cdot 2 + 1\cdot\frac{3}{8}\cdot(3-\log_2 3) \\
&= \frac{1}{2} + 1 + \frac{9}{8} - \frac{3}{8}\log_2 3 = \frac{21}{8} - \frac{3}{8}\log_2 3 \\
&= \frac{3}{8}\,(7-\log_2 3) \\
&= H(X),
\end{aligned}$$

and

$$\begin{aligned}
H(g_2(X)) &= H(Z) = -\sum_{z\in\mathcal{A}_\mathcal{Z}} p_Z(z)\,\log_2 p_Z(z) \\
&= -\frac{3}{8}\cdot\log_2\left(\frac{3}{8}\right) - \frac{1}{2}\cdot\log_2\left(\frac{1}{2}\right) - \frac{1}{8}\cdot\log_2\left(\frac{1}{8}\right) \\
&= \frac{3}{8}\cdot(3-\log_2 3) + \frac{1}{2} + \frac{3}{8} \\
&= \frac{1}{8}\,(16 - 3\log_2 3) \\
&= H(X) - \frac{5}{8}.
\end{aligned}$$

(c) Proof that the entropy $H(g(X))$ of a function $g(x)$ of a random variable X is not greater than the entropy of the random variable X,

$$H(g(X)) \le H(X)$$

Determine the condition under which equality is achieved.

<u>Solution:</u>
Using the chain rule, $H(X,Y) = H(X) + H(Y|X)$, we can write

$$\begin{aligned}
H(X, g(X)) &= H(g(X), X) \\
H(X) + H(g(X)|X) &= H(g(X)) + H(X|g(X)) \\
H(g(X)) &= H(X) + H(g(X)|X) - H(X|g(X)).
\end{aligned}$$

Since the random variable $g(X)$ is a function of the random variable X, the value of $g(X)$ is known if the value of X is known. Hence, the conditional probability mass function $p_{g(X)|X}(y|x)$ is given by

$$p_{g(X)|X}(y|x) = \left\{ \begin{array}{lcl} 1 & : & y = g(x) \\ 0 & : & y \ne g(x) \end{array} \right. .$$

Let $\mathcal{A}_X$ denote the alphabet of the random variables X with $\forall x \in \mathcal{A}_X, p_X(x) > 0$. Similarly, let $\mathcal{A}_Y$ denote the alphabet of the random variable $Y = g(X)$ with $\forall y \in \mathcal{A}_Y, p_Y(y) > 0$. The conditional entropy $H(g(X)|X)$ is given by

$$\begin{aligned} H(g(X)|X) &= -\sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} p_{X,g(X)}(x,y) \, \log_2 p_{g(X)|X}(y|x) \\ &= -\sum_{x \in \mathcal{A}_X} p_X(x) \left(\sum_{y \in \mathcal{A}_Y} p_{g(X)|X}(y|x) \, \log_2 p_{g(X)|X}(y|x) \right). \end{aligned}$$

The terms with $p_{g(X)|X}(y|x) = 0$ do not contribute to the sum in parenthesis and can be ignored. We obtain,

$$\begin{aligned} H(g(X)|X) &= -\sum_{x \in \mathcal{A}_X} p_X(x) \, p_{g(X)|X}(g(x)|x) \, \log_2 p_{g(X)|X}(g(x)|x) \\ &= -\big(1 \cdot \log_2 1\big) \cdot \sum_{\forall x} p_X(x) = -\log_2 1 \\ &= 0. \end{aligned}$$

Hence, the conditional entropy $H(g(X)|X)$ is always equal to 0, and we obtain for $H(g(X))$,

$$H(g(X)) = H(X) - H(X|g(X)).$$

Since the entropy is always greater than or equal to 0, we have proved that

$$H(g(X)) \leq H(X).$$

If and only if $g(X)$ is an injective function for all letters of the alphabet $\mathcal{A}_X$, i.e., if $\forall a, b \in \mathcal{A}_X$, $a \neq b$ implies $g(a) \neq g(b)$, we can define an inverse function $h(y)$, so that $h(g(x)) = x$, $\forall x \in \mathcal{A}_X$. In this case, we obtain

$$H(X|g(X)) = H(h(g(X))|g(X)) = 0,$$

Consequently, if $g(X)$ is an injective function for all letters of the alphabet $\mathcal{A}_Y$, the entropy $H(g(X))$ is equal to the entropy $H(X)$,

$$H(g(X)) = H(X).$$

If $g(X)$ is not an injective function for all letters of the alphabet $\mathcal{A}_Y$, i.e., if there are two alphabet letters a and $b \neq a$, with $g(a) = g(b)$, the entropy $H(g(X))$ is less than the entropy $H(X)$,

$$H(g(X)) < H(X).$$