Signal Theory

Prof. dr. ir. Gerd Vandersteen

December 13, 2018
# Contents

1 Information Theory  
1.1 Parts of a communication system .......................... 5  
1.2 Important achievements in Information theory ............... 7  

2 Discrete communication systems  
2.1 Entropy of independent discrete sources ...................... 10  
2.1.1 Notation ................................................. 10  
2.1.2 Entropy as introduced by Shannon .......................... 11  
2.1.3 Alternative approach ..................................... 14  
2.1.4 Interpretation of entropy ................................... 16  
2.1.5 Binary entropy function ................................... 17  
2.1.6 Entropy for alphabet with length \( n \) ....................... 18  
2.2 Information rate (entropy rate) of independent sources ...... 19  
2.2.1 General definition of information rate ...................... 19  
2.2.2 Information rate for binary coders ........................ 20  
2.3 Redundancy of independent discrete sources ................... 20  
2.4 Examples of independent discrete source ....................... 21  
2.4.1 Academic example ......................................... 21  
2.4.2 Letter/symbol frequency in different languages ............. 24  
2.4.3 Morse code ................................................. 26  
2.5 Markov processes ............................................. 30  
2.5.1 State diagram ............................................. 31  
2.5.2 Trellis diagram ............................................. 33  
2.5.3 Irreducible Markov chain ................................... 35  
2.5.4 Ergodic Markov processes .................................. 37  
2.6 Entropy of an ergodic discrete Markov information sources ... 40  
2.7 Discrete joint entropy, conditional entropy, and mutual information 41  
2.7.1 Discrete joint entropy ..................................... 42  
2.7.2 Discrete conditional entropy (=equivocation) ............... 43  
2.7.3 Discrete mutual information (=transinformation) ........... 48  
2.7.4 Binary Symmetric Channel .................................. 51  
2.8 Discrete channel capacity ..................................... 53  
2.8.1 Channel capacity of a Binary Symmetric Channel .......... 53  
2.8.2 Generalization ............................................ 54
CONTENTS

2.9 Discrete encoding and decoding in transceivers .......................... 58
  2.9.1 Transducers .................................................. 58
  2.9.2 Data processing inequality .................................. 58

2.10 Shannon theorem .................................................. 60

3 Continuous communication systems ....................................... 62
  3.1 Extending information concepts to continuous random variables . 63
    3.1.1 Properties of the entropy for continuous random variables 64
    3.1.2 Changing coordinate system .................................. 64
    3.1.3 The entropy of a uniform distribution .......................... 66
    3.1.4 The entropy of a Gaussian distribution ........................ 66
    3.1.5 The entropy of a multi-dimensional Gaussian distribution case ........................................ 66
  3.2 Probability density functions \( f(x) \) that maximize the entropy \( H(X) \) 68
    3.2.1 Calculus of variations to determine \( f(x) \) that maximizes \( H(X) \) .................................................. 68
    3.2.2 Uniform distribution maximizes the entropy under bounded amplitude constraint .......................... 70
    3.2.3 Gaussian distribution maximizes the entropy under fixed power constraint .................................. 72
  3.3 Additive White Gaussian Noise (AWGN) channels .................... 74
  3.4 Complex-valued continuous random variables ........................ 74

4 Continuous time-signals .................................................. 77
  4.1 Nyquist-Shannon sampling theorem for band-limited systems ... 77
  4.2 Band-limited Gaussian channel .................................. 79
    4.2.1 Nyquist Inter Symbol Interference (ISI) criterion ......... 80
    4.2.2 Raised-cosine filter ......................................... 81
    4.2.3 Root-Raised-Cosine filter .................................. 84
  4.3 Complex-valued representation of passband signals ................ 85
    4.3.1 Analog modulation .......................................... 86
    4.3.2 Digital modulation ......................................... 91
    4.3.3 Nyquist-Shannon sampling theorem for passband signals . 100

5 Shannon-Hartley theorem on channel capacity .......................... 101
  5.1 General introduction .......................................... 102
    5.1.1 Bounded amplitude constraint ................................ 103
    5.1.2 Fixed power constraint ...................................... 103
  5.2 Shannon-Hartley theorem for independent continuous variables 104
    5.2.1 Single Gaussian channel .................................... 104
    5.2.2 Independent parallel Gaussian channels .................... 105
    5.2.3 Power distribution for independent parallel Gaussian channels ........................................ 106
    5.2.4 General multi-variable case .................................. 108
    5.2.5 Shannon-Hartley theorem for Multiple-Input Multiple-Output (MIMO) systems .......................... 110
5.3 Shannon-Hartley theorem ........................................ 112
  5.3.1 Shannon-Hartley theorem .................................. 112
  5.3.2 Shannon-Hartley theorem for independent parallel channels113
  5.3.3 Power distribution for independent parallel channels ... 114
5.4 Shannon’s limit of the channel capacity .................... 115
  5.4.1 Spectral efficiency .......................................... 116
  5.4.2 Upper bound for attainable spectral efficiency .......... 116
  5.4.3 Band-limited Shannon limit ................................ 118
  5.4.4 Ultimate Shannon limit .................................... 119
  5.4.5 Power and bandwidth-limited regime .................... 120
  5.4.6 Application of the Shannon limit:
      modulation schemes and forward error correction ........ 120
6  Optimal binary encoding ....................................... 123
  6.1 Uniquely decodable codes ................................. 124
  6.2 Prefix codes .................................................. 126
  6.3 Kraft-McMillan inequality ................................. 128
  6.4 Binary prefix coding techniques .......................... 130
      6.4.1 Shannon coding ......................................... 130
      6.4.2 Shannon-Fano coding .................................. 132
      6.4.3 Huffman coding ......................................... 132
  6.5 Shannon’s fundamental source coding theorem ............. 135
7  Basic concepts in signal theory ................................ 140
  7.1 Classification of signals .................................... 141
      7.1.1 Deterministic signals .................................. 141
      7.1.2 Stochastic signals ..................................... 141
      7.1.3 Periodic signals ....................................... 141
  7.2 Properties of stochastic signals ............................ 142
  7.3 Ensemble averages ........................................... 144
      7.3.1 Stationary signals ...................................... 146
      7.3.2 Properties for wide-sense stationary signals ....... 148
      7.3.3 Auto-correlation after linear time-invariant (LTI) filtering 152
      7.3.4 Properties for periodic signals ....................... 154
      7.3.5 Modulation of WSS process .............................. 157
  7.4 Time averages or realization means ........................ 159
  7.5 Ergodicity ..................................................... 161
8  Spectral analysis of stochastic processes .................... 164
  8.1 Power spectra using Wiener-Kinchin theorem ............... 165
  8.2 Properties of power spectra ................................ 167
  8.3 Power spectrum after linear time-invariant (LTI) filtering ... 170
  8.4 Power spectrum of periodic auto-correlation ............... 172
  8.5 Periodogram .................................................. 173
      8.5.1 Power representation in the frequency domain ....... 174
      8.5.2 Ergodicity in the spectrum ............................. 175
## CONTENTS

8.6 Applications and examples .................................. 176
  8.6.1 Stationary white noise ................................. 176
  8.6.2 Stationary Gaussian noise .............................. 176
  8.6.3 Filtered white noise .................................... 179
  8.6.4 Binary synchronous Non-Return-to-Zero (NRZ) signal .... 180
  8.6.5 Random Telegraph Signal .............................. 188
  8.6.6 Flicker or 1/f noise .................................... 194
  8.6.7 Power spectra of modulated processes .................... 196

9 Karhunen–Loève expansion .................................. 201
  9.1 Fourier expansion ........................................... 202
  9.2 Karhunen-Loève expansion ................................. 203
  9.3 Karhunen-Loève expansion applied to band-limited signals ... 207
    9.3.1 Discrete Karhunen-Loève expansion applied to band-limited signals ..................... 209

10 Wiener filter ................................................. 213
  10.1 Orthogonality principle .................................. 214
  10.2 Wiener filter .............................................. 215
  10.3 Wiener filter for uncorrelated signal and noise .......... 216

11 Matched filter ............................................. 218
  11.1 Derivation of matched filter .............................. 218
  11.2 Examples of matched filters .............................. 222
    11.2.1 Radar / Sonar ....................................... 222
    11.2.2 GPS .................................................. 224
    11.2.3 Timing synchronization in digital communication .... 225
    11.2.4 Integrate and dump ................................... 227
    11.2.5 Root-raised-cosine filter as matched filter ........ 230
Chapter 1

Information Theory

1.1 Parts of a communication system

A communication system considered in this chapter consists essentially of the five parts shown in Fig. 1.1:

![Schematic diagram of a general communication system](image)

Figure 1.1: Schematic diagram of a general communication system [1].

An information source produces a message or sequence of messages to be communicated to the receiving terminal. The message may be of various types:

- a sequence of characters or bytes;
- a single function of time $f(t)$ as in audio;
- two or more functions of time, e.g. stereo audio signals or systems intended to service several individual channels in multiplex;
- a function of time and other variables as analog television where the message may be thought of as a function of two space coordinates and time;
• combinations of various types of information sources.

A transmitter which operates on the message in some way to produce a signal suitable for transmission over the channel. The transmitter therefore takes an information source as input and transforms this into a different information stream. Different parts can be identified in the transmitter to code the original message towards a signal:

Compression aims to reduce the size of the message. This can be done with or without loss. Compressing a file in a zip-file is an example of a compression without loss of information, while the jpeg compression of an image is an example of a compression with loss of information.

Channel coding (also known as Forward Error Correction (FEC) coding) avoids the message of getting corrupt due to transmission errors. This is achieved by add information that allows to detect and even correct errors at the receiver side.

Encryption avoids the message from being read by others, for example when sending passwords or personal data. Encryption will make the message unreadable for outsiders who don’t know the key to decode the original message from the received signal.

Analog and digital modulation transforms the analog or digital information from the source into a signal that is more appropriate to transport over the physical channel by converting e.g. to a higher frequency. Examples are the analog AM and FM modulation techniques and the digital PAM, BPSK, PSK, QAM, OFDM,... digital modulation techniques.

The channel is merely the medium used to transmit the signal from transmitter to receiver. It can take different forms, including the communication of data over a distance using an electrical (wire communication), optical (fiber optics), acoustical (sound), electromagnetic (radio waves), and the storage/retrieval of data (hard disk drives, DVD players, QR-codes). The original signal of the transmitter is disturbed in the channel by a noise source, introducing errors in the received signal.

The receiver ordinarily performs the inverse operation of the transmitter, i.e. reconstructing the message from the received signal.

The destination is the person (or device) for whom the message is intended.

Communication systems can be classified in three main categories: discrete, continuous, and mixed. A discrete system is such that both the the message and the signal are a sequence of discrete symbols. A typical case is communication and storage of digital data.

A continuous system treats both the message and the signal as continuous functions, e.g., radio or analog television. Mixed system are dependent on both discrete and continuous variables, e.g., PCM (Pulse Code Modulation) transmission of speech.
1.2 Important achievements in Information theory

Information theory aims to quantify the following concepts:

**Entropy: a quantitative measure of information** This is done through the concept of entropy, \( H \), that expresses the amount of information as a function of the probability that a message occurs. The concept of entropy is introduced on the bases of axioms for discrete random variables (Chapter 2) that is afterwards extended to continuous variables (Chapter 3). The unity of the entropy is a bit. The average entropy of a symbol is e.g. expressed in bits per symbol (\( \text{bit/sym} \)).

**Information rate: average entropy** The information rate \( R \) measures the amount of information (expressed using the entropy) that is transmitted on average per time unit. Hence, the information rate is often expressed in bits per seconds (\( \text{bit/s} \)).

**Mutual information: measure of information transport over a channel** Information can get lost (=transmission errors) when transporting it over a channel. The amount of information that is transported over a channel is measured using the concept of entropy and is expressed using the mutual information of the channel, \( I \).

**Channel capacity: maximum information rate over channel** The channel capacity, \( C \), is by definition the maximum information rate that one can possibly attain for a given channel and for an optimal coding of the information at the source side. Hence, the information rate \( R \) of a channel will always be smaller or equals to the channel capacity \( C \) if no loss of information is allowed:

\[
R \leq C
\]

**Shannon-Hartley theorem** This theorem expresses the channel capacity under some very general conditions (additive white Gaussian noise channel),

\[
C = W \log (1 + SNR)
\]

and is function of the bandwidth \( W \) of the channel and the Signal-to-Noise-Ratio (SNR) of the received channel. This theorem is not only used to compute the capacity, but also gives insight in how the signal power can be distributed optimally when using multiple channels and it also gives the fundamentals for Multiple-Input Multiple-Output (MIMO) communication.

**Fundamental Shannon limits** The Shannon-Hartley theorem provides insight in limits of the channel. Hence, interpreting this theorem makes it possible to determine (a) what the channel capacity is when the bandwidth is not an issue, (b) determine the smallest SNR to obtain a reliable
communication when e.g. transmitting 4 bits per symbols (as in 16-QAM modulation), (c) determine the gap between the information rate (possibly after channel coding) and the fundamental limit.
Chapter 2

Discrete communication systems

The discrete communication systems studied in this chapter represent the information using a (discrete) sequence of discrete random variables. Hence, both the $k$ index representing the progress in time takes discrete values, as well as the possible values that can be realized, $x(k)$, take.

Discrete information sources are introduced by first determining a way of measuring the amount of information in a message: the entropy for independent events (Section 2.1). Afterwards, a measure for the amount of information generated per time unit is introduced in Section 2.2. The amount of information, i.e. the entropy, and the information rate of a source can be compared with the maximum possible entropy/rate. This leads to the definition of redundancy and efficiency introduced in Section 2.3. These (relatively) abstract definitions finally illustrated on some practical examples in Section 2.4.

As most information sources have some kind of correlation, the concept of correlation is introduced through the use of Markov processes (Section 2.5). The entropy for independent events are extended to Markov processes by considering the probability of each state of the Markov process and the resulting entropy generated by the state transitions in each state (Section 2.6).

After introducing the information properties of a information source, it is time to study the impact of a channel. This requires the extension of the concept of entropy on a single random variable $X$ (i.e. the information source), towards two random variables $X$ and $Y$ (i.e. the received signal). This leads to the important concepts such as joint / conditional entropy and mutual information (Section 2.7).

It is possible to determine a source $X$ such that the mutual information over a channel is maximized. This maximum mutual information is known as the channel capacity and represents the maximum amount of information that can be transmitted over the channel (Section 2.8).

After describing (discrete) information sources and the channels, it is nec-
necessary to known how to process information in order to modulate/demodulate the signals in the transmitter/receiver. This includes the encoder and decoders techniques which are encountered in e.g. compression techniques, channel coding,... The so-called transducers take an information source as input, process the information, and generate an information sequence as output (Section 2.9). This will give insight in the necessary conditions that need to be fulfilled to have a lossless communication.

Last but not least, Shannon theorem is studied in Section 2.10 which puts a constraint on the information rate for a given channel capacity.

2.1 Entropy of independent discrete sources

There are two ways to introduce a quantitative measure for the amount of information produced by the source, also known as the entropy. The first approach (used by Shannon [1]) starts with tree axioms and then proof that there is a (unique) function that fulfills these properties. An alternative way is through the introduction of self-information and will be discussed to clarify various properties of entropy.

2.1.1 Notation

The following notations will be used within this chapter.

Overall convention used

- Capital Roman characters: used to indicate stochastic processes, e.g. random variable $X$ which is the distribution of the roll of a dice
- Small Roman characters: the outcome of a stochastic realization, e.g. $x$ represents a particular realization of rolling a dice
- Greek characters: Stochastic moments of the random variable, i.e. $\sigma^2_X$ is the variance of $X$.

Set of possible symbols

- Discrete random variable $X$ with $X \in \{x_i\}$ for $i = 1, \ldots, n$
- Discrete random variable $Y$ with $Y \in \{y_j\}$ for $j = 1, \ldots, m$

Probabilities

Probability of the symbol $x_i$

$$p(x_i) = P[X = x_i]$$

Joint probabilities

$$p(x_i, x_j) = P[X = x_i \cap X = x_j]$$
\[ p(x_i, y_j) = P[X = x_i \cap Y = y_j] \]

Conditional probabilities
\[ p(x_i | x_j) = P[X = x_i | X = x_j] \]
\[ p(x_i | y_j) = P[X = x_i | Y = y_j] \]

Sequences of symbols
Discrete sequence, e.g. symbol at instance \( k \): \( x(k) \) with
\[ p(x_i(k)) = P[X(k) = x_i(k)] \]
Continuous sequence over time, e.g. symbol at time instance \( t \): \( x(t) \) with
\[ p(x_i(t)) = P[X(t) = x_i(t)] \]

2.1.2 Entropy as introduced by Shannon

Assume a set of \( n \) possible events, \( x_1, \ldots, x_n \), of the discrete random variable \( X \) with probability of occurrence \( p(x_1), \ldots, p(x_n) \). The measure \( H(X) = H(p(x_1), \ldots, p(x_n)) \), which is a function of the probabilities of \( p(x_i) \) will be called the entropy, and must fulfill the following assumptions:

1. \( H \) must be continuous in \( p(x_i) \).
2. If all \( p(x_i) \) are equal \( (p(x_i) = 1/n) \), then \( H \) should be a monotonic increasing function of \( n \). There is more choice, and hence uncertainty, when there are more possible equally likely events.
3. If a choice is broken down into two successive choices, the original \( H \) should be the weighted sum of the individual values of \( H \).

The latter can be illustrated using an example.
Figure 2.1: Decomposition of a choice from three possibilities [1].

Consider three possible events \( n = 3 \) with \( p(x_1) = \frac{1}{2}, p(x_2) = \frac{1}{3}, \) and \( p(x_3) = \frac{1}{6} \). Fig. 2.1 shows the breaking down of \( H(p(x_1), p(x_2), p(x_3)) \) into two successive choices. The final \( x_i \) have the same probabilities (and hence information content) as before. It is therefore required that

\[
H(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}) = H(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2} H(\frac{2}{3}, \frac{1}{3})
\]

The coefficient \( \frac{1}{2} \) before the second term is required since this second choice only occurs half the time.
This can be expressed in a more general way using Fig. 2.2. It can be seen that $H(p(x_1), p(x_2), p(x_3))$ should be equal to

$$H(p(x_1), 1 - p(x_1)) + (1 - p(x_1))H\left(\frac{p(x_2)}{1 - p(x_1)}, \frac{p(x_3)}{1 - p(x_1)}\right)$$

**Definition 1 (Entropy).** The entropy $H(X)$ of the source $X$ that satisfies the three above assumptions is of the form

$$H(X) = -K \sum_{i=1}^{n} p(x_i) \log p(x_i) \quad (2.1)$$

where $K$ is a positive constant.

The constant $K$ merely amounts to a choice of a unit of measure or, equivalently, the base used for the logarithm:

$$\log_b p = \frac{\log_a p}{\log_a b} = \frac{\ln p}{\ln b}$$

It is common to use the base 2 of the logarithm, leading to an entropy that is expressed in the unit 'bits'. We will therefore assume implicitly though the notes that $\log x$ represents $\log_2 x$, also known as the binary logarithm. In that case, the entropy is expressed in bit/symbol. The natural logarithms will be represented by $\ln x$.

The definition of the entropy can hence be simplified to
\[ H(X) = - \sum_{i=1}^{n} p(x_i) \log p(x_i) \]  

Equation (2.2) satisfies the assumed axioms:

1. \( H(X) \) is a continuous function in \( p(x_i) \).
2. For \( p(x_i) = \frac{1}{n} \), the entropy is monotonously increasing with \( n \). Therefore, consider

\[
H = - \sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} = \log n
\]  

which is increasing monotonously.
3. For the case of Fig. 2.2, we can express

\[
H(p(x_1), 1 - p(x_1)) + (1 - p(x_1))H\left(\frac{p(x_2)}{1 - p(x_1)}, \frac{p(x_3)}{1 - p(x_1)}\right)
\]

as

\[
\begin{align*}
-p(x_1) \log p(x_1) - (1 - p(x_1)) \log (1 - p(x_1)) & \\
+ (1 - p(x_1)) \left[ - \frac{p(x_2)}{1 - p(x_1)} \log \left( \frac{p(x_2)}{1 - p(x_1)} \right) - \frac{p(x_3)}{1 - p(x_1)} \log \left( \frac{p(x_3)}{1 - p(x_1)} \right) \right] & \\
= -p(x_1) \log p(x_1) - (1 - p(x_1)) \log (1 - p(x_1)) & \\
- p(x_2) \log \left( \frac{p(x_2)}{1 - p(x_1)} \right) - p(x_3) \log \left( \frac{p(x_3)}{1 - p(x_1)} \right) & \\
= -p(x_1) \log p(x_1) - p(x_2) \log p(x_2) - p(x_3) \log p(x_3) & \\
- (1 - p(x_1)) \left[ p(x_2) - p(x_3) \right] \log (1 - p(x_1)) & \\
= - \sum_{i=1}^{3} p(x_i) \log p(x_i) & \text{since } \sum_{i=1}^{3} p(x_i) = 1 \\
= H(p(x_1), p(x_2), p(x_3))
\end{align*}
\]

The proof for a general \( n \) can be done through induction. A general proof (including uniqueness of \( H(X) \)) can be found in Shannon [1].

### 2.1.3 Alternative approach

An alternative approach is by first defining the information in a single symbol, the so-called self-information, and then average this self-information over all the symbols, taking into account their probability of occurrence.
Introduction to self-information

Consider a source that can generate \( i = 1, \ldots, n \) symbols, \( x_i \), and let \( p(x_i) \) be the probability that the \( i \)th message is selected for transmission. One associates the self-information to the \( i \)th message as function of its probability of occurrence \( S(x_i) = f(p(x_i)) \). This function needs to fulfill the following properties

1. The information of each message needs to contribute in a non-negative way
   \[ 0 \leq f(p(x_i)) \]

2. If the \( i \)th message is selected surely, then there is no information present in this message since we known that the \( i \)th message will be selected anyway. Hence,
   \[ \lim_{p(x_i) \to 1} f(p(x_i)) = 0 \]

3. The lower the probability that the message is selected, the higher the information gained from that selection
   \[ f(p(x_i)) > f(p(x_j)) \text{ if } p(x_i) < p(x_j) \]

4. Consider a message \( x_i, x_j \) that is composed out of two independent messages \( x_i \) and \( x_j \). The probability of this composed message equals \( p(x_i)p(x_j) \) since \( P[A \cap B] = P[A]P[B] \) if \( A \) and \( B \) are independent. If two messages are independent, then we expect that the total information equals the sum of the two messages:
   \[ f(p(x_i, x_j)) = f(p(x_i)) + f(p(x_j)) \]

One and only one function can obey to the above conditions: the logarithm \( f(x) = -\log(x) \). The above four conditions can easily be verified:

1. A probability is bounded between zero and one, \( 0 \leq p \leq 1 \), and hence
   \[ -\log(p) \geq 0 \]

2. \(-\log(1) = 0\)

3. \(-\log(x) > -\log(y) \text{ if } x < y\)

4.
   \[
   f(p(x_i, x_j)) = -\log(p(x_i, x_j)) \\
   = -\log(p(x_i)p(x_j)) \\
   = -\log p(x_i) - \log p(x_j) \\
   = f(p(x_i)) + f(p(x_j))
   \]

The self-information of the symbol \( x_i \) is therefore defined as
\[
S(x_i) = -\log(p(x_i))
\]
Averaging the self-information of the different symbols in a message

Auto-information has been defined in function of the individual messages, or symbols that a source can produce. Communication systems are not intended to transmit one single message only, but must to be able to send all possible messages. It is therefore necessary to distinction between what a source is producing at a certain moment versus what it could possibly generate. Hence, it is recommended to describe the source in terms of the average information that it produces. This average information defines the entropy of a source.

Consider a discrete source $X$ with $n$ statistically independent symbols, where $n$ is the length of the alphabet. When the $i^{th}$ symbol is selected for transmission, this will yield $S_i$ information. A long message with a large number of symbols, $N \gg$, contains the $i^{th}$ symbol on average $Np(x_i)$ times. The total information of this long message is therefore given by

$$\sum_{i=1}^{n} Np(x_i)S(x_i)$$

When this total information is divided by $N$, the average or mean information per symbol equals

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

$H(X)$ is the entropy of the source expressed in $\text{bit/symbol}$. The defined average is the expected value with respect to the probability space (population mean). We assume that the sources are stationary. When the source would be non-stationary, the probabilities of the symbols can vary in function of time.

### 2.1.4 Interpretation of entropy

The entropy $H$ can be interpreted as follows: Although one is not be able to predict without uncertainty which will be the next symbol that a source will generate, it is known that on the average $H$ bit per symbol are made available by the source. Alternatively, if the source produces symbols that are statistically independent and if one has observed $N \gg$ symbols, then about $NH$ bits of information are acquired.

Another interpretation links the probability of a long message of a large number independent symbols, $N \gg$. This long message will contain with high probability about $Np(x_1)$ occurrences of the first symbol $x_1$, $Np(x_2)$ occurrences of the second $x_2$, etc. Hence, the probability of this particular message will be roughly equal to

$$p = p(x_1)^{Np(x_1)}p(x_2)^{Np(x_2)} \ldots p(x_n)^{Np(x_n)}$$
or
\[
\log p = N \sum_{i=1}^{n} p(x_i) \log p(x_i) = -NH(X)
\]

Hence, the probability of this particular message equals
\[
p = 2^{-NH(X)}.
\]

### 2.1.5 Binary entropy function

The binary entropy function \(H_b(p)\) is defined as the entropy of a Bernoulli process. The Bernoulli process is modeled as a random variable \(X\) that can take on only \(n = 2\) values: 0 and 1 with the probability of success \(P[X = 1] = p\) and probability of failure of \(P[X = 0] = 1 - p\).

The entropy is then given by

\[
H(X) = H_b(p) = -p \log p - (1 - p) \log(1 - p)
\]

\[
= p \log \frac{1 - p}{p} - \log(1 - p)
\]

\[
= \frac{1}{\ln 2} \left( p \log \frac{1 - p}{p} - \ln(1 - p) \right)
\]

![Figure 2.3: Entropy \(H_b(p)\) for a Bernoulli process with probability \(p\).](image)
The entropy equals zero when $p = 0$ or $p = 1$. In those two cases, one is sure of always having respectively a 0 or a 1 as output of the process. This implies a deterministic output and hence an entropy equal to zero.

A Bernoulli process has a maximal entropy for $p = 1/2$ with $H_b(1/2) = 1 \text{ bit/symbol}$. This can easily be proven since

$$\frac{\partial H_b(p)}{\partial p} = \frac{1}{\ln 2} \left( \ln \frac{1-p}{p} + \frac{p}{1-p} \frac{-p - (1-p)}{p^2} - \frac{1}{1-p} \right)$$

which equals zero if $p = 1/2$. This implies that the maximum entropy of the binary entropy function equals $H_b(1/2) = 1$ which basically means that the entropy of a binary process equals 1 bit/symbol bit if both states (0 and 1) are equally probable ($p = 1/2$). If the probability is e.g. $p = 0.2$, then the entropy is reduced to $H_b(0.2) = 0.72 \text{ bit/symbol}$.

### 2.1.6 Entropy for alphabet with length $n$

The entropy of a discrete source with alphabet length $n$ depends on the probabilities of the symbols. It has a lower bound and upper bound

$$0 \leq H(X) \leq \log n$$

The lower bound (0) expresses that the source on the average does not produce any information. An entropy $H(X) = 0$ implies that there is no uncertainty about the generated symbols. This can only be associated to a source that repeatedly generates the same symbol, i.e. all probabilities $p(x_i)$ are zero except for a single symbol $j$ where $p(x_j) = 1$.

The upper bound of the entropy corresponds to maximum uncertainty about which symbol a source will produce next. Hence, it is attained if the symbols are independent and uniformly distributed, i.e. with equal probability: $p(x_i) = 1/n$ for all $i$. This implies that

$$H_{max}(X) = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} = \log n$$

#### Examples of alphabets

- Binary symbols ($n = 2$): $H_{max} = H_b(1/2) = 1 \text{ bit/symbol}$
- Roll of dice ($n = 6$): $H_{max} = 2.585 \text{ bit/symbol}$
- Decimal numbers ($n = 10$): $H_{max} = 3.32 \text{ bit/symbol}$
- Morse symbols (characters + space) ($n = 27$): $H_{max} = 4.755 \text{ bit/symbol}$
2.2 Information rate (=entropy rate) of independent sources

The entropy $H(X)$ measures the amount of information/entropy within a single symbol $X$. Hence, it is expressed in bit/symbol. However, the study of the rate at which data is transferred requires another concept. This concept is the information rate (also known as the entropy rate).

2.2.1 General definition of information rate

The information rate can be defined in a general way by considering the duration of the signal that is used to represent each symbol $x_i$. Take e.g. Morse code (Section 2.4.3) where each of the character is represented by a series of dashes and dots, and where the timing of the dashes and dots are not necessary synchronized with a central clock. Hence, the signal that represents $i^{th}$ symbol $x_i$ in Morse has a variable duration $\tau(x_i)$.

**Definition 2** (Information rate). The information rate $R(X)$ equals the self-information $S(x_i)$ of each symbol $x_i$ (in bit), multiplied with its average frequency of occurrences, $f(x_i)$ (in s$^{-1}$)

$$R(X) = \sum_{i=1}^{n} f(x_i) S(x_i)$$

$$= -\sum_{i=1}^{n} f(x_i) \log p(x_i)$$

The average frequency of occurrences $f(x_i)$ can be expressed as the ratio of the probability of a symbol $p(x_i)$ divided by the average duration of a symbol, $\langle \tau \rangle$,

$$f(x_i) = \frac{p(x_i)}{\langle \tau \rangle}$$

where

$$\langle \tau \rangle = \sum_{i=1}^{n} p(x_i) \tau(x_i)$$

represents the average duration of a symbol with $\tau(x_i)$ the duration of the $i^{th}$ symbol $x_i$.

This enable the reformulation of the information rate using the entropy and the average duration of a symbol since

$$R(X) = -\sum_{i=1}^{n} f(x_i) \log p(x_i)$$

$$= -\sum_{i=1}^{n} \frac{p(x_i)}{\langle \tau \rangle} \log p(x_i)$$

$$= \frac{1}{\langle \tau \rangle} H(X)$$
2.2.2 Information rate for binary coders

Symbols \( x_i \) are often translated into a sequence of \( N(x_i) \) bits, where the number of bits \( N(x_i) \) can be variable. We are then speaking about a variable length code. As example, consider an alphabet with \( n = 4 \), and consider that \( x_1 \) is represented by transmitting a single bit \('0'\), \( x_2 \) by transmitting \('10'\), \( x_3 \) by transmitting \('110'\), and \( x_4 \) by transmitting \('111'\). In this case we have that \( N(x_1) = 1, N(x_2) = 2, N(x_3) = 3, \) and \( N(x_4) = 3 \). Hence, it is as if we transformed our discrete source \( X \) into an equivalent binary source which represents each \( x_i \) with a (variable) number of \( N(x_i) \) bits.

A binary coder often transmits its data with a fixed sampling period \( T_s \). In this situation, it is possible to have alternative definition of the information rate which is related to the sampling period \( T_s \), namely

\[
\tau(x_i) = N(x_i)T_s \\
\langle \tau \rangle = \langle N \rangle T_s
\]

with \( \langle N \rangle \) the average value

\[
\langle N \rangle = \sum_{i=1}^{n} p(x_i)N(x_i)
\]

Hence,

\[
R(X) = \frac{1}{\langle \tau \rangle} H(X) = \frac{1}{\langle N \rangle T_s} H(X)
\]

This makes it possible to define a second kind of information rate, namely an information rate

\[
R(X) = \frac{1}{\langle N \rangle} H(X)
\]

expressed as \( \text{bit/symbol} \) instead of \( \text{bit/s} \). This representation is often used when the efficiency of data compression is studied.

2.3 Redundancy of independent discrete sources

Redundancy measures the fractional difference between the information rate \( R(X) \) of a source \( X \), and its maximum possible information rate \( R_{\text{max}}(X) \). This maximum possible value equals \( \log n / \langle \tau \rangle \) if the symbols are independent and uniformly distributed.

**Definition 3** (Redundancy and efficiency of a discrete source). The **absolute redundancy** of a discrete source equals

\[
R_X = R_{\text{max}}(X) - R(X)
\]
which is expressed in bit/s (or bit/sample for binary sources). The relative redundancy is defined as

\[ r_X = 1 - \frac{R(X)}{R_{\text{max}}(X)} \]

and can be expressed in percentages since \( 0 \leq r_X \leq 1 \).

Complementary to the concept of relative redundancy is the concept of efficiency which is defined as

\[ e_X = \frac{R(X)}{R_{\text{max}}(X)} \]

and which is also expressed in percentages \( (0 \leq e_X \leq 1) \).

2.4 Examples of independent discrete source

The following examples are provided to clarify the theory:

1. An academic example to demonstrate the concepts on a fixed and variable length binary coder;
2. The letter frequency in different languages;
3. Morse code to show the power of variable length coding techniques.

2.4.1 Academic example

Consider a discrete source \( X \) with an alphabet of length \( n = 4 \) for which the probabilities for the occurrences of the symbols are given by \( \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8} \). Assume furthermore that the source generates a binary sequence with one sample every \( T_s = 5 \text{ ms} \). Hence, up to \( 200 \text{ bit/s} \) can be transmitted (if the bits are uniformly distributed). This maximum information rate is also known as the channel capacity \( C = 200 \text{ bit/s} \). In a later chapter, it will be shown that it is possible to introduce coding to approach this information rate. This section aims to illustrate this with a simple example.

Fixed length code (= block code)

A first proposal is to use a binary code as shown in Table 2.1. The entropy of this source equals

\[ H(X) = \frac{1}{2} \log 2 + \frac{1}{4} \log 4 + \frac{1}{8} \log 8 + \frac{1}{8} \log 8 \]

\[ = 1.75 \text{ bit/symbol} \]

The average symbol length of the binary source expressed in number of bits/samples equals \( \langle N \rangle = 2 \text{ sample/symbol} \) samples/symbol. The average time duration of a
symbol equals \( \langle \tau \rangle = 10 \text{ ms/symbol} \). The information rate therefore equals

\[
R(X) = \frac{H(X)}{\langle N \rangle} = \frac{1.75}{2} \text{ bit/sample}
\]

or

\[
R(X) = \frac{H(X)}{\langle \tau \rangle} = 175 \text{ bit/s}.
\]

<table>
<thead>
<tr>
<th>(i)</th>
<th>code</th>
<th>( p(x_i) )</th>
<th>( S(x_i) ) (bit)</th>
<th>( N(x_i) ) (sample/symbol)</th>
<th>( \tau(x_i) ) (ms/symbol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>00</td>
<td>( 1/2 )</td>
<td>1</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>01</td>
<td>( 1/4 )</td>
<td>2</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>( 1/8 )</td>
<td>3</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>( 1/8 )</td>
<td>3</td>
<td>2</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 2.1: Binary code with fixed length.

Based on these values, one can conclude that the encoding technique is not optimal: the data rate of binary encoded source is 0.875 bit/sample which is less than the optimal rate for binary encoding, i.e. 1 bit/sample. Furthermore, the information rate of 175 bit/s is also smaller than the channel capacity of 200 bit/s.

The difference can be partly explained by the fact that the binary digits '0' and '1' do not have the same probability. Consider therefore a large number of symbols, \( N \). The number of binary digits '0' will then converge (for \( N \to \infty \)) towards

\[
2\frac{N}{2} + 1\frac{N}{4} + 1\frac{N}{8} + 0\frac{N}{8} = \frac{11}{8}N
\]

while the total number of binary digits converges to

\[
2\frac{N}{2} + 2\frac{N}{4} + 2\frac{N}{8} + 2\frac{N}{8} = \frac{16}{8}N
\]

The probability of a binary digit '0' therefore equals \( P(0') = \frac{11}{16} \). Using a similar reasoning for binary digit '1' can be used to proof that \( P(1') = \frac{5}{16} \). Hence, the binary source from the encoder is not running at its maximum entropy, as this would require that \( P(0') = P(1') = 1/2 \).

Furthermore, it is possible to compute

- the absolute redundancy \( R_X = 200 - 175 = 25 \text{ bit/s} \) or \( 1 - 0.875 = 0.125 \text{ bit/sample} \),
- the relative redundancy \( r_X = 1 - 175/200 = 0.125 \), and
- the efficiency \( e_X = 175/200 = 0.875 \).
Variable length code

Consider the variable length code as shown in Table 2.2. The entropy of this source is still equal to $H(X) = 1.75 \text{bit/symbol}$, but the average length $\langle N \rangle$ and duration $\langle \tau \rangle$ of a symbol now equals

$$\langle N \rangle = \frac{1}{2}1 + \frac{1}{4}2 + \frac{2}{8}3 = \frac{14}{8} = 1.75 \text{sample/symbol}$$

$$\langle \tau \rangle = \langle N \rangle T_s = \frac{14}{8}5 = 8.75 \text{ms/symbol}.$$

| $i$ | code $| p(x_i) | S(x_i) \text{ (bit)} | N(x_i) \text{ (sample/symbol)} | \tau(x_i) \text{ (ms/symbol)} |
|----|---|---|---|---|
| 1  | 0  | $\frac{1}{2}$ | 1  | 1  | 5  |
| 2  | 10 | $\frac{1}{4}$ | 2  | 1  | 10 |
| 3  | 110| $\frac{1}{8}$ | 3  | 3  | 15 |
| 4  | 111| $\frac{1}{8}$ | 3  | 3  | 15 |

Table 2.2: Binary code with variable length

The information rate $R(X)$ therefore equals

$$R(X) = \frac{H(X)}{\langle N \rangle} = 1 \text{bit/sample}$$

or

$$R(X) = \frac{H(X)}{\langle \tau \rangle} = 200 \text{bit/s}.$$  

One can conclude that this is an optimal code since the number of produced binary symbols per second (bit/s) corresponds exactly to the information rate of the source that has to be encoded.

Consider once more a large number of symbols, $N$. The number of binary digits ‘0’ will then converge (for $N \rightarrow \infty$) towards

$$\frac{N}{2} + \frac{N}{4} + \frac{N}{8} + \frac{N}{8} = \frac{7}{8}N$$

while the total number of binary digits converges to

$$\frac{N}{2} + \frac{2N}{4} + \frac{3N}{8} + \frac{3N}{8} = \frac{14}{8}N$$

The probability of a binary digit ‘0’ therefore equals $P[0'] = \frac{1}{2}$. Using a similar reasoning for binary digits ‘1’ can be used to proof that $P[1'] = \frac{1}{2}$.

Note that in this case, the binary digits ‘0’ and ‘1’ have the same probability (equi-probable symbols). Moreover, the alternative coding has been formulated such that the resulting binary representation (code) is the shortest for the most probable symbol of the source. The search for methods to yield optimal codes will be continued in Chapter 6.
2.4.2 Letter/symbol frequency in different languages

Consider that the different letters/symbols within a language are independent and that each letter from 'a' to 'z' represent a symbol. The letter frequency in larger texts (= the frequency or probability that a certain letter occurs [8]) enabled various applications. A first application is to use this letter frequency analysis as a rudimentary technique for language identification. Other applications that use the letter frequency are typesetting systems (such as the Linotype machine) and communication systems such as morse code.

Consider first the symbol frequency of three different languages, English, Dutch and French. The symbol probability (originating from [8]) and the self-information for each symbol \((-\log p(x_i))\) are shown in Figure 2.4 and Table 2.3.

![Graph showing symbol probability for English, Dutch, and French.](image)

Figure 2.4: Symbol probability for English, Dutch and French.
## Table 2.3: Letter probability for English, Dutch, and French together with the self-information of each letter.

<table>
<thead>
<tr>
<th>i</th>
<th>Letter</th>
<th>English $x_i$</th>
<th>Dutch $x_i$</th>
<th>French $x_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>0.082</td>
<td>0.075</td>
<td>0.079</td>
</tr>
<tr>
<td>2</td>
<td>b</td>
<td>0.015</td>
<td>0.016</td>
<td>0.009</td>
</tr>
<tr>
<td>3</td>
<td>c</td>
<td>0.028</td>
<td>0.012</td>
<td>0.034</td>
</tr>
<tr>
<td>4</td>
<td>d</td>
<td>0.043</td>
<td>0.059</td>
<td>0.038</td>
</tr>
<tr>
<td>5</td>
<td>e</td>
<td>0.127</td>
<td>0.189</td>
<td>0.151</td>
</tr>
<tr>
<td>6</td>
<td>f</td>
<td>0.022</td>
<td>0.008</td>
<td>0.011</td>
</tr>
<tr>
<td>7</td>
<td>g</td>
<td>0.020</td>
<td>0.034</td>
<td>0.009</td>
</tr>
<tr>
<td>8</td>
<td>h</td>
<td>0.061</td>
<td>0.024</td>
<td>0.008</td>
</tr>
<tr>
<td>9</td>
<td>i</td>
<td>0.070</td>
<td>0.065</td>
<td>0.077</td>
</tr>
<tr>
<td>10</td>
<td>j</td>
<td>0.002</td>
<td>0.015</td>
<td>0.006</td>
</tr>
<tr>
<td>11</td>
<td>k</td>
<td>0.008</td>
<td>0.022</td>
<td>0.001</td>
</tr>
<tr>
<td>12</td>
<td>l</td>
<td>0.040</td>
<td>0.036</td>
<td>0.056</td>
</tr>
<tr>
<td>13</td>
<td>m</td>
<td>0.024</td>
<td>0.022</td>
<td>0.031</td>
</tr>
<tr>
<td>14</td>
<td>n</td>
<td>0.067</td>
<td>0.100</td>
<td>0.073</td>
</tr>
<tr>
<td>15</td>
<td>o</td>
<td>0.075</td>
<td>0.061</td>
<td>0.060</td>
</tr>
<tr>
<td>16</td>
<td>p</td>
<td>0.019</td>
<td>0.016</td>
<td>0.026</td>
</tr>
<tr>
<td>17</td>
<td>q</td>
<td>0.001</td>
<td>0.000</td>
<td>0.014</td>
</tr>
<tr>
<td>18</td>
<td>r</td>
<td>0.060</td>
<td>0.064</td>
<td>0.069</td>
</tr>
<tr>
<td>19</td>
<td>s</td>
<td>0.063</td>
<td>0.037</td>
<td>0.082</td>
</tr>
<tr>
<td>20</td>
<td>t</td>
<td>0.091</td>
<td>0.068</td>
<td>0.075</td>
</tr>
<tr>
<td>21</td>
<td>u</td>
<td>0.028</td>
<td>0.020</td>
<td>0.065</td>
</tr>
<tr>
<td>22</td>
<td>v</td>
<td>0.010</td>
<td>0.028</td>
<td>0.019</td>
</tr>
<tr>
<td>23</td>
<td>w</td>
<td>0.024</td>
<td>0.015</td>
<td>0.001</td>
</tr>
<tr>
<td>24</td>
<td>x</td>
<td>0.002</td>
<td>0.000</td>
<td>0.004</td>
</tr>
<tr>
<td>25</td>
<td>y</td>
<td>0.020</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>26</td>
<td>z</td>
<td>0.001</td>
<td>0.014</td>
<td>0.003</td>
</tr>
</tbody>
</table>

The most frequent symbols in Dutch are the characters ‘e’ (with 18.9% frequency), the ‘n’ (about 10%) and the ‘a’, ‘t’, and ‘i’ (each with a probability around 7%).

This is completely different in English: the ‘e’ is still the most frequent symbol but the probability is only 12.7%. The ‘e’ is followed by the ‘t’ (9.1%), the ‘a’ (8.2%) and the ‘o’, ‘i’, and ‘n’ (each with a probability around 7%).

The probabilities and the self-information make it possible to compute the entropy for each language using $H(X) = \sum_i p(x_i)S(x_i)$. The numerical results can be found in Table 2.4.
If all symbols would have the same probability, namely \( \forall i, p_i = \frac{1}{26} \), then the entropy \( H_{max} = \log 26 = 4.7004 \text{ bit/symbol} \). This makes it possible to compute the redundancy \( R_X = H_{max} - H(X) \) and the efficiency \( e_X = H(X)/H_{max} \) as shown in Table 2.4.

Figure 2.5 compares the probability of the uniform distribution to the sorted probability of occurrence for symbols in English.

Figure 2.5: Comparison a plain English text with a source of the same alphabet running at its maximum entropy. The symbols have been ordered from highest to lowest probability of occurrence to be able to compare them with the uniform distribution.

### 2.4.3 Morse code

Morse code encodes the most frequent letters with the shortest symbols. The Morse alphabet arranges the letters into groups of letters that require equal amounts of time to transmit, and then sorting these groups in increasing order, yielding the sequence
Similar ideas are used in modern data-compression techniques such as Huffman coding.

It can be of interest to check to what extent the Morse code approaches the optimal code, although it has been developed about a century before the coding theorems in the Information Theory were published.

The Morse code is a method of transmitting text information as a series of on-off tones, lights, or clicks that can be directly understood by a skilled listener or observer without special equipment. The International Morse Code encodes the ISO basic Latin alphabet, some extra Latin letters, the Arabic numerals and a small set of punctuation and procedural signals as standardized sequences of short and long signals called "dots" and "dashes". The International Morse Code [9] is represented in Figure 2.6.

**International Morse Code**

1. The length of a dot is one unit.
2. A dash is three units.
3. The space between parts of the same letter is one unit.
4. The space between letters is three units.
5. The space between words is seven units.

![International Morse Code Diagram](image)

Figure 2.6: The International Morse Code [9].

The previous section analyzed the probability that a symbol from our alphabet occurs in an English text. The list of symbols is now extended with a blank (also called space) as these are used to separate words. It is also possible
CHAPTER 2. DISCRETE COMMUNICATION SYSTEMS

28
to represent numbers and punctuation marks with the Morse code but these are not considered in this study.

First, an attempt will be made to transform the asynchronous character of the transmission from the Morse code into a synchronous one, i.e. in a practical signal of which the duration of the symbols can be related to a clock. The Morse code consists of dots and dashes, as can be observed in Figure 2.6. Assume that a dot is composed of closing and opening a switch and that both actions take the same amount of time. Hence, a dot therefore takes two time units. A dash closes the switch three times longer than a dot and is therefore represented by four time units (three from the dash and one for opening the switch; i.e. the Morse key). A separation of the different characters within one word is represented by three time units (the switch remains open). Hence, the length of a regular character (time units) equals

\[\text{#dots} \times 2 + \text{#dashes} \times 4 + 3\]

A word is separated from another word by six time units (blank).

The asynchronous transmission of the telegraph signal can be approached by a synchronous one using the above assumptions. Consequently, the duration of each symbol \(\tau(x_i)\) can be determined as shown in Table 2.5. This table can then be used to compute

- the entropy of the source \(H(X) = \sum p(x_i)S(x_i) = 4.073\text{ bit/symbol}\),
- the average duration of a symbol \(\langle \tau \rangle = \sum_i p(x_i)\tau(x_i) = 9.2662\text{ time unit/symbol}\),
- the information rate \(R(X) = \frac{H(X)}{\langle \tau \rangle} = 0.4396\text{ bit/time unit}\),
- the absolute redundancy \(R_X = \frac{\log 27}{\langle \tau \rangle} - R(X) = 0.0736\text{ bit/symbol}\),
- the relative redundancy \(r_X = 1 - H(X) / \log 27 = 0.1434\), and
- the efficiency \(e_X = H(X) / \log 27 = 0.8566\).
CHAPTER 2. DISCRETE COMMUNICATION SYSTEMS

One can compare this result now with the direct binary encoding. In order to represent binary 27 different symbols, a 5-bit representation is sufficient ($2^4 < 27 < 2^5$). The 5-bit code is a code with fixed length and is therefore a block code with a length of 5. As each dot is represented by 2 time units, one can assume that a single block code of 5-bit takes 10 time units.

This implies that

- the entropy of the source $H(X) = \sum p(x_i)S(x_i) = 4.073\text{ bit/symbol}$, as it is independent of the binary coding used,

- the average symbol rate $\langle N \rangle = 5\text{ bit/symbol},$

- the average duration of a symbol $\langle \tau \rangle = 10\text{ time unit/symbol},$

- the information rate $R(X) = \frac{H(X)}{\langle N \rangle} = 0.817\text{ bit/sample}$ or $R(X) = \frac{H(X)}{\langle \tau \rangle} = 0.407\text{ bit/time unit},$

### Table 2.5: Letter frequency, Morse code and its duration.

<table>
<thead>
<tr>
<th>i</th>
<th>Letter</th>
<th>$p(x_i)$</th>
<th>$S(x_i)$</th>
<th>$p(x_i)S(x_i)$</th>
<th>code</th>
<th>$\tau(x_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>0.065</td>
<td>3.939</td>
<td>0.257</td>
<td>-</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>b</td>
<td>0.012</td>
<td>6.334</td>
<td>0.079</td>
<td>-...</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>c</td>
<td>0.022</td>
<td>5.526</td>
<td>0.120</td>
<td>--</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>d</td>
<td>0.035</td>
<td>4.837</td>
<td>0.169</td>
<td>-...</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>e</td>
<td>0.104</td>
<td>3.264</td>
<td>0.095</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>f</td>
<td>0.020</td>
<td>5.659</td>
<td>0.112</td>
<td>-...</td>
<td>13</td>
</tr>
<tr>
<td>7</td>
<td>g</td>
<td>0.016</td>
<td>5.975</td>
<td>0.095</td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>8</td>
<td>h</td>
<td>0.049</td>
<td>4.342</td>
<td>0.214</td>
<td>...</td>
<td>11</td>
</tr>
<tr>
<td>9</td>
<td>i</td>
<td>0.056</td>
<td>4.164</td>
<td>0.232</td>
<td>..</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>j</td>
<td>0.001</td>
<td>10.118</td>
<td>0.009</td>
<td>-...</td>
<td>17</td>
</tr>
<tr>
<td>11</td>
<td>k</td>
<td>0.005</td>
<td>7.615</td>
<td>0.039</td>
<td>-...</td>
<td>13</td>
</tr>
<tr>
<td>12</td>
<td>l</td>
<td>0.033</td>
<td>4.917</td>
<td>0.163</td>
<td>-...</td>
<td>13</td>
</tr>
<tr>
<td>13</td>
<td>m</td>
<td>0.020</td>
<td>5.630</td>
<td>0.114</td>
<td>--</td>
<td>11</td>
</tr>
<tr>
<td>14</td>
<td>n</td>
<td>0.056</td>
<td>4.146</td>
<td>0.234</td>
<td>--</td>
<td>9</td>
</tr>
<tr>
<td>15</td>
<td>o</td>
<td>0.060</td>
<td>4.069</td>
<td>0.242</td>
<td>-...</td>
<td>15</td>
</tr>
<tr>
<td>16</td>
<td>p</td>
<td>0.014</td>
<td>6.179</td>
<td>0.085</td>
<td>-...</td>
<td>15</td>
</tr>
<tr>
<td>17</td>
<td>q</td>
<td>0.001</td>
<td>10.118</td>
<td>0.009</td>
<td>-...</td>
<td>17</td>
</tr>
<tr>
<td>18</td>
<td>r</td>
<td>0.050</td>
<td>4.328</td>
<td>0.216</td>
<td>..</td>
<td>11</td>
</tr>
<tr>
<td>19</td>
<td>s</td>
<td>0.052</td>
<td>4.277</td>
<td>0.221</td>
<td>...</td>
<td>9</td>
</tr>
<tr>
<td>20</td>
<td>t</td>
<td>0.073</td>
<td>3.778</td>
<td>0.275</td>
<td>--</td>
<td>7</td>
</tr>
<tr>
<td>21</td>
<td>u</td>
<td>0.022</td>
<td>5.474</td>
<td>0.123</td>
<td>-...</td>
<td>11</td>
</tr>
<tr>
<td>22</td>
<td>v</td>
<td>0.008</td>
<td>6.913</td>
<td>0.057</td>
<td>-...</td>
<td>13</td>
</tr>
<tr>
<td>23</td>
<td>w</td>
<td>0.017</td>
<td>5.870</td>
<td>0.100</td>
<td>-...</td>
<td>13</td>
</tr>
<tr>
<td>24</td>
<td>x</td>
<td>0.001</td>
<td>9.481</td>
<td>0.013</td>
<td>-...</td>
<td>15</td>
</tr>
<tr>
<td>25</td>
<td>y</td>
<td>0.015</td>
<td>6.098</td>
<td>0.089</td>
<td>-...</td>
<td>17</td>
</tr>
<tr>
<td>26</td>
<td>z</td>
<td>0.001</td>
<td>10.288</td>
<td>0.008</td>
<td>-...</td>
<td>15</td>
</tr>
<tr>
<td>27</td>
<td>space</td>
<td>0.192</td>
<td>2.382</td>
<td>0.457</td>
<td></td>
<td>6</td>
</tr>
</tbody>
</table>
• the absolute redundancy $R_X = \frac{5}{\langle \tau \rangle} - R(X) = 0.0930 \text{bit/time unit},$

• the relative redundancy $r_X = 1 - \frac{4.073}{5} = 0.1854,$ and

• the efficiency $e_X = \frac{4.073}{5} = 0.8146 = 81.46\%.$

The efficiency of the direct binary encoding is lower than the one obtained from the Morse code. Hence, the Morse code is outperforming the block code with length 5. However, the Morse code is not optimal since the least common character 'z' ($p = 0.0008$) has a shorter duration of 15 than 'j' or 'q' ($p = 0.0009$) which have duration of 17.

Later in the course, we will study the Huffman coding that encodes each symbol in a uniquely decodable binary symbol with variable length (Section 6.4.3). This encoding starts with the probabilities of Table 2.5 and results into a binary code with a code length ranging from 3 to 10 bit, and with an average symbol rate $\langle N \rangle = 4.113 \text{bit/symbol}.$ The information rate of the Huffman coding is therefore $R(X) = H(X)/\langle N \rangle = 0.990 \text{bit/sample}$ or $R(X) = H(X)/\langle \tau \rangle = 0.495 \text{bit/time unit}$ which is better than the information rate of $R(X) = 0.4396 \text{bit/time unit}$ for Morse code and which is close to the $1 \text{bit/sample}$ that is to be expected for an optimal binary encoding.

2.5 Markov processes

A formalism is necessary to capture the dependence of the symbols over time. However, we need to limit the class of processes somehow in order to be able to describe the processes mathematically. This is done through the concept of discrete-time Markov gains. The properties of Markov processes are used in Section 2.6 to determine its entropy.

A discrete-time Markov chain is a sequence of random variables $X(k)$ which represent the state of the system at instance $k$. The discrete-time chain can change from one state $x_i(k-1)$ to another $x_j(k)$ for increasing time index $k$. If, additionally, the chain satisfies the Markov property, then the chain will be called a discrete-time Markov chain.

**Definition 4** (Markov property). The Markov property is fulfilled if the probability of moving to the next state (at time instance $k$) only depends only on the present state (at time instance $k-1$) and not on the previous states (at time instance before $k-1$). This can be expressed mathematically using a conditional probability

$$P \left( X(k) = x_j(k) | X(k-1) = x_{i_{k-1}}(k-1) \cap \ldots \cap X(1) = x_{i_1}(1) \right)$$

$$= P \left( X(k) = x_j(k) | X(k-1) = x_{i_{k-1}}(k-1) \right)$$

or abbreviated

$$P \left( x_j(k) | x_{i_{k-1}}(k-1) \cap \ldots \cap x_{i_1}(1) \right) = P \left( x_j(k) | x_{i_{k-1}}(k-1) \right).$$
As the transition from one state to another is only determined by the present state, it become possible to uniquely define the Markov process using a single matrix.

**Definition 5** (State-transition probability). The state-transition probability of Markov processes equals

\[ P_{ij}(k) = P(x_j(k)|x_i(k-1)) = P[X(k) = x_j(k)|X(k-1) = x_i(k-1)] \]

with

\[ \sum_j P_{ij}(k) = 1. \]

This can be represented by a so-called \( n \times n \) transition matrix \( \mathbf{P}(k) \) with the elements \( P_{ij}(k) \) which uniquely defines the behavior of the Markov process.

The Markov process does not required that \( \mathbf{P}(k) \) stays constant over time. However, if \( \mathbf{P}(k) \) stays constant, we say that the Markov process is stationary.

**Definition 6** (Stationary Markov process). A stationary Markov process (=time-homogeneous) has a state-transition probability that is constant over \( k \): \( P_{ij}(k) = P_{ij} \) or \( \mathbf{P}(k) = \mathbf{P} \) for all \( k \).

### 2.5.1 State diagram

A Markov chain is usually shown by a state transition diagram (sometimes abbreviated as state diagram). This is a directed graph where each state is represented by one node and where conditional probability of moving from one state to another is represented by a directed arcs.

Fig. 2.7 shows the general representation of a state diagram for \( n = 3 \) states \( x_1, x_2, x_3 \) of a stationary Markov process. The state-transition probabilities are shown next to the arrows and correspond to the transition matrix

\[
\mathbf{P} = \begin{bmatrix}
P_{11} & P_{12} & P_{13} \\
P_{21} & P_{22} & P_{23} \\
P_{31} & P_{32} & P_{33}
\end{bmatrix}
\]
Example 1. Binary differential coding

Differential coding strategies code the information in the difference of the signal. Hence, a differential coding needs to remember the previous symbol (=state) to be able to construct the transmission signal.

Consider a binary differential code where the probability of changing state equals $p$. This probability of change can come from the data source where the probability of e.g. a bit that equals one is $p$. The state transition matrix equals

$$ P = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix} $$

The state diagram of this Markov process is shown in Fig. 2.8.

![Figure 2.8: State diagram of (binary) differential coding.](image)

A more complex, although artificial example is the Markov model for the weather of the Land of Oz.

Example 2. Weather Markov model for the weather of the Land of Oz

Consider that the weather in the land of Oz can be one of $n = 3$ states: $x_1 = R$: rainy weather, $x_2 = N$: nice weather, and $x_3 = S$: snowy weather. The state diagram of a stationary Markov process that is characterized by its state-transition matrix

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

![Figure 2.8: State diagram of (binary) differential coding.](image)
is then given by Fig. 2.9.

Figure 2.9: State diagram of the land of Oz example.

Note that the sum over a row of $P$ equals one and that the sum of all the arrows that leave a state $x_i$ in the state diagram also equals one. Furthermore, some arrows are not drawn when the transition probability equals zero ($P_{22} = 0$ in this example). The lack of this transition from $x_2$ to itself also implies that it is impossible to have two days of nice weather in a row in the Land of Oz.

2.5.2 Trellis diagram

A trellis diagram is a graph whose nodes are ordered into vertical slices (time), and with each node at each time connected to at least one node at an earlier and at least one node at a later time. The earliest and latest times in the trellis have only one node. This implies that the trellis graph draws the state transitions (and their probabilities) from an initial state (the earliest time) to the final state (the latest time). Fig. 2.10.
Figure 2.10: Generic trellis diagram that starts in the initial state $x_1(1)$ and shows all the (probabilistic) transitions until the final state $x_2(5)$.

Trellises are used in encoders and decoders for communication theory and encryption and use a central rule in e.g. the Viterbi algorithm for Hidden Markov Models (also used in e.g. speech recognition). There, the trellis diagram is used to compute the probability that a certain sequence can occur and then determines the sequence that is the most likely (Maximum Likelihood principle).

The trellis for the weather of the land of Oz is shown in Fig. 2.11. Note that e.g. the transition from $x_2(k - 1)$ to $x_2(k)$ is non-existing (never two days of nice weather on a row).

Figure 2.11: Trellis example for the weather of the land of Oz.
2.5.3 Irreducible Markov chain

A Markov chain is said to be irreducible if it is possible to get to any state from any state. State $j$ is said to be accessible from a state $i$ if a system started in state $i$ has a non-zero probability of transitioning into state $j$ at some point.

**Definition 7** (Irreducible Markov chain). A Markov chain is irreducible if there exists an integer $n_{ij} > 1$ such that

$$P(X(n_{ij}) = x_j | X(1) = x_i) > 0$$

This integer is allowed to be different for each pair of states, hence the subscripts in $n_{ij}$.

Fig. 2.12 and 2.13 show some typical cases of reducible Markov chains.

![Reducible Markov chain](image1)

Figure 2.12: Reducible Markov chain since the state $x_2$ is unreachable when starting from $x_1$ or $x_3$ (and vice versa).

![Reducible Markov chain](image2)

Figure 2.13: Reducible Markov chain since neither $x_1$ nor $x_3$ are reachable when starting from state $x_2$. 
The probability that a stationary Markov process is in state $x_i$ is independent of $k$ due to the stationary. Hence, the probability of a certain state $x_i$ can be represented by $p(x_i) = \pi_i = p(x_i(k))$ for all $k$. These probabilities must satisfy the

$$\pi_j = p(x_j(k)) = \sum_{i=1}^{n} p(x_i(k-1))p(x_j(k)|x_i(k-1)) = \sum_{i=1}^{n} \pi_i P_{ij}$$

Hence,

$$\pi_j = \sum_{i=1}^{n} \pi_i P_{ij}$$

This can be expressed in matrix format by considering the vector $\pi$ containing the elements $\pi_i$

$$\pi = \pi P$$

which is an eigenproblem where the eigenvectors are the probabilities of the state. This can be seen by considering a generic eigenvalue problem with an eigenvector $e$ with eigenvalue $\lambda$

$$\lambda e = eP.$$

The main difference is the scaling of the eigenvectors. Indeed, the probability vector $\pi$ is a left eigenvector $e$ of the transition matrix $P$ with an eigenvalue of $\lambda = 1$, and is normalized using

$$\pi = \frac{e}{\sum_{i=1}^{n} e_i} \quad (2.5)$$

such that $\sum_{i=1}^{n} \pi_i = 1$.

Let $W$ be the matrix of eigenvectors (each vector normalized using a 2-norm) where each row is a left eigenvector of $P$, i.e. the (left) eigendecomposition with $WP = DW$ with $D$ the diagonal matrix of left eigenvalues

$$D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

with $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$. Hence,

$$P = W^{-1}DW.$$

Since $P$ is a row stochastic matrix, its largest left eigenvalue $\lambda_1 = 1$. If there is a unique stationary distribution (only one eigenvalue equal to one, or
1 = \lambda_1 > \lambda_2$, then the largest eigenvalue and the corresponding eigenvector \((W_{[1,:]})) are both unique. Furthermore we have that
\[
W_{[1,:]} P = \lambda_1 W_{[1,:]} = W_{[1,:]} 
\]
and hence that \(\pi = W_{[1,:]}\).

If the Markov chain is stationary, and hence \(P\) is constant over \(k\), then the \(l\)-step transition probability equals
\[
P^l_{ij}(k) = p(x_j(k)|x_i(k-l))
\]
which be computed as the \(l\)th-power of the transition matrix,
\[
P^l = W^{-1} D W W^{-1} DW ... W^{-1} DW = W^{-1} D^l W
\]
with
\[
D^l = \text{diag}(\lambda_1^l, \lambda_2^l, ..., \lambda_n^l)
\]
If there is a unique stationary distribution, then \(\lambda_1 > \lambda_2\) and hence
\[
\lim_{l \to \infty} D^l = \text{diag}(1, 0, ..., 0).
\]

2.5.4 Ergodic Markov processes

Among the possible discrete Markoff processes there is a group with special properties: the “ergodic” processes. We shall call the corresponding sources ergodic sources. Although a rigorous definition of an ergodic process is somewhat involved, the general idea is simple: every sequence realized by an ergodic (Markov) process has the same statistical properties. This implies that it is possible to determine averages of the ensemble of all possible sequences using an average over time using an (infinitely) long realization.

**Theorem 1** (Ergodic Markov chain). A finite-state irreducible Markov chain is ergodic if it has an aperiodic state, or more generally, if there is a number \(l\) such that any state can be reached from any other state in any number of steps greater than or equal to a number \(l\).

Fig. 2.14 show some typical examples of an ergodic Markov chain. The examples in Fig. 2.12 and 2.13 are non ergodic as they are reducible, while Fig. 2.15 shows an example of a non-ergodic Markov chain as it is periodic.
Lemma 1. An ergodic Markov process has a unique probability vector \( \pi \) such that \( \pi P = \pi \) is strictly positive. This corresponds to the (unique) left eigenvector which corresponds with the eigenvalue \( \lambda_1 = 1 \). All other eigenvalues are smaller than one \( 1 = \lambda_1 > \lambda_2 \geq \lambda_3 \geq \ldots \)

Proof. Consider a row vector \( p(x(1)) \) that represents a valid probability distribution of the initial state at \( k = 1 \). Since the eigenvectors of \( P \) span the complete space, we can write \( p(x(1)) \) as a linear combination of the eigenvectors \( W \) using a row vector \( a \)

\[
p(x(1)) = aW
\]

Considering \( k \) steps in the Markov process using the state transition matrix results in

\[
P^k = W^{-1}D^kW
\]

Starting from the arbitrary initial state vector \( x(1) \), it is possible to determine
the state probability after $k \to \infty$ steps using

$$p(x(\infty)) = \lim_{k \to \infty} p(x(1))P^k$$

$$= \lim_{k \to \infty} aWW^{-1}\lambda^k$$

$$= a \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} W$$

$$= a_1 W_{[1,:]}$$

Note that the vector of probabilities $p(x(\infty))$ only depends on $W_{[1,:]}$ as the effect of the single element $a_1$ is removed when scaling the probability vector such that the sum of its elements equals one. Additionally we observed previously that that $\pi = W_{[1,:]}$. This implies that $p(x(\infty)) = \pi$. \qed

The requirement that $\lambda_1 > \lambda_2$ is required to guarantee that $p(x(\infty))$ is determined by only. If two eigenvalues would equal one $1 = \lambda_1 = \lambda_2 \geq \lambda_3 \geq \ldots$, then all linear combinations of $W_{[1,:]}$ and $W_{[2,:]}$ will make up $p(x(\infty))$, implying that this asymptotic solution is non-unique.

**Example 3.** The weather of the Land of Oz is an ergodic process.

First assumed that an initial distribution satisfying $p(x(1)) = [0.8 \ 0.1 \ 0.1]$. The successive distribution can then be computed using $p(x(k+1)) = p(x(1))P^k$ (in matlab: x * P^k), results in

$$p(x(2)) = [0.4750 \ 0.2250 \ 0.3000]$$

$$p(x(3)) = [0.4250 \ 0.1938 \ 0.3813]$$

$$p(x(4)) = [0.4047 \ 0.2016 \ 0.3938]$$

When starting from the initial probabilities $p(x(1)) = [0.1 \ 0.2 \ 0.7]$ one obtains

$$p(x(2)) = [0.3250 \ 0.2000 \ 0.4750]$$

$$p(x(3)) = [0.3812 \ 0.2000 \ 0.4187]$$

$$p(x(4)) = [0.3953 \ 0.2000 \ 0.4047]$$

It can be seen that both cases converge towards $[0.4 \ 0.2 \ 0.4]$ irrespective of the initial $x(1)$. 
The asymptotic value can be computed using the Matlab function: \([V, D, U] = \text{eig}(P)\) returns full matrix \(W\) whose columns are the corresponding left eigenvectors such that \(U^T P = DU^T\). For the Land of Oz example, we obtain that

\[
D = \begin{bmatrix}
1 & 0 & 0 \\
0 & 0.25 & 0 \\
0 & 0 & -0.25
\end{bmatrix}
\]

and

\[
W = U^T = \begin{bmatrix}
-0.6667 & -0.3333 & -0.6667 \\
-0.7071 & 0 & 0.7071 \\
0.4082 & -0.8165 & 0.4082
\end{bmatrix}
\]

Considering the eigenvector which corresponds with the eigenvalue \(\lambda_1 = 1\) gives \(e = [-0.6667, -0.3333, -0.6667]\). Normalizing using 2.5 finally results in \(\pi = [0.4, 0.2, 0.4]\).

### 2.6 Entropy of an ergodic discrete Markov information sources

The properties of Markov processes introduced in the previous section are now used to determine the entropy of a Markov information source.

Consider a discrete source of a stationary Markov process. For each possible state \(x_i\) there will be a set of probabilities \(P_{ij}\) of producing the next state \(x_j\). Consider now the information generated when changing from state \(x_i\) towards the next state \(x_j\). This means the information is captured by the state transitions and hence in its state transition matrix. Take e.g. the weather announcement of today in the Land of Oz: the news/information about the weather for today is captured by the transition probabilities conditioned on the weather of yesterday.

The entropy that is generated when the Markov process is in state \(x_i\) can be seen as an entropy conditioned on that state

\[
H(X|x_i(k-1)) = -\sum_{j=1}^n p(x_j(k)|x_i(k-1)) \log p(x_j(k)|x_i(k-1))
\]

\[
= -\sum_{j=1}^n P_{ij} \log P_{ij}
\]

The entropy of the source will be defined as the average of this conditional entropy \(H(X|x_i(k-1))\) weighted in accordance with the probability of occurrence of the considered state \(\pi_i\).

\[
H(X) = \sum_{i=1}^n \pi_i H(X|x_i) = -\sum_{i=1}^n \sum_{j=1}^n \pi_i P_{ij} \log P_{ij}
\]
Definition 8 (Entropy Markov process). The entropy of a stationary and ergodic Markov process is given by

\[ H(X) = -\sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i P_{ij} \log P_{ij}. \]

Example 4. The entropy of the weather in the Land of Oz can be computed using (??) with

\[ -P_{ij} \log P_{ij} = \begin{bmatrix} 0.5 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0.5 \end{bmatrix} \]

and \( \pi = [0.4 \ 0.2 \ 0.4] \), resulting \( H(X|x_i) = [1.5 \ 1.0 \ 1.5] \) bit and hence an entropy of \( H(X) = 1.4 \) bit.

Definition 9 (Information rate Markov process). The information rate \( R \) expresses the entropy per second for a stationary and ergodic Markov process at a definite time rate

\[ R(X) = \sum_{i=1}^{n} f_i H(X|x_i) = -\sum_{i=1}^{n} \sum_{j=1}^{n} f_i P_{ij} \log P_{ij} \]

where \( f_i \) is the average frequency (occurrences per second) of state \( x_i \).

\( H(X) \) measures the amount of information generated by the source per symbol (in \( \text{bit/symbol} \)), while the information rate \( R \) measures the rate of information in bits per second (\( \text{bit/s} \)).

2.7 Discrete joint entropy, conditional entropy, and mutual information

The study of the information transfer over a channel requires new concepts that make it possible to measure the transfer of information from the input of the channel (\( X \)) to the output of the channel (\( Y \)). Hence, the concept of entropy will be extended towards two random process, \( X \) and \( Y \). The following concepts will be introduced:

joint entropy \( H(X,Y) \): an entropy considering both \( X \) and \( Y \) simultaneously;

conditional entropy \( H(X|Y) \) and \( H(Y|X) \): the entropy of one random variable conditioned on the other random variable;

mutual information \( I(X;Y) \): the amount of information obtained about one random variable, through the other random variable.
There are various dependencies between all the different kinds of entropy and mutual information. These relationships can be represented graphically in Fig. 2.16.

Figure 2.16: Diagram showing additive and subtractive relationships for various information measures associated with correlated variables $X$ and $Y$. The area contained by both circles is the joint entropy $H(X,Y)$. The circle on the left (red and violet) is the individual entropy $H(X)$, with the red being the conditional entropy $H(X|Y)$. The circle on the right (blue and violet) is $H(Y)$, with the blue being $H(Y|X)$. The violet is the mutual information $I(X;Y)$ [3].

2.7.1 Discrete joint entropy

The joint entropy is the natural extension of the entropy towards two random variables $X$ and $Y$.

**Definition 10** (Discrete joint entropy). The joint entropy of two random variables $X$ and $Y$ with alphabets $x_i$ for $i = 1, \ldots, n$ and $y_j$ for $j = 1, \ldots, m$ is defined by

$$H(X,Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log p(x_i, y_j)$$

with $p(x_i, y_j)$ the joint probability that $x_i$ and $y_j$ occur together. The definition of joint entropy can be extended to more than two random variables.

**Properties**

**Lemma 2.** $H(X, X) = H(X)$

**Proof.** This can be proven directly using

$$p(x_i, x_j) = P[X = x_i \cap X = x_j] = p(x_i)\delta_{ij}$$
and the definitions of $H(X, X)$ and $H(X)$

$$H(X, X) = - \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i) \delta_{ij} \log (p(x_i)\delta_{ij})$$

$$= - \sum_{i=1}^{n} p(x_i) \log p(x_i)$$

$$= H(X)$$

Lemma 3. $H(X, Y) = H(Y, X)$

Proof. Commutativity follows directly from the commutativity of probabilities $p(x_i, y_j) = p(y_j, x_i)$. □

Lemma 4. $X$ and $Y$ are statistically independent if and only if $H(X, Y) = H(X) + H(Y)$

Proof. This can easily be proven using by the fact that independent variable $x_i$ and $y_j$ imply that $p(x_i, y_j) = p(x_i)p(y_j)$ and hence,

$$H(X, Y) = - \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log p(x_i, y_j)$$

$$= - \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i)p(y_j) \log (p(x_i)p(y_j))$$

$$= - \sum_{j=1}^{m} p(y_j) \sum_{i=1}^{n} p(x_i) \log p(x_i) - \sum_{i=1}^{n} p(x_i) \sum_{j=1}^{m} p(y_j) \log p(y_j)$$

$$= - \sum_{i=1}^{n} p(x_i) \log p(x_i) - \sum_{j=1}^{m} p(y_j) \log p(y_j)$$

$$= H(X) + H(Y)$$

□

2.7.2 Discrete conditional entropy (=equivocation)

If $H(Y|X = x_i)$ is the entropy of the discrete random variable $Y$ conditioned on the discrete random variable $X$ taking a certain value $x_i$. The entropy $H(Y|X)$, which is conditioned over the complete random process $X$, is the result of averaging $H(Y|X = x_i)$ over all possible values $x_i$ that $X$ may take. Hence, the conditional entropy of $Y$ given $X$ is defined as the weighted sum of
CHAPTER 2. DISCRETE COMMUNICATION SYSTEMS

$H(Y|X = x_i)$ for each possible value of $x_i$ using $p(x_i)$ as the weights [1].

\[
H(Y|X) = \sum_{i=1}^{n} p(x_i) H(Y|X = x_i)
\]

\[
= - \sum_{i=1}^{n} p(x_i) \sum_{j=1}^{m} p(y_j|x_i) \log p(y_j|x_i)
\]

\[
= - \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i) p(y_j|x_i) \log p(y_j|x_i)
\]

\[
= - \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log p(y_j|x_i).
\]

**Definition 11** (Discrete conditional entropy). The conditional entropy of two random variables $X$ and $Y$ with alphabets $x_i$ for $i = 1, \ldots, n$ and $y_j$ for $j = 1, \ldots, m$ is defined by

\[
H(Y|X) = - \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log p(y_j|x_i)
\]

with $p(y_j|x_i)$ the conditional probability that $y_j$ occurs when $x_i$ already took place.

**Definition 12** (Equivocation). The equivocation $H(Y|X)$ is a measure for the uncertainty at the output when the input is known.

**Definition 13** (Ambiguity). The ambiguity $H(X|Y)$ is a measure for the uncertainty at the input when the output is known.

Intuitively, if entropy $H(X)$ is regarded as a measure of uncertainty about a random variable $X$, then $H(X|Y)$ is "the amount of uncertainty remaining about $X$ after $Y$ is known".

**Lemma 5.** A conditional entropy is non-negative

\[
H(Y|X) \geq 0, \\
H(X|Y) \geq 0.
\]

**Proof.** Consider the definition of the conditional entropy

\[
H(Y|X) = - \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log p(y_j|x_i)
\]

Both $p(x_i, y_j)$ and $p(y_j|x_i)$ represent probabilities that are bounded in the range $[0, 1]$. Hence, all the individual terms are non-negative. \qed
Lemma 6. Relation between the different entropy definitions:

\[
H(Y|X) = H(X,Y) - H(X), \\
H(X|Y) = H(X,Y) - H(Y).
\]

Proof. We can rewrite the above definition of the condition entropy as follows:

\[
H(Y|X) = - \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log p(y_j|x_i)
\]

\[
= - \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)}
\]

\[
= - \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log p(x_i) \log p(y_j|x_i) + \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log p(x_i)
\]

\[= H(X) + \sum_{i=1}^{n} p(x_i) \log p(x_i) = H(X,Y) - H(X)
\]

where (a) uses the marginal probability

\[
p(x_i) = \sum_{j=1}^{m} p(x_i, y_j).
\]

Similarly, it is possible to proof that \(H(X|Y) = H(X,Y) - H(Y)\). \(\square\)

Lemma 7. Conditional entropy for an ideal channel:

\[H(X|Y) = H(Y|X) = 0\]

if and only if the value of \(Y\) is completely determined by the value of \(X\).

Proof. An ideal channel implies that there is no uncertainty on \(x_i\) given \(y_j\) and vice versa. Hence, \(p(x_i|y_j) = p(y_j|x_i) = 1\) for all \(i\) and \(j\). Using the definition of the conditional entropy, it becomes clear that \(H(X|Y) = H(Y|X) = 0\) since \(\log p(x_i|y_j) = \log p(y_j|x_i) = 0\). \(\square\)

The entropy for an ideal channel can be interpreted as follows: Consider the interpretation that the entropy is a measure of the randomness of a process (= the amount of information). An ideal channel implies that \(X\) is completely determined once \(Y\) is known. Hence, if we condition \(X\) on \(Y\), then there is no randomness left, implying that the conditional entropy \(H(X|Y)\) equals zero.

Lemma 8. Conditional entropy for independent random variables:

\[H(Y|X) = H(Y)\] and \(H(X|Y) = H(X)\) if and only if \(Y\) and \(X\) are independent random variables.
Proof. If $X$ and $Y$ are independent then $p(x_i, y_j) = p(x_i)p(y_j)$ and hence

$$H(Y|X) = - \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)}$$

$$= - \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i)p(y_j) \log \frac{p(x_i)p(y_j)}{p(x_i)}$$

$$= - \sum_{i=1}^{n} p(x_i) \sum_{j=1}^{m} p(y_j) \log p(y_j)$$

$$= H(Y)$$

An alternative proof is using

$$H(Y|X) \overset{(a)}{=} H(X, Y) - H(X)$$

$$\overset{(b)}{=} H(X) + H(Y) - H(X)$$

$$= H(Y)$$

where (a) uses Lemma 6 and (b) Lemma 4 for the joint entropy of independent random variables. Similarly, it is possible to proof that $H(X|Y) = H(X)$. □

The entropy of independent random variables can be interpreted as follows: If $X$ is independent of $Y$, then conditioning $X$ on $Y$ will have no influence on the randomness / amount of information. Hence, the conditional entropy $H(X|Y)$ equals the entropy without the condition $H(X)$.

**Lemma 9.** The condition entropy $H(Y|X)$ is bounded by $H(Y)$. Likewise, the condition entropy $H(X|Y)$ is bounded by $H(X)$.

$$H(Y|X) \leq H(Y),$$

$$H(X|Y) \leq H(X).$$

**Proof.** (For your information) This can be proven by first rewriting the definition of $H(Y|X)$

$$H(Y|X) = - \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log p(y_j|x_i)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} p(y_j) p(x_i, y_j) \log \frac{p(x_i)}{p(x_i, y_j)}$$

$$= \sum_{j=1}^{m} p(y_j) \left[ \sum_{i=1}^{n} p(x_i, y_j) \log \frac{p(x_i)}{p(x_i, y_j)} \right]$$
An upper bound can be found using Jensen’s inequality: for any concave function $f(x)$, values $t_1, \ldots, t_k$, and $\lambda_1, \ldots, \lambda_k \in [0, 1]$ with $\sum_{i=1}^{k} \lambda_i = 1$, it holds that

$$
\sum_{i=1}^{k} \lambda_i f(t_i) \leq f \left( \sum_{i=1}^{k} \lambda_i t_i \right)
$$

A function $f(x)$ on an interval is said to be concave if, for any $x$ and $y$ in that interval and for any $\alpha \in [0, 1]$ the function satisfies

$$
f((1 - \alpha)x + \alpha y) \geq (1 - \alpha)f(x) + \alpha f(y)
$$

The logarithm $f(x) = \log(x)$ is concave on its domain $(0, \infty)$. Using Jensen’s inequality, it can therefore be shown that

$$
H(Y|X) = \sum_{j=1}^{m} p(y_j) \left[ \sum_{i=1}^{n} \frac{p(x_i, y_j)}{p(y_j)} \log \frac{p(x_i)}{p(x_i, y_j)} \right]
$$

\[ \leq \sum_{j=1}^{m} p(y_j) \log \left[ \sum_{i=1}^{n} \frac{p(x_i, y_j)}{p(y_j)} \frac{p(x_i)}{p(x_i, y_j)} \right] \]

\[ \leq \sum_{j=1}^{m} p(y_j) \log \left[ \frac{1}{p(y_j)} \sum_{i=1}^{n} p(x_i) \right] \]

\[ \leq \sum_{j=1}^{m} p(y_j) \log \left[ \frac{1}{p(y_j)} \right] \]

\[ \leq H(Y) \]

where (a) uses Jensen’s inequality.

**Lemma 10.** $H(X, Y)$ is bounded

$$
0 \leq \max \{H(X), H(Y)\} \leq H(X, Y) \leq H(X) + H(Y)
$$

The lower bound of the joint entropy is the maximum of the individual entropies. The upper bound of the joint entropy equals to sum of the two individual entropies.

**Proof.** The upper bound is proven using

$$
H(X, Y) \overset{(a)}{=} H(X|Y) + H(Y) \overset{(b)}{\leq} H(X) + H(Y)
$$

where (a) uses Lemma 6, and (b) uses Lemma 9 where the equality holds when $X$ and $Y$ are independent.

The lower bound is proven by

$$
H(X, Y) = H(X|Y) + H(Y) \geq H(Y)
$$
and

\[ H(X,Y) = H(Y|X) + H(X) \geq H(X). \]

The equality holds for an ideal channel implying that \( H(X|Y) = H(Y|X) = 0 \) by Lemma 7.

### 2.7.3 Discrete mutual information (=transinformation)

The mutual information quantifies the amount of information transmitted over a channel. It therefore starts with the amount of transmitted information \( H(X) \) and takes into account the amount of information which is missing in the received signal \( Y \). Since entropy as a measure of uncertainty, it is reasonable to use the conditional entropy of the message, knowing the received signal, as a measure of this missing information: \( H(X|Y) \) basically describes what the uncertainty is on \( X \) when \( Y \) was received, hence it is a measure of the amount of information that got lost.

**Definition 14 (Mutual information).** The mutual information is defined as

\[ I(X;Y) = H(X) - H(X|Y) \]

For two discrete random variables \( X \) and \( Y \), the mutual information equals

\[ I(X;Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log \left( \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \right) \]

where \( p(x, y) \) is the joint probability function of \( X \) and \( Y \), and \( p(x) \) and \( p(y) \) are the marginal probability distribution functions of \( X \) and \( Y \) respectively. This can be shown using

\[
I(X;Y) \\
= H(X) - H(X|Y) \\
= -\sum_{i=1}^{n} p(x_i) \log p(x_i) + \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log p(x_i|y_j) \\
= -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log p(x_i) + \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log \left( \frac{p(x_i, y_j)}{p(y_j)} \right) \\
= \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log \left( \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \right). \]

**Lemma 11.** Mutual information for an ideal channel:

\( I(X;Y) = H(X) = H(Y) \) if the value of \( Y \) is completely determined by the value of \( X \).
CHAPTER 2. DISCRETE COMMUNICATION SYSTEMS

Proof. An ideal channel implies that there is no uncertainty on \( y_j \) given \( x_i \) and vice versa. Hence, \( p(x_i|y_j) = 1 \) and \( p(y_j|x_i) = 1 \). It can therefore be seen that

\[
I(X;Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log \left( \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \right)
\]

\[
= \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log \left( \frac{p(x_i,y_j)}{p(x_i)} \right)
\]

\[
= - \sum_{i=1}^{n} p(x_i) \log p(x_i) = H(X)
\]

and

\[
I(X;Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log \left( \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \right)
\]

\[
= \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log \left( \frac{p(y_j|x_i)}{p(y_j)} \right)
\]

\[
= - \sum_{j=1}^{m} p(y_j) \log p(y_j) = H(Y)
\]

An alternative proof is using the definition

\[
I(X;Y) = H(X) - H(X|Y)
\]

and Lemma 7 that the conditional entropy of an ideal channel \( H(X|Y) = 0 \). This immediately implies that \( I(X;Y) = H(X) \) for an ideal channel. \( \square \)

Lemma 12. Mutual information for independent random variables:

\( I(X;Y) = 0 \) if \( Y \) and \( X \) are independent random variables.

Proof. If \( X \) and \( Y \) are independent then \( p(x_i, y_j) = p(x_i)p(y_j) \) and hence

\[
I(X;Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log \left( \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \right)
\]

\[
= \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log \left( \frac{p(x_i)p(y_j)}{p(x_i)p(y_j)} \right)
\]

\[
= 0.
\]

An alternative proof uses the definition

\[
I(X;Y) = H(X) - H(X|Y)
\]

and Lemma 8 that for the conditional entropy of independent random variables \( H(X|Y) = H(X) \). This immediately implies that \( I(X;Y) = 0 \) for independent random variables. \( \square \)
Lemma 13. Mutual information written in various ways

\[ I(X; Y) = H(X) - H(X|Y) \]

\[ = H(X) + H(Y) - H(X, Y) \]

\[ = H(Y) - H(Y|X) \]

\[ = H(X, Y) - H(X|Y) - H(Y|X) \]

Proof. All these relations can be proven either from the definitions of the mutual information/various entropies

\[ I(X; Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log \left( \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \right) \]

and the fact that the logarithm of a product equals the sum of the individual logarithms.

Alternatively, it is possible to show

\[ I(X; Y) = H(X) - H(X|Y) \]

\[ = H(X) + H(Y) - H(X, Y) \]

\[ = H(Y) - H(Y|X) \]

\[ = H(X, Y) - H(X|Y) - H(Y|X) \]

from (a) the definition, (b) (c) and (d) using Lemma 6, respectively (b) \( H(X|Y) = H(X, Y) - H(Y) \), (c) \( H(X, Y) - H(X) = H(Y|X) \), and (d) \( H(Y) = H(X, Y) - H(X|Y) \).

Lemma 14. The mutual information is non-negative and bounded by the entropy of the source \( H(X) \)

\[ 0 \leq I(X; Y) \leq H(X) \]

Proof. Using the definition

\[ I(X; Y) = H(X) - H(X|Y) \]

and the fact that the conditional entropy \( H(X|Y) \) is upper bounded by \( H(X) \):

\[ H(X|Y) \leq H(X) \] (Lemma 9) implies that \( 0 \leq I(X; Y) \). This lower bound corresponds to a useless channel where \( X \) and \( Y \) are independent.

The upper bound is given by an ideal channel (Lemma 7) with \( H(X|Y) = 0 \) and hence \( I(X; Y) = H(X) \). The mutual information for an ideal channel therefore equals the entropy of the source.

Combining these two bounds implies that \( 0 \leq I(X; Y) \leq H(X) \).
2.7.4 Binary Symmetric Channel

**Definition 15 (Binary Symmetric Channel).** A Binary Symmetric Channel (BSC) with crossover probability $p$, denoted by $BSC_p$, is a channel with binary input $X$ and binary output $Y$. The probability that the output equals the complement of the input equals $p$ (Fig. 2.17).

![Diagram of Binary Symmetric Channel](#)

Figure 2.17: Representation of a Binary Symmetric Channel (BSC) with probability $p$ of having a transmission error.

Hence,

\[
\begin{align*}
p(y_1|x_1) &= 1 - p \\
p(y_1|x_2) &= p \\
p(y_2|x_1) &= p \\
p(y_2|x_2) &= 1 - p
\end{align*}
\]

Note that the processes $y_j$ conditioned on $x_1$ (i.e. $y_j|x_2$) and on $x_2$ (i.e. $y_j|x_1$) can both be seen a Bernoulli processes. An ideal channel corresponds with $p = 0$. It is furthermore assumed that $0 \leq p \leq 1/2$. If $p > 1/2$, then the receiver can swap the output symbols and obtain an equivalent channel with crossover probability of $1 - p$.

**Example 5.** Binary Amplitude Shift Keying (ASK) and Binary Phase Shift Keying (BPSK)

The demodulation of Binary ASK and BPSK start with a continuous source that is then quantized (converted into a binary output value) by comparing the input of the demodulator with a threshold put symmetrical between the two values of the noiseless symbol representations. The symmetry with respect to the threshold implies that a binary symmetric channel can be assumed.

The probability of having a bit error in the BSC channel, $p$, is also known as Bit-Error-Rate (BER), $P_e$, for the considered modulation. The probability of an error can be compute assuming an appropriate model for the channel (e.g. an Additive White Gaussian Noise (AWGN) channel). The BER for a Binary ASK where both symbols are a factor $A$ away from the threshold, and where the
symbols is perturbed by additive Gaussian noise with noise variance \( \sigma^2 \) equals

\[
P_e = Q\left(\frac{A}{\sigma}\right) = \frac{1}{2} \text{erfc}\left(\frac{A}{\sqrt{2} \sigma}\right)
\]

where the function \( Q(x) \) gives the probability that a single sample taken from a standard normal distribution (zero mean, unity variance) will be greater or equal to \( x \). \( Q(x) \) is related to the complementary error function

\[
\text{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt
\]

by

\[
Q(x) = \frac{1}{2} \text{erfc}\left(\frac{x}{\sqrt{2}}\right) \quad \text{for } x \geq 0.
\]

**Lemma 15.** The mutual information of a Binary Symmetric Channel (BSC\(_p\)):

\[
I_{BSC}(X; Y) = H(Y) - H_b(p)
\]

**Proof.** The mutual information for a BSC\(_p\) can be reformulated as

\[
I_{BSC}(X; Y) \stackrel{(a)}{=} H(Y) - H(Y|X) \\
\stackrel{(b)}{=} H(Y) - \sum_{i=1}^2 p(x_i) H(Y|X = x_i) \\
\stackrel{(c)}{=} H(Y) - \sum_{i=1}^2 p(x_i) H_b(p) \\
\stackrel{(d)}{=} H(Y) - H_b(p) \sum_{i=1}^2 \underbrace{p(x_i)}_{=1} \\
= H(Y) - H_b(p).
\]

The steps (a) and (b) follows from the definition of mutual information and conditional entropy respectively. Equality (c) expresses that the entropy at the output for a given and fixed input symbol \( H(Y|X = x_i) \). This expression is the further simplified (d) as the symmetry of the Binary Symmetric Channel implies that the individual entropies conditioned on the input \( x_i \) equal \( H(Y|X = x_i) = H_b(p) \) independently of \( x_i \).
CHAPTER 2. DISCRETE COMMUNICATION SYSTEMS

2.8 Discrete channel capacity

The capacity of the channel equals the maximum of the mutual information between the input and output of the channel, where the maximization is with respect to the input distribution.

Definition 16 (Channel capacity). The channel capacity is defined as

\[ C = \sup_X I(X;Y) \]

where the supremum is taken over all possible choices of the source \( X \).

Alternative definition in bits/sec can be found by multiplying with the average number of symbols / second.

A general theory can be found in [1], but this introduction will only consider two particular cases. First, the particular case of a Binary Symmetric Channel (BSC) is studied in order to illustrate the maximization with respect to the source \( X \). Next, a general result is given that can be applied in a lot of cases.

2.8.1 Channel capacity of a Binary Symmetric Channel

Theorem 2. The capacity of a BSC is

\[ C_{\text{BSC}}(p) = 1 - H_b(p) \]

which is attained for a uniform distribution at the input \( X \).

Proof. To obtain the channel capacity, it is necessary to take the supremum of the mutual information \( I(X;Y) \) over all possible inputs \( X \). Lemma 15 shows us that the mutual information of a BSC satisfies \( I(X;Y) = H(Y) - H_b(p) \). Hence,

\[ C_{\text{BSC}}(p) = \sup_X I(X,Y) = \sup_X (H(Y) - H_b(p)) \]

A first observation is that only the term \( H(Y) \) depends on the input distribution.

\[ C_{\text{BSC}}(p) = \sup_X (H(Y)) - H_b(p) \]

Additionally, it is shown in Section 2.1.5 that the entropy \( H(Y) \) of a binary variable \( Y \) is maximal and equal to one when the stochastic variable \( Y \) is uniformly distributed \( (p(y_1) = p(y_2) = 1/2) \). The case of an uniform distribution of \( Y \) can only be reached by a uniform distribution at the input \( X \) due to the symmetry of the channel. Hence, the capacity of a BSC is

\[ C_{\text{BSC}}(p) = 1 - H_b(p) \]

which is attained for a uniform distribution at the input \( X \). \( \square \)
Fig. 2.18 shows the channel capacity as function of the probability of having an error $p$. It is also important to note that the channel capacity of $C_{BSC} = 1 - H_b(p)$ is only attained when the transmitted symbols are equally probable. If not, the amount of mutual information over the channel will always be lower than the channel capacity shown in Fig. 2.18.

2.8.2 Generalization

A general solution for the channel capacity can be found in Section 16 of [1], showing how to compute both the channel capacity and the corresponding optimal source when the transition probabilities $p(y_j|x_i)$ of the channel are known. This leads to a set of equations that can be evaluate numerically. However, the equations are quite difficult to interpret in general. We will therefore focus on a special case which is encountered frequently in practice (or can be used as an approximation).

**Assumption 1.** Consider that each input symbol has the same set of probabilities on the lines emerging from it, and each output symbol has the same set of probabilities on the lines emerging to it. Examples of such channels are shown in Fig. 2.19.
Under these assumptions, the capacity can be easily calculated due to the symmetries.

Due to the assumed symmetry, the conditional entropy \( H(Y|X) \) becomes independent of the distribution of probabilities on the input symbols as \( p(y_j|x_i) \) only depends on \( y_j \) and does not depend on \( x_i \). Hence,

\[
H(Y|X) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log p(y_j|x_i) \\
= -\sum_{i=1}^{n} \sum_{j=1}^{m} p(y_j|x_i) p(x_i) \log p(y_j|x_i) \\
= -\sum_{i=1}^{n} p(x_i) \sum_{j=1}^{m} p(y_j|x_i) \log p(y_j|x_i) \\
= -\sum_{j=1}^{m} p(y_j|x_i) \log p(y_j|x_i)
\]

where \( p(y_j|x_i) \) are the transition probabilities from any input symbol \( x_i \). The channel capacity then equals

\[
C = \sup_X (H(Y) - H(Y|X)) \\
= \sup_X H(Y) + \sum_{j=1}^{m} p(y_j|x_i) \log p(y_j|x_i).
\]

The maximum of \( H(Y) \) equals \( \log m \) with \( m \) the number of output symbols, since it is possible to make all outputs \( Y \) equally probable by making the input symbols \( X \) equally probable. The channel capacity is therefore equal to

\[
C = \log m + \sum_{j=1}^{m} p(y_j|x_i) \log p(y_j|x_i)
\]
and this is attained for a uniform distribution for the input symbols $X$.

**Example 6.** Numerical examples for channels $(a, b, c)$ represented in Fig. 2.19

\[
C_a = \log 4 - \frac{2}{2} \log 2 = \log 2 \\
C_b = \log 4 - \frac{2}{3} \log 3 - \frac{2}{6} \log 6 \\
C_c = \log 3 - \frac{1}{2} \log 2 - \frac{1}{3} \log 3 - \frac{1}{6} \log 6
\]

**Example 7.** Fig. 2.20

Figure 2.20: Representation of a QPSK channel with a probability $p$ of having a transmission error with the neighboring symbol. In this case the symbols $x_i, y_j$ are presented as follows: 00 or 0 rad; 01 or $\pi/2$ rad; 11 or $\pi$ rad; and 10 or $3/2\pi$ rad.

represents the state transition diagram of a Quadrature Phase Shift Keying (QPSK) modulation when a probability $p$ is assumed of having a transmission error with the neighboring symbol. The channel capacity equals
\[ C_{QPSK} = \log 4 + (1 - p) \log(1 - p) + 2 \left(\frac{p}{2}\right) \log \left(\frac{p}{2}\right) \]
\[ = 2 + (1 - p) \log(1 - p) + p \log \left(\frac{p}{2}\right) \]

Example 8. Fig. 2.21

Figure 2.21: Representation of a bit-shift right operation as used in a 2-bit convolutional coder where the most significant bit is set to 1 (full arrow) or 0 (dashed arrow) with probability \( \frac{1}{2} \).

\[ \text{represents the state transition diagram for a 2-bit convolutional coder whose state is generated using a bit-shift left operation where the right-most bit is set to the data bit. The channel capacity is maximized for an input with a uniform distribution with } p = \frac{1}{2} \]

\[ C = \log 4 - \frac{2}{2} \log 2 = 1. \]

Hence, this convolutional coder can be used to transmit at most 1 bit at the time.
2.9 Discrete encoding and decoding in transceivers

This section introduces the mathematical operations performed by the transmitter and receiver in encoding and decoding the information. Either of these will be called a discrete transducer.

2.9.1 Transducers

The input to the transducer is a sequence of input symbols and outputs a sequence of output symbols. The transducer may have an internal memory so that its output depends not only on the present input symbol but also on the past history. We assume that the internal memory is finite, i.e. there exist a finite number \( n_a \) of possible states \( a \) of the transducer and that its output \( y(k) \) is a function \( f(\cdot) \) of the present state \( a(k) \) and the present input symbol \( x(k) \). The next state \( a(k+1) \) will be determined by a second function \( g(\cdot) \) of the same two quantities.

**Definition 17.** A transducer is described by two functions \( f \) and \( g \) with

\[
\begin{align*}
 y(k) &= f(x(k), a(k)) \\
 a(k+1) &= g(x(k), a(k))
\end{align*}
\]

where

- \( x(k) \) is the \( k^{th} \) input symbol,
- \( a(k) \) is the state of the transducer when the \( k^{th} \) input symbol is introduced,
- \( y(k) \) is the output symbol (or sequence of output symbols) produced when \( x(k) \) is introduced if the transducer’s state is \( a(k) \).

If the output symbols of one transducer can be identified with the input symbols of a second transducer, they can be connected in tandem and this cascade is also a transducer.

**Definition 18.** Non-singular and inverse transducers

If there exists a second transducer which operates on the output of the first and recovers the original input, then the first transducer will be called **non-singular** and the second transducer will be called the **inverse** of the first one.

2.9.2 Data processing inequality

**Theorem 3** (Data processing inequality). The output \( Y \) of a finite state transducer driven by a finite state statistical source (= a stationary and ergodic Markov process) \( X \), is a finite state statistical source with information rate (= entropy per unit time) \( R(Y) \) that less than or equal to that of the input entropy \( R(X) \)

\[
R(Y) \leq R(X).
\]

The equality holds if the transducer is non-singular.
The proof in [1] starts with the observation that both the source and the transducer are both represented by a state-space representation. Consider that the source is a Markov process which produces a sequence of symbols $x(k)$ starting from the state $b(k)$. The transition from $b(k)$ to $b(k+1)$ is defined by the Markov process and the number of states $n_b$ is finite. The transducer, on the other hand produces the blocks of output symbols $y(k) = f(x(k), a(k))$ using the input $x(k)$ and the internal states $a(k)$ of the transducer. The transition from $a(k)$ to $a(k+1)$ is defined by the transducer equation $a(k+1) = g(x(k), a(k))$. The number of states $n_a$ is also finite.

The combined system (transducer+source) can be represented by the “product state space” of the pairs $(a, b)$. This product state space contains a finite number of $n_a \times n_b$ states. One needs to take the product of the two spaces as all combinations of the states of spaces $a$ and $b$ must be considered. Each state in this product space will occur with a probability of $p_{a_i}f_{b_j}$ with $p_{a_i}$ the probability of states $a_i$ and $f_{b_j}$ the frequency of occurrence of the (Markov) state $b_j$.

The information rate of the output $Y$ can be calculated as the weighted sum over the states

$$R(Y) = \sum_{i=1}^{n_a} \sum_{j=1}^{n_b} p_{a_i}f_{b_j} H(Y|a_i, b_j).$$

The input symbol $x$ is a deterministic function of the states $b$ (the Markov process) and independent of $a$, hence

$$H(X|a_i, b_j) = H(X|b_j).$$

The output $y(k) = f(x(k), a(k))$ and the state update $a(k+1) = g(x(k), a(k))$ are produced in a deterministic way. This implies that the transducer can’t generate a higher entropy for the output $y(k)$ compared with its input $x(k)$

$$H(Y|a_i, b_j) \leq H(X|a_i, b_j).$$

This implies that

$$H(Y|a_i, b_j) \leq H(X|b_j).$$

The output rate of the transducer can therefore be bounded by

$$R(Y) \leq \sum_{i=1}^{n_a} \sum_{j=1}^{n_b} p_{a_i}f_{b_j} H(Y|a_i, b_j) \leq \sum_{i=1}^{n_a} \sum_{j=1}^{n_b} p_{a_i}f_{b_j} H(X|b_j) \leq \sum_{i=1}^{n_a} \left( \sum_{j=1}^{n_b} f_{b_j} \right) H(X|b_j) = R(X).$$
where the inequality in (a) holds if the transducer is non-singular.

To prove that \( R(X) = R(Y) \) for a non-singular transducer, consider that such non-singular transducer is connected with its output to its inverse transducer. If \( R(X), R(Y), \) and \( R(Z) \) are respectively the output information rate (=entropy per second) of the source, the first (=non-singular), and second (=the inverse) transducers, then \( R(X) \geq R(Y) \geq R(Z) = R(X) \). The output entropy of a non-singular transducer therefore equals the entropy of its input source \( R(X) = R(Y) \).

### 2.10 Shannon theorem

One of the major achievements of Shannon [1] is to set the conditions for communication without information loss.

**Theorem 4** (Shannon theorem). Let a discrete channel have the capacity \( C \) and a discrete source with information rate \( R \).

1. If \( R \leq C \) there exists a coding system such that the output of the source can be transmitted over the channel with an arbitrarily small frequency of errors (=arbitrarily small equivocation).
2. If \( R > C \) it is possible to encode the source such that the frequency of errors (=equivocation) is less than \( R - C + \varepsilon \) where \( \varepsilon \) is arbitrarily small.
3. There is no method of encoding which gives a frequency of errors (=equivocation) less than \( R - C \).

**Proof.** An extensive proof can be found in [1]. It uses the probabilistic interpretation of the entropy (Section 2.1.4) and the ergodicity of the source (=Markov process) to relate the information rate \( R \) with the probability of a transmission error for a channel with channel capacity \( C \). More details can be found in [1].

Shannon’s theorem is extremely important from both theoretical and practical point of view. It gives the answer to three important questions (given in the same order as in the theorem):

1. If the channel capacity (i.e. the maximum mutual information that can be transported without errors) is larger than the information rate \( R \) of the source, then it is possible to transmit the information of the source over the channel without any errors.

2. If the information rate \( R \) is larger than the channel capacity \( C \), then their exists a coding with a minimum number of errors / equivocation \( R - C + \varepsilon \) where \( \varepsilon \) tends to zeros when the length of the message tends to infinity. Hence, it is possible to determine the minimal information loss introduced by the channel.

3. If the information rate \( R \) is larger than the channel capacity \( C \), then it is just not possible to have an error rate less than \( R - C \).
Note that this theorem states that “there exists a coding system”, and “it is possible to encode”. How to engineer this optimal coding system is not elaborated in the theorem or its proof.
Chapter 3

Continuous communication systems

A discrete communication system assumes that the information is captured by a (discrete) sequence of discrete random variables $x(k)$. However, a large class of communication channels are continuous over time (instead of being a discrete sequence) and originate from a continuous random variable $x(t)$. Note that the discretization can take place in two dimensions: First, the signal can be discrete or continuous over time. Second, the random variable can either originate from a discrete or a continuous random distribution. These differences will be stressed in the text by talking about discrete-time / continuous-time, and discrete-variable / continuous-variable. The previous chapter on discrete communication systems was studying information theory for discrete-time discrete-variables.

This chapter will extend the fundamental concepts of information theory (entropy, mutual information, ...) for discrete random variables towards independent continuous random variables (Section 3.1). These concepts are then used to proof that the average information of a source is maximized by

- uniform distributions for bounded amplitude constraints (Section 3.2.2), and
- Gaussian distributions for fixed power constraints (Section 3.2.3).

Both are proven using a method from the calculus of variations that is introduced in Section 3.2.1.

The introduction of continuous variables makes is possible to define channels where the signals are perturbed using additive noise. The Additive White Gaussian Noise (AWGN) channels will be introduced and studied in Section 3.3.
3.1 Extending information concepts to continuous random variables

The entropy of a discrete set of probabilities $p(x_1), \ldots, p(x_n)$ has been defined as

$$H(X) = - \sum_{i=1}^{n} p(x_i) \log p(x_i).$$

In an analogous manner we define the entropy of a continuous distribution with the probability density function $f_X(x)$ by

$$H(X) = - \int_{-\infty}^{+\infty} f_X(x) \log f_X(x) dx.$$  

The subscript $X$ in $f_X(x)$ will be dropped for notational simplicity. Hence, the simplified notation is

$$H(X) = - \int_{-\infty}^{+\infty} f(x) \log f(x) dx.$$  

For an $n$-dimensional probability density function $f(x_1, \ldots, x_n)$ we have

$$H(X) = - \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(x_1, \ldots, x_n) \log f(x_1, \ldots, x_n) dx_1 \cdots dx_n.$$  

Similarly, the joint and conditional entropies can be defined when considering two continuous random variables $X$ and $Y$ (both may themselves be multidimensional). The joint entropy then equals

$$H(X,Y) = - \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y) \log f(x,y) dx dy$$

while the conditional entropies equal

$$H(Y|X) = - \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y) \log \left( \frac{f(x,y)}{f(x)} \right) dx dy$$

$$H(X|Y) = - \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y) \log \left( \frac{f(x,y)}{f(y)} \right) dx dy$$

where $f(x)$ and $f(y)$ represent the marginal distributions

$$f(x) = \int_{-\infty}^{+\infty} f(x,y) dy$$

$$f(y) = \int_{-\infty}^{+\infty} f(x,y) dx.$$
CHAPTER 3. CONTINUOUS COMMUNICATION SYSTEMS

3.1.1 Properties of the entropy for continues random variables

The entropies of continuous distributions have most (but not all) of the properties of the discrete case. Most of the properties will be restated below, and their proofs are similar to the discrete cases. An important property that differentiates a continues random variable, namely the effect of a change in coordinate system, will be treated in Section 3.1.2.

- $H(X, X) = H(X)$
- $H(X, Y) = H(Y, X)$
- $X$ and $Y$ are statistically independent if and only if $H(X, Y) = H(X) + H(Y)$
- A conditional entropy is non-negative $H(Y|X) \geq 0$ and $H(X|Y) \geq 0$
- $H(Y|X) = H(X, Y) - H(X)$ and $H(X|Y) = H(X, Y) - H(Y)$
- Conditional entropy for an ideal channel: $H(X|Y) = H(Y|X) = 0$ if and only if the value of $Y$ is completely determined by the value of $X$.
- Conditional entropy for independent random variables: $H(Y|X) = H(Y)$ and $H(X|Y) = H(X)$ if and only if $Y$ and $X$ are independent random variables.
- The condition entropy $H(Y|X)$ is bounded by $H(Y)$: $H(Y|X) \leq H(Y)$, likewise the condition entropy $H(X|Y)$ is bounded by $H(X)$: $H(X|Y) \leq H(X)$
- $H(X, Y)$ is bounded $0 \leq \max [H(X), H(Y)] \leq H(X, Y) \leq H(X) + H(Y)$
- Mutual information

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

- The mutual information is bounded $0 \leq I(X; Y) \leq H(X)$ with the lower bound for the useless channel and the upper bound for the ideal channel.

3.1.2 Changing coordinate system

There is one important difference between the continuous and discrete entropies. In the discrete case the entropy measures in an absolute way the randomness of the chance variable. In the continuous case the measurement is relative to the coordinate system. A change in coordinates will in general change the entropy.
This will be illustrated using a bijective linear coordinate transformation from \( x_1, \ldots, x_n \) to \( y_1, \ldots, y_n \)

\[
y_i = \sum_{j=1}^{n} a_{ij} x_j
\]

This can be represented in matrix form as

\[
y = Ax
\]

where \( A \) is a regular square matrix. Multivariate statistics shows that a change in coordinates implies that

\[
dy = |\det(A)| \, dx
\]

and

\[
f_Y(y) = \frac{1}{|\det(A)|} f_X(A^{-1}y)
\]

where \(|\det(A)|\) in the denominator is necessary to account for the conservation of the probability.

The entropy in the new coordinates therefore equals

\[
H(Y) = - \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_Y(y) \log(f_Y(y)) \, dy
\]

\[
= - \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \frac{1}{|\det(A)|} f_X(A^{-1}y) \log \left( \frac{1}{|\det(A)|} f_X(A^{-1}y) \right) \, dy
\]

\[
= - \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \frac{1}{|\det(A)|} f_X(x) \log \left( \frac{1}{|\det(A)|} f_X(x) \right) |\det(A)| \, dx
\]

\[
= - \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_X(x) \log \left( \frac{1}{|\det(A)|} f_X(x) \right) \, dx
\]

\[
= - \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_X(x) \log(f_X(x)) \, dx
\]

\[
+ \log(|\det(A)|) \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_X(x) \, dx
\]

\[
= H(X) + \log(|\det(A)|)
\]

In spite of this dependence on the coordinate system the entropy concept is important in the continuous case as derived concepts such as information rate and channel capacity depend on the difference of two entropies. Hence, they don't depend on the coordinate frame.

Note that the entropy of a continuous distribution can be negative! However, information rates and channel capacities will always be non-negative.
3.1.3 The entropy of a uniform distribution

Consider a uniform distribution in \([-a, a]\). Hence, the probability density function equals

\[
f(x) = \begin{cases} \frac{1}{2a} & \text{for } x \in [-a, a] \\ 0 & \text{elsewhere.} \end{cases}
\]

Using the definition of the entropy, it can easily be seen that

\[
H(X) = -\int_{-a}^{a} \frac{1}{2a} \log \left( \frac{1}{2a} \right) dx = \log(2a)
\]

Note the striking resemblance with the entropy of a discrete random variable with an alphabet length of \(n\) whose entropy equals \(\log n\).

3.1.4 The entropy of a Gaussian distribution

When substituting the Gaussian distribution

\[
f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}
\]

in the expression of the entropy

\[
H(X) = -\int f(x) \log f(x) dx
\]

one obtains

\[
H(X) = -\int_{-\infty}^{\infty} f(x) \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \right) dx
\]

\[
= -\int_{-\infty}^{\infty} f(x) \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right) dx + \int_{-\infty}^{\infty} f(x) \frac{x^2}{2\sigma^2} \log(e) dx
\]

\[
= \log \left( \sqrt{2\pi\sigma^2} \right) \int_{-\infty}^{\infty} f(x) dx + \frac{\log(e)}{2\sigma^2} \int_{-\infty}^{\infty} x^2 f(x) dx
\]

\[
= \frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2} \log(e)
\]

and hence

\[
H(X) = \frac{1}{2} \log(2\pi e \sigma^2).
\]

3.1.5 The entropy of a multi-dimensional Gaussian distribution case

Assume zero mean \(n\)-dimensional random variables \(x_1, \ldots, x_n\) that are stacked in a column vector \(x\). Let \(x^T\) represents the transpose of the vector \(x\). The
The $i,j$th element $C_{ij}$ of the covariance matrix $C = E[xx^T]$ is then given by

$$C_{ij} = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} x_i x_j f(x_1, \ldots, x_n) dx_1 \cdots dx_n$$

The probability density function of the $n$-dimensional zero-mean Gaussian distribution is given by

$$f(x_1, \ldots, x_n) = \frac{1}{\sqrt{(2\pi)^n \det C}} e^{-\frac{1}{2}x^TC^{-1}x}.$$ 

The entropy of this $n$-dimensional Gaussian distribution can be computed similarly to the one-dimensional case and equals

$$H(X) = \frac{1}{2} \log((2\pi)^n \det C) + \frac{1}{2} \log(e)$$

$$= \frac{1}{2} \log((2\pi)^n e \det C).$$

Note that $H(X)$ has two contributions:

1. $\frac{1}{2} \log((2\pi)^n \det C)$ is a contribution that depends on the number of random variable and their co-variance matrix, 
2. $\frac{1}{2} \log(e)$ is a constant term, independent of the stochastic properties of the process. 

The latter contribution can be seen as the constant term that comes with the coordinate system used (Section 3.1.2).

**Example 9.** Consider a set of $n$ independent, zero-mean Gaussian distributed random variables $x_1, \ldots, x_n$ with variances $\sigma^2_{x_1}, \ldots, \sigma^2_{x_n}$. The co-variance matrix then equals

$$C = \begin{bmatrix} \sigma^2_{x_1} & 0 & \cdots & 0 \\ 0 & \sigma^2_{x_2} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2_{x_n} \end{bmatrix}.$$ 

Using

$$\det(C) = \prod_{i=1}^{n} \sigma^2_{x_i},$$

the entropy $H(X)$ can be written as

$$H(X) = \frac{1}{2} \log((2\pi)^n \det C) + \frac{1}{2} \log(e)$$

$$= \sum_{i=1}^{n} \frac{1}{2} \log(2\pi \sigma^2_{x_i}) + \frac{1}{2} \log(e).$$

This equation clearly shows the contributions of the individual (independent) sources and the constant term generated by the coordinate system used.
3.2 Probability density functions $f(x)$ that maximize the entropy $H(X)$

This section determines the probability density functions that maximize the entropy under two constraints, namely a bounded amplitude and a fixed power constraint. We therefore need to determine a function that maximizes another function. This is different from classical maximization problems where a maximum value (a real number) is determined which maximizes the function.

In order to determine a function which maximizes another function, Section 3.2.1 introduces a method from the calculus of variations. Afterwards, this method will be applied to prove that the average information of a source is maximized by

- uniform distributions for bounded amplitude constraints (Section 3.2.2), and
- Gaussian distributions for fixed power constraints (Section 3.2.3).

3.2.1 Calculus of variations to determine $f(x)$ that maximizes $H(X)$

The calculus of variations is determining the extrema of so-called functionals, which can be described as "functions of functions". The main difference with classical extrema of regular functions is the following: Extrema for regular functions look for the numerical values or expressions for the extrema

$$\max_x g(x) \text{ and } \min_x g(x)$$

while extrema of functionals are looking for the extreme with respect to the elements $f(x)$ of a given function space defined over a given domain

$$\max_{f(x)} \Phi(f(x)) \text{ and } \min_{f(x)} \Phi(f(x)).$$

A simplified form of the one-dimensional Euler-Lagrange equation can be used to determine the probability density function $f(x)$ which maximizes the entropy. As the entropy is defined through the integral

$$-\int_{-\infty}^{\infty} f(x) \log f(x) \, dx$$

it is required to look for an extremum over a function of the form

$$\Phi(f(x)) = \int_a^b \phi(x, f(x)) \, dx.$$
Theorem 5 (Simplified Euler-Lagrange equation). The extrema of $\Phi(f(x))$ with respect to the elements $f(x)$

$$\max_{f(x)} \int_a^b \phi(x, f(x)) dx$$

(and equivalently for minima) is given by

$$\frac{\partial \phi(x, f(x))}{\partial f(x)} = 0$$

(3.1)

Proof. We wish to find a function $f(x)$ which satisfies the boundary conditions $f(a) = A$, $f(b) = B$, and which extremizes the functional $\Phi(f(x))$ under the assumption that $\Phi(f(x))$ is twice continuously differentiable.

Consider that $f(x)$ extremizes the functional $\Phi(f(x))$ and satisfies the boundary conditions. Any slight perturbation of $f(x)$ that preserves the boundary values must either increase $\Phi$ (if $f(x)$ is a minimizer) or decrease $\Phi$ (if $f(x)$ is a maximizer).

Let

$$g_\varepsilon(x) = f(x) + \varepsilon \eta(x)$$

(3.2)

be the result of such a perturbation $\varepsilon \eta(x)$ of $f(x)$, where $\varepsilon$ is small and $\eta(x)$ is a differentiable function satisfying $\eta(a) = \eta(b) = 0$.

Consider

$$\Phi_\varepsilon = \int_a^b \phi(x, g_\varepsilon(x)) dx = \int_a^b \phi_\varepsilon dx$$

where $\phi_\varepsilon$ depends on $x$ and $\varepsilon$. The total derivative of $\Phi_\varepsilon$ with respect to $\varepsilon$ equals

$$\frac{d\Phi_\varepsilon}{d\varepsilon} = \frac{d}{d\varepsilon} \int_a^b \phi_\varepsilon dx = \int_a^b \frac{d\phi_\varepsilon}{d\varepsilon} dx$$

The total derivative of the integrand $\phi_\varepsilon$ that can be computed using

$$\frac{d\phi_\varepsilon}{d\varepsilon} = \frac{\partial \phi_\varepsilon}{\partial x} \frac{dx}{d\varepsilon} + \frac{\partial \phi_\varepsilon}{\partial g_\varepsilon} \frac{dg_\varepsilon}{d\varepsilon}$$

(a)

$$= \frac{\partial \phi_\varepsilon}{\partial g_\varepsilon} \frac{dg_\varepsilon}{d\varepsilon}$$

(b)

$$\eta(x) \frac{\partial \phi_\varepsilon}{\partial g_\varepsilon}$$

where (a) uses the fact that $x$ does not depend on $\varepsilon$, and (b) uses (3.2). Hence,

$$\frac{d\Phi_\varepsilon}{d\varepsilon} = \int_a^b \eta(x) \frac{\partial \phi_\varepsilon}{\partial g_\varepsilon} dx$$

It was assumed that $f(x)$ was an extremum, implying that $\Phi_\varepsilon$ is extreme in $\varepsilon = 0$, implying that $g_\varepsilon(x) = f(x)$ and $\phi_\varepsilon(x, g_\varepsilon(x)) = \phi(x, f(x))$. Since $\Phi_\varepsilon$ has
an extremum in $\varepsilon = 0$, we have that

$$\left. \frac{d\Phi}{d\varepsilon} \right|_{\varepsilon=0} = \int_a^b \eta(x) \frac{\partial \phi}{\partial f} \, dx = 0$$

The fundamental lemma of calculus of variations state that if a continuous function $f(x)$ on an open interval $(a, b)$ satisfies the equality

$$\int_a^b f(x) h(x) \, dx = 0$$

for all $h(x)$ on $(a, b)$, then $f(x)$ must be identically zero. This finally leads to the (simplified) Euler-Lagrange equation that $f(x)$ needs to satisfy

$$\frac{\partial \phi(x, f(x))}{\partial f(x)} = 0.$$

### 3.2.2 Uniform distribution maximizes the entropy under bounded amplitude constraint

The aim is to determine the probability density function $f(x)$ that results in a maximum entropy assuming that $x$ is bounded between $-a$ and $a$. A symmetric range is assumed as an asymmetric range only implies an offset in the distribution, and such offset does not transport any information/entropy.

The bounded amplitude constraint implies that $f(x)$ is a non-zero probability density function in $[-a, a]$ and zero elsewhere. Hence, all integrals will have their integration ranges from $-a$ to $a$.

The problem at hand needs to maximize the entropy

$$- \int_{-a}^a f(x) \log f(x) \, dx$$

under the constraint that

$$\int_{-a}^a f(x) \, dx = 1$$

This maximization problem with equality constraints can be expressed using the Lagrange multiplier $\lambda$

$$\Phi(f(x), \lambda) = - \int_{-a}^a f(x) \log f(x) \, dx + \lambda \left[ \int_{-a}^a f(x) \, dx - 1 \right]$$

$$= \int_{-a}^a [-f(x) \log f(x) + \lambda f(x)] \, dx - \lambda$$

$$= \int_{-a}^a \phi(x, f(x), \lambda) \, dx - \lambda$$
CHAPTER 3. CONTINUOUS COMMUNICATION SYSTEMS

with

\[ \phi(x, f(x), \lambda) = -f(x) \log f(x) + \lambda f(x) \]
\[ = -f(x) \frac{\ln f(x)}{\ln 2} + \lambda f(x). \]

The problem at hand is not looking for the extreme of a function (e.g. a real value), but for the function \( f(x) \) that maximizes \( \Phi \). This requires the use of the calculus of variations where the Euler-Lagrange equation (Section (3.2.1)) states that if

\[ \Phi = \int \phi(x, f(x)) dx \]

that the function \( f(x) \) which maximizes \( \Phi \) must satisfy

\[ \frac{\partial \phi(x, f(x))}{\partial f(x)} = 0. \]

This ultimately leads to the following set of equations

\[ \frac{\partial \phi(x, f(x))}{\partial f(x)} = - \frac{1}{\ln 2} - \frac{\ln f(x)}{\ln 2} + \lambda = 0 \]
\[ \frac{\partial \Phi}{\partial \lambda} = \int f(x) dx - 1 = 0 \]

The first equation can be rewritten as \( \ln f(x) = (\lambda \ln 2 - 1) \), or equivalently

\[ f(x) = e^{(\lambda \ln 2 - 1)} \]

\( f(x) \) is therefore a constant over the range of \(-a\) to \(a\). This constant is fixed by the constraint that

\[ \int_{-a}^{a} f(x) dx = 1 \]

and hence

\[ f(x) = \begin{cases} \frac{1}{2a} & \text{for } x \in [-a, a] \\ 0 & \text{elsewhere.} \end{cases} \]

This proofs that the uniform distribution maximizes the entropy under a bounded amplitude constraint. The maximum entropy equals

\[ H(X) = - \int_{-a}^{a} \frac{1}{2a} \log \left( \frac{1}{2a} \right) dx \]
\[ = \log(2a). \]

Note the resemblance with the uniform discrete distribution: it also maximizes the entropy (but this time for a discrete random variable) and has an entropy of \( \log n \). Also note that the discrete variable does not depend on a coordinate system (always equal to \( \log n \)), while the \( H(X) = \log(2a) \) depends on the coordinate system \( (x) \) used.
3.2.3 Gaussian distribution maximizes the entropy under fixed power constraint

The aim is to determine the probability density function \( f(x) \) that results in a maximum entropy subject to the condition that the power \( \sigma^2 \) is fixed. This distribution does not depend on the mean value, as the mean value does not convey any information/entropy. We can therefore assume that \( f(x) \) is zero mean.

It is necessary to maximize the entropy
\[
- \int_{-\infty}^{\infty} f(x) \log f(x) dx
\]
under the constraints
\[
\int_{-\infty}^{\infty} f(x) dx = 1 \quad \text{and} \quad \int_{-\infty}^{\infty} x^2 f(x) dx = \sigma^2.
\]

This maximization problem with equality constraints can be expressed using Lagrange multipliers \( \lambda_1 \) en \( \lambda_2 \)

\[
\Phi(f(x), \lambda_1, \lambda_2) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx + \lambda_1 \left[ \int_{-\infty}^{\infty} f(x) dx - 1 \right] + \lambda_2 \left[ \int_{-\infty}^{\infty} x^2 f(x) dx - \sigma^2 \right]
\]

\[
= \int_{-\infty}^{\infty} \left[ -f(x) \log f(x) + \lambda_1 f(x) + \lambda_2 x^2 f(x) \right] dx - \lambda_1 - \lambda_2 \sigma^2
\]

with
\[
\phi(x, f(x), \lambda_1, \lambda_2) = -f(x) \log f(x) + \lambda_1 f(x) + \lambda_2 x^2 f(x)
\]
\[
= -f(x) \frac{\ln f(x)}{\ln 2} + \lambda_1 f(x) + \lambda_2 x^2 f(x).
\]

The problem at hand is not looking for the extreme of a function (e.g. a real value), but for the function \( f(x) \) that maximizes \( \Phi \). This requires the use of the calculus of variations where the Euler-Lagrange equation (Section (3.2.1)) states that if
\[
\Phi = \int_{-\infty}^{\infty} \phi(x, f(x)) dx
\]
that the function \( f(x) \) which maximizes \( \Phi \) must satisfy
\[
\frac{\partial \phi(x, f(x))}{\partial f(x)} = 0.
\]

This ultimately leads to the following set of equations.
\[
\frac{\partial \phi(x, f(x))}{\partial f(x)} = -\frac{1}{\ln 2} - \frac{\ln f(x)}{\ln 2} + \lambda_1 + \lambda_2 x^2 = 0
\]
\[
\frac{\partial \Phi}{\partial \lambda_1} = \int_{-\infty}^{\infty} f(x) dx - \sigma^2 = 0
\]
\[
\frac{\partial \Phi}{\partial \lambda_2} = \int_{-\infty}^{\infty} x^2 f(x) dx - 1 = 0.
\]

The first equation can be rewritten as
\[
\ln f(x) = (\lambda_1 \ln 2 - 1) + \lambda_2 x^2 \ln 2.
\]

where \(\lambda_1\) and \(\lambda_2\) are constants that are determined using the two other equations. Hence, it is possible to introduce two new constants \(\lambda'_1 = e^{(\lambda_1 \ln 2 - 1)}\) \(\lambda'_2 = \lambda_2 \ln 2\) such that
\[
f(x) = \lambda'_1 e^{\lambda'_2 x^2}.
\]

Using the constraints, it can easily be seen that the Gaussian distribution with
\[
f(x) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{x^2}{2\sigma^2}}
\]
satisfies both constraints. This proofs that the Gaussian distribution maximizes the entropy under a fixed power constraint.

The entropy of this one-dimensional distribution equals
\[
H(X) = \frac{1}{2} \log(2\pi e \sigma^2)
\]
as shown in Section 3.1.4.

Similarly, it can be proven for the \(n\)-dimensional case that the maximum entropy occurs when \(f(x_1, \ldots, x_n)\) is an \(n\)-dimensional zero-mean Gaussian distribution
\[
f(x_1, \ldots, x_n) = \frac{1}{\sqrt{(2\pi)^n \det C}} e^{-\frac{1}{2} x^T C^{-1} x}
\]
with \(C\) the covariance matrix with the \(i, j^{th}\) element
\[
C_{ij} = \int \cdots \int x_i x_j f(x_1, \ldots, x_n) dx_1 \cdots dx_n.
\]

The entropy of this distribution (Section 3.1.5) equals
\[
H(X) = \frac{1}{2} \log((2\pi)^n e \det C). \quad (3.3)
\]
3.3 Additive White Gaussian Noise (AWGN) channels

Additive White Gaussian Noise (AWGN) is a basic noise model used in information theory to mimic the effect of many random processes. It basically assumes the following properties:

**Additive** noise \( n \) is added to the original signal \( x \), producing \( y = x + n \).

**White** noise which has a uniform power across the frequency band of the system. It is an analogy to the color white which has uniform emissions at all frequencies in the visible spectrum.

**Gaussian** (=normal distribution) noise as wideband noise often comes from many various (independent) sources, implying that the summation of all these noise sources will tend to a Gaussian (=normal) distribution as stated by the central limit theorem in statistics.

Both single and multiple channels can be considered using this AWGN representation. A single channel will be represented by \( y = x + n \), where \( n \) is independent of \( x \) and zero-mean Gaussian distributed with

\[
f_N(n) = \frac{1}{\sqrt{2\pi \sigma_n^2}} e^{-\frac{n^2}{2\sigma_n^2}}
\]

where \( \sigma_n^2 \) represents the noise power.

Multiple channels will be represented using vectors which are represented in bold as \( y = x + n \), where \( n \) is a column vector, independent of \( x \), and zero-mean \( (E[n] = 0) \) normally distributed with

\[
f_N(n) = \frac{1}{\sqrt{(2\pi)^n \det C_n}} e^{-\frac{1}{2} x^T C_n^{-1} x}
\]

with covariance matrix \( C_n = E[nn^T] \).

3.4 Complex-valued continuous random variables

Complex variables are often used when representing telecom signals, e.g. the in-phase (I) and quadrature (Q) component of an IQ modulated signal represent the real and imaginary part of a complex valued signal.

There are two possible views on complex random variables. A first view is that the real \( re \) and imaginary \( im \) parts of the complex value \( z = re + j im \) are both considered as random variable. A single complex valued \( z \in \mathbb{C} \) can then be expressed using a two-dimensional random variable with \( x \in \mathbb{R}^2 \) with \( x_1 = \Re(z) \) and \( x_2 = \Im(z) \) with co-variance matrix \( C = E[xx^T] \) given by

\[
C = \begin{bmatrix}
\sigma_{x_1}^2 & \sigma_{x_1,x_2}^2 \\
\sigma_{x_1,x_2}^2 & \sigma_{x_2}^2
\end{bmatrix}
\]
Another point of view is to consider $z$ as a complex random variable where

$$E(z) = E(re) + j E(im)$$

and the variance equals

$$\sigma^2_z = E(zz^*) - E(z)E(z^*)$$

$$= \sigma^2_{re} + \sigma^2_{im}.$$  

Note the complex conjugation $\cdot^*$ of the second factors in the definitions.

This can be extended to the multivariate case by defining the co-variance matrix $C = E[zz^H] - E[z]E[z^H]$ with $z^H$ the Hermitian transpose of the vector $z$, i.e. $z^H = (z^*)^T$. Hence, the co-variance between two complex-valued random variables $z_1$ and $z_2$ is defined as

$$\sigma^2_{z_1z_2} = E(z_1z_2^*) - E(z_1)E(z_2^*).$$

Random variables $z_1$ and $z_2$ are said to be uncorrelated if $\sigma^2_{z_1z_2} = 0$.

Complex random variables also have the so-called relation matrix

$$E[zz^T] - E[z]E[z^T]$$

which is defined similar to the co-variance matrix, but now using transposes instead of Hermitian transposes.

**Definition 19** (Circularly-symmetric complex normal distribution). Circularly-symmetric complex normal distributed random variables $z = re + j im$ are complex-valued with

- $re$ and $im$ are jointly normal,
- $re$ and $im$ are zero mean $E(re) = E(im) = 0$,
- the relation matrix $E(zz^T)$ equals zero.

The latter implies that $E(re^2) - E(im^2) = 0$ and $E(re \cdot im) = 0$ and hence $\sigma^2_{re} = \sigma^2_{im}$ and $\sigma^2_{re,im} = 0$.

The complex variance $E(zz^H)$ equals

$$\sigma^2_z = E(re^2) + E(im^2)$$

$$= 2\sigma^2_{re}.$$  

Circular symmetric complex normal random distribution can therefore be seen as a bivariate normal distribution. This distribution is extensively used in signal processing.

**Lemma 16.** The entropy of (circularly-symmetric) complex normal noise process $Z$ equals

$$H(Z) = \log (\pi \sigma^2_z) + \frac{1}{2} \log(e).$$
Proof. Consider the complex-valued process as two real-valued ones (the $re$ and $im$ part). The $Z$ is circularly-symmetric complex normal distribution with covariance matrix

$$C = \begin{bmatrix} \sigma_{re}^2 & 0 \\ 0 & \sigma_{re}^2 \end{bmatrix}.$$

The entropy of the equivalent two-dimensional ($n = 2$) Gaussian distribution (3.3) equals

$$H(Z) = \frac{1}{2} \log((2\pi)^2 \det C) + \frac{1}{2} \log(e)$$

$$= \frac{1}{2} \log \left((2\pi \sigma_{re}^2)^2\right) + \frac{1}{2} \log(e)$$

$$= \log (2\pi \sigma_{re}^2) + \frac{1}{2} \log(e)$$

$$= \log (\pi \sigma_{z}^2) + \frac{1}{2} \log(e).$$

\qed
Chapter 4

Continuous time-signals

This chapter will start with the Nyquist-Shannon sampling theorem for band-limited systems as this theorem is used to transfer the knowledge of channel capacity for discrete-time systems towards continuous-time systems (Section 4.1). Section 4.2 introduces the band-limited Gaussian channel. Afterwards, the Nyquist sampling theorem is discussed and illustrated with the raised-cosine filter in Section 4.2.1. Last, the complex-valued representation of passband signals is discussed in Section 4.3.

4.1 Nyquist-Shannon sampling theorem for band-limited systems

Expanding the concepts such as channel capacity from the discrete-time, towards continuous-time signals demands a theorem that links both domains. This is done using the Nyquist-Shannon sampling theorem that specifies the conditions to perfectly reconstruct a continuous-time signal $x(t)$ with a given bandwidth $W$ using a discrete-time sequence $x(n)$ containing $2W$ samples.

**Theorem 6** (Nyquist-Shannon sampling theorem [1]). Assume that the continuous time signal $x(t)$ is band limited with $f_{\text{max}} \leq W$ and that $x(t)$ is equidistantly samples at the instances $t = kT_s$ for an integer $k$, with $T_s = 1/F_s$ the sampling period, and $F_s = 2W$ the sampling frequency. Then, $x(t)$ can be perfectly reconstructed by

$$x(t) = \sum_{k=-\infty}^{\infty} x(nT_s) \text{sinc} \left( \frac{t}{T_s} - k \right)$$

with

$$\text{sinc}(x) = \frac{\sin \pi x}{\pi x}.$$

$F_s$ is also called the Nyquist sampling rate and $T_s$ the Nyquist sampling period. This sinc function is visualized in Fig. 4.1.
CHAPTER 4. CONTINUOUS TIME-SIGNALS

The sinc function is given by

$$\text{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$$

for non-zero integer values of $x$.

The Nyquist-Shannon sampling theorem can be seen as a series expansion expressing the continuous time signal $x(t)$ as a sum of orthogonal sinc functions. The orthogonality comes from the fact that

$$\text{sinc}\left(\frac{t}{T_s} - n\right) = \begin{cases} 1 & t = kT_s \\ 0 & t \neq kT_s \end{cases}$$

for integer values $k$ and $n$. The coefficients $x(kT_s)$ can be considered as the coefficients of this series expansion.

Summarized: If a function of time $x(t)$ is limited to the band from 0 to $W$ cycles per second, then $x(t)$ is completely determined by giving its ordinates at a series of discrete points spaced with the Nyquist sampling period

$$T_s = \frac{1}{F_s} = \frac{1}{2W}$$

Hence, the continuous time signal $x(t)$ contains exactly the same amount of information that its sampled version $x(nT_s)$.

Figure 4.1: The sinc($x$) = sin($\pi x$)/($\pi x$) function which equals zero for non-zero integer values of $x$. 
4.2 Band-limited Gaussian channel

The Nyquist-Shannon sampling theorem demands that the signals are band limited. This is an issue when considering a white noise source. White noise sources have, by definition, an infinite bandwidth. Hence, in order to make using of the sampling theorem sampling theorem, we need to adapt our definition of Additive White Gaussian Noise (AWGN) channel towards a band-limited version.

**Definition 20** (Band-limited Gaussian channel). A band-limited Gaussian channel with a bandwidth $W$ consists of the addition of a band-limited input signal $x(t)$ with $f_{\text{max}} = W$, and a white Gaussian noise source $\zeta(t)$ that passes through an ideal low-pass filter $h_{\text{ideal}}(t)$ with a bandwidth $W$, resulting in the noise contribution $n(t) = h_{\text{ideal}}(t) * \zeta(t)$ where $*$ represents the convolution product. Hence,

$$y(t) = x(t) + n(t) = x(t) + h_{\text{ideal}}(t) * \zeta(t)$$

The band-limited noise source will still be called a white noise source as the power spectral density of $n(t)$ is constant in the frequency band up to $W$.

Consider the power spectrum of the white noise source

$$\Sigma_\zeta(f) = \frac{N_0}{2}, \quad \forall f$$

where $N_0$ equals the noise power in a 1 Hz bandwidth, measured in watts per hertz or joules. The ideal low-pass filter will filter out all contribution above $W$, hence

$$\Sigma_n(f) = \begin{cases} \frac{N_0}{2} & -W \leq f \leq W \\ 0 & \text{otherwise} \end{cases}$$

The factor $\frac{1}{2}$ originates from the fact that the total power is computed by integrating over both the negative and the positive frequency axes

$$P_n = \int_{-\infty}^{\infty} \Sigma_n(f) df = \int_{-W}^{W} \frac{N_0}{2} df = N_0 W$$

**Lemma 17.** The noise samples $n(kT_s)$ are independent, zero-mean Gaussian distributed with variance $N_0 W$.

**Proof.** It will be proven in Section 8.6.1 that an ideal low-pass filtered white Gaussian noise with bandwidth $W$ has a noise auto-correlation function which equals

$$E(n(t + \tau)n(t)) = \rho_n(\tau) = \frac{N_0}{2} 2W \text{sinc}(2W\tau).$$
The Nyquist-Shannon sampling theorem is used to get a discrete-time sequence of the noise signal by sampling at the Nyquist rate $F_s = 2W$. Hence, the samples need to be studied at the instances $\tau = kT_s = k/(2W)$. The auto-correlation sequence for the noise then becomes

$$E(n(t + kT_s)n(t)) = \rho_n(kT_s) = \begin{cases} N_0W, & k = 0 \\ 0, & \text{otherwise.} \end{cases}$$

This implies that there is no correlation between the different noise samples over time as the covariance

$$E(n(t + kT_s)n(t)) - E(n(t + kT_s))E(n(t))$$

equals zero for non-zero integers $k$. This proofs that the noise samples $n(kT_s)$ are uncorrelated, zero-mean normal distributed with variance $N_0W$.

Finally, it is known that uncorrelated Gaussian distributions implies that independence (see Theorem 37) which completes the proof. \qed

### 4.2.1 Nyquist Inter Symbol Interference (ISI) criterion

The Nyquist Inter Symbol Interference ISI gives the criterion in order to avoid inter symbol interference or ISI. It is a constructive method as it provides a method for constructing such band-limited filters.

The impulse response of the channel causes that transmitted consecutive symbols will be spread in time. This causes inter symbol interference as previously transmitted symbols affect the current received one.

The Nyquist theorem relates this time-domain condition (avoiding ISI) to an equivalent frequency-domain condition. It should be noted that this Nyquist ISI criterion is closely related to the Nyquist-Shannon sampling theorem (Theorem 6). The main difference is that the Nyquist-Shannon sampling gives the criterion to reconstruct the complete continuous-time signal using discrete-time values, while the Nyquist ISI criterion looks for frequency limited re-constructors to reconstruct the continuous-time signal at discrete time instance, namely at $nT_s$.

**Theorem 7** (Nyquist ISI criterion). Consider the channel impulse response $h(t)$. The condition for an Inter Symbol Interference (ISI)-free response requires that

$$h(nT_s) = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases}$$

for all integers $n$ and where $T_s$ represents the symbol period. The Nyquist ISI theorem says that this is equivalent to

$$\sum_{k=-\infty}^{\infty} H\left(f - \frac{k}{T_s}\right) = 1 \quad \forall f$$

where $H(f)$ is the Fourier transform of $h(t)$.
Note that this criterion intuitively requires that frequency-shifted replicas of $H(f)$ must add up to a constant value.

This criterion is practically used in baseband filtering by regarding the symbol sequence as weighted impulses (Dirac delta function). When the baseband filters in the communication system satisfy the Nyquist ISI criterion, symbols can be transmitted over a channel without ISI with flat response within a limited frequency band. Examples of such baseband filters are the sinc filter (see Theorem 6) and the raised-cosine filter described below.

4.2.2 Raised-cosine filter

The raised-cosine filter is a filter frequently used for pulse-shaping in digital modulation due to its ability to minimise inter symbol interference (ISI). Its name stems from the fact that the band-pass characteristic in its simplest form is a cosine function which is 'raised' up to form a constant plateau at low frequencies.

The transfer function of the raised-cosine filter is mathematically expressed as

$$H_{RC}(f, \beta) = \begin{cases} 1 & |f| \leq \frac{1-\beta}{2T_s} \\ \frac{1}{2} \left( 1 + \cos \left( \frac{\pi T_s}{\beta} \left[ |f| - \frac{1-\beta}{2T_s} \right] \right) \right) & \frac{1-\beta}{2T_s} < |f| \leq \frac{1+\beta}{2T_s} \\ 0 & \text{otherwise.} \end{cases} \quad (4.1)$$

where $0 \leq \beta \leq 1$ is the roll-off factor and $T_s$ equals the symbol period. Note that (4.1) can be interpreted as the smooth transition from a constant level 1 to 0 using half a cosine function.

The roll-off factor $\beta$ is a measure of the excess bandwidth of the filter, i.e. the bandwidth occupied beyond the Nyquist bandwidth of $1/(2T_s)$. If we denote the excess bandwidth as $\Delta f$, then $\beta = 2T_s \Delta f$ since

$$f_{max} = \frac{1}{2T_s} + \Delta f = \frac{1+\beta}{2T_s}.$$ 

The bandwidth of a raised-cosine filter is most commonly defined as the width of the non-zero portion of its spectrum, i.e. $W = (1 + \beta)/(2T_s)$.

Fig. 4.2 shows the transfer function of a raised-cosine filter for $\beta = 0.25$. It has a flat plateau 1 from zero to $0.75/(2T_s)$, a cosine transition centered around $1/(2T_s)$ from the plateau 1 to 0, and the flat plateau 0 above $1.25/(2T_s)$. The transfer functions for other values of $\beta$ are shown in Fig. 4.3. Note that $\beta = 0$ corresponds to the rectangular filter in the frequency domain.
The impulse response of the raised-cosine filter can be computed analytically.
and equals

$$h_{RC}(t) = \frac{\cos\left(\pi \beta \frac{t}{T_s}\right) \sin\left(\frac{\pi t}{T_s}\right)}{1 - (2\beta \frac{t}{T_s})^2 \pi \frac{t}{T_s}}$$

$$\cos\left(\pi \beta \frac{t}{T_s}\right) \frac{1}{1 - (2\beta \frac{t}{T_s})^2 \pi \frac{t}{T_s}} \sin\left(\frac{t}{T_s}\right)$$

It is important to notice that, due to the sinc term, the impulse response satisfies

$$h_{RC}(nT_s) = \begin{cases} 
1 & n = 0 \\
0 & n \neq 0
\end{cases}$$

for integer $n$. This is also demonstrated using Fig 4.4 which shows the impulse response of a raised-cosine filter for different values of $\beta$.

![Figure 4.4: Impulse response of a raised-cosine filter for various values of $\beta$. Note the zeros at $nT_s$ for non-zero integer $n$.](image)

Note that if $\beta = 0$, that the raised-cosine filter becomes a sinc filter with a rectangular filter characteristic in the frequency domain and hence has a sinc impulse response.

A practical issue when using this impulse response is that it has an infinite support in time. Practical systems will need to truncate the support over time, resulting in (small) non-zero side lobes in the frequency domain.
4.2.3 Root-Raised-Cosine filter

A root-raised-cosine filter (RRC), also known as square-root-raised-cosine filter (SRRC), is frequently used as the transmit and receive filter in a digital communication system to perform matched filtering (see Chapter 11). The combined response of two such RRC filters (one at the transmit and one at the receive side) results in an overall response of a raised-cosine filter. This raised-cosine filter is preferred to remove the inter symbol interference (ISI), as the raised-cosine filter satisfies the Nyquist ISI criterion. If both the transmitter and the receiver use an identical filter, the overall transfer function must satisfy $H(f)H(f) = H_{RC}(f)$ and hence

$$H_{RRC}(f) = \sqrt{H_{RC}(f)}.$$

Its name comes from the fact that its frequency response is the square root of the frequency response of the raised-cosine filter. Fig. 4.5 shows the frequency response of the root-raised-cosine and the raised-cosine filter.

![Figure 4.5: Transfer function of the raised-cosine and root-raised-cosine filter.](image)

Fig. 4.6 shows the impulse response of a root-raised-cosine filter. Unlike the raised-cosine filter, this impulse response is not zero at $nT_s$. 
4.3 Complex-valued representation of passband signals

Analog and digital modulation methods often use a small frequency band to transmit the data. Modulation schemes such as ASK, PSK, QAM, and FSK use these band-pass signals with constant or varying carrier frequency to transmit the data. As the carrier frequency is often much higher than the data-rate / bandwidth of the modulation, there is a need to represent this type of signals in a more efficient way: a complex-valued representation using the equivalent baseband of the signal (also known as equivalent lowpass).

**Definition 21** (Equivalent baseband/lowpass representation). Consider a signal $X_{BP}(t)$ which is a carrier that is modulated with a bandwidth which is small compared to the carrier frequency. This so-called band-pass signal

$$X_{BP}(t) = I(t) \cos(2\pi f_c t) - Q(t) \sin(2\pi f_c t)$$

can be represented efficiently using its equivalent baseband (=equivalent lowpass) representation where $I(t)$ represents the in-phase signal and $Q(t)$ the quadrature-phase signal. The equivalent baseband then considers the complex-valued signal

$$Z(t) = I(t) + j Q(t).$$
The link with the physical signal then corresponds to
\[ X_{BP}(t) = \Re \left( Z(t)e^{j2\pi ft} \right) \]

Instead of an in-phase / quadrature (IQ) representation, it is also possible to rewrite the modulation in an instantaneous amplitude (=envelope) / phase format
\[ X_{BP}(t) = A(t) \cos (2\pi f_c t + \phi(t)) \]
with the instantaneous amplitude
\[ A(t) = \sqrt{I^2(t) + Q^2(t)} \]
and phase
\[ \phi(t) = \arctan \left( \frac{Q(t)}{I(t)} \right). \]

Similarly, it is possible to write the modulation in an instantaneous amplitude / frequency format
\[ X_{BP} = A(t) \cos (2\pi(f_c + f_m(t))t). \]

The phase and frequency modulation are related by
\[ f_m(t) = \frac{1}{2\pi} \frac{d\phi(t)}{dt} \]
or equivalently
\[ \phi(t) = 2\pi \int_0^t f_m(t)dt + \phi_0. \]

These relationships enable the usage of IQ modulated signals to obtain amplitude, phase, and frequency modulation. How an analog or digital information source is converted in the in-phase and quadrature-phase signals is determined by the (wide-range) of possible analog and digital modulation schemes.

4.3.1 Analog modulation
Consider the analog modulation of a zero-mean signal \( X(t) \) with baseband bandwidth \( W \).

Amplitude modulation
\( \text{AM} \) Amplitude Modulation does not use the phase to convey information. Hence, the phase \( \phi(t) \) is constant value \( \phi_0 \)
\[ X_{BP}(t) = A(t) \cos (2\pi f_c t + \phi_0). \]

A classical amplitude modulation (analog AM) typically required that the amplitude remains positive, i.e.
\[ A(t) = A + K_A X(t) \geq 0 \]
with $A$ the carrier amplitude, and $K_A$ the amplitude modulation gain. The modulation index

$$m_{AM} = \frac{\max K_A |X(t)|}{A}$$

determines how close the instantaneous amplitude can come to zero by demanding that $|X(t)| \leq mA$. If $m = 0.5$, the carrier amplitude varies by 50% above (and below) its unmodulated level (see Fig. 4.7). For $m = 1$, it varies by 100% and the instantaneous amplitude sometimes reaches zero. This represents full modulation using standard AM and is often a target as it has the highest possible signal-to-noise ratio. However, $m$ must never exceed one in standard AM. Increasing the modulating signal beyond that point is known as over-modulation.

An AM results in a power spectrum comprising a carrier signal and two side-bands, one left and one right of the carrier (Fig. 4.8). The overall bandwidth equals $2W$, i.e. twice the bandwidth of the original signal.
Figure 4.7: AM signal for various modulation indexes (50% modulation, fully modulated and over-modulated).
DSB Double Side Band modulation is basically an AM modulation with over-modulation. Hence, the instantaneous amplitude $A(t)$ can become negative. This will reduce the carrier amplitude and hence improve the power efficiency. A DSB modulation results in a (reduced) carrier signal and two side-bands, one left and one right of the carrier, with a total bandwidth of $2W$.

DSB-SC DSB-Suppressed Carrier does not transmit the carrier signal at all. Hence, only the two sidebands with a total bandwidth of $2W$ remains. Although more power efficient, the DSC-SC has the disadvantage that recovering the original carrier signal is not straightforward.

SSB Single Side Band modulation uses the in-phase component to carry the signal, while the quadrature component is generated through a so-called Hilbert filter. This Hilbert filter is constructed such that one of the two sidebands is completely suppressed. Only one sideband is used to transport the information: the Upper SideBand (USB) upper or the Lower SideBand

Figure 4.8: Left: illustration of the spectrum of AM and SSB signals. The lower side band (LSB) spectrum is inverted compared to the baseband and the upper side band (USB). Right: illustration of the spectrum of VSB signals.
(LSB). The total bandwidth equals \( W \), the same bandwidth as the original signal. A small carrier can be present in the signal in order to ease the recovery of the carrier.

**SSB-SC** SSB-Suppressed Carrier does not transmit the carrier signal. Hence, only a single sidebands with a total bandwidth of \( W \) remains. Although this modulation is the most efficient in both power and bandwidth efficiency, the SSB-SC has the disadvantage that recovering the original carrier signal is not straightforward.

**VSB** Vestigial sideband suppresses one of the sidebands only partially and is used in e.g. analog video modulation schemes. This compromise between SSB and DSB as been used when the signal \( X(t) \) has significant low-frequency content. This makes the realization of good Hilbert filters extremely difficult for the low frequency range.

**Phase / frequency modulation**

**FM** Frequency Modulation codes the information through the deviation of the instantaneous frequency of the carrier with respect to the center frequency \( f_c \). It is assumed that the frequency modulated carrier has a constant amplitude \( A(t) = A \). The instantaneous frequency then equals

\[
f(t) = f_c + K_f X(t)
\]

with

\[
f_m = K_f X(t)
\]

and the instantaneous phase

\[
\phi(t) = 2\pi K_f \int_0^t X(t)dt + \phi_0.
\]

The device that generates a variable frequency which depends on a input signal \( X(t) \) is a so-called Voltage Controlled Oscillator (VCO).

The modulation index for FM modulation \( (m_{FM}) \) relates the variation in the carrier frequency \( K_f X(t) \) with the higher frequency component present in the modulating signal, \( W \):

\[
m_{FM} = \frac{\max K_f |X(t)|}{W}.
\]

For a sine wave modulation, the modulation index is seen to be the ratio of the peak frequency deviation of the carrier wave to the frequency of the modulating sine wave. If \( m_{FM} \ll 1 \), the modulation is called narrowband FM, and its bandwidth is approximately \( 2W \). If \( m_{FM} \gg 1 \), the modulation is called wideband FM and its bandwidth is approximately \( 2m_{FM}W \). A general expression for the bandwidth of an FM modulation is given by Carson’s rule [5], stating that nearly all \((\approx 98\%)\) of the power of a frequency-modulated signal lies within a bandwidth of \( 2W(1+m_{FM}) \).
4.3.2 Digital modulation

Figure 4.9: Relationship between the different digital modulation schemes.

**Definition 22** (Constellation diagram). Constellation diagram displays the signal as a two-dimensional complex-valued IQ-plane at symbol sampling instants.

The angle of a point represents the phase shift of the carrier wave from a reference phase. The distance of a point from the origin represents a measure of the amplitude or power of the signal.

**Definition 23** (Gray code). A Gray code (named after Frank Gray), is an ordering of the binary numeral system such that two successive values differ in only one bit (binary digit).
Table 4.1: Example of a 4-bit Binary and Gray code.

Examples of constellation diagrams and Gray codes can be found in Figures 4.11, 4.12, 4.13, and 4.14.

**Definition 24** (Eye diagram). An eye diagram is an oscilloscope display in which a digital-modulated signal is repetitively sampled and applied to the vertical input, while the data rate is used to trigger the horizontal sweep. It is so called because, for several types of digital modulation, the eye diagram looks like a series of eyes between a pair of rails.
AM Amplitude Modulation

PAM Pulse-amplitude modulation encodes the analog or digital information in the amplitude of a series of signal pulses. Demodulation is performed by detecting the amplitude level of the carrier at every single period. The received signal is only required at an integer number of sampling periods, and hence the channel must fulfill the Nyquist ISI criterion (Section 4.2.1). That is why raised-cosine characteristics are often used in practice (Section 4.2.2).

Digital PAM uses a fixed number of pulse amplitude $M$ (PAM-M), which is usually a power of two (Fig. 4.11). Examples are some Ethernet communication standards such as 100BASE-T4 (PAM-3), 1000BASE-T (PAM-5), and 10GBASE-T a version of PAM-16.
Figure 4.11: The constellation diagram for PAM-2, PAM-4, and PAM-8 modulation, together with the Gray-codes that are associated with each of the symbols. Note that neighboring symbols only differ by 1 bit.

**ASK** Amplitude-shift keying (ASK) is another name for PAM for digital communication.

**OOK** On-off keying is the simplest form of amplitude-shift keying (ASK) modulation that represents digital data at the presence or absence of a carrier wave.

**QAM** Quadrature Amplitude Modulation conveys the information through the amplitude modulation of two orthogonal carrier: the sine and cosine.

**Rectangular QAM** is frequently used and combines 2 orthogonal PAM-modulated channel using the sine and cosine. An example of such rectangular QAM-16 modulation and its (two) PAM-4 modulations are shown in Fig. 4.12. Fig. 4.13 shows constellation diagrams for different values of $M$. 
Figure 4.12: Example of how a rectangular QAM-16 can be seen as the transmission of two PAM-4 signals over two orthogonal channels (the sine and cosine carriers of the IQ modulation). The two PAM-4 constellations are shown left and above the QAM-16 constellation.
Figure 4.13: Example of QAM for various $M$: QAM-4 (=QPSK), QAM-16, QAM-64, and QAM-256. Note that a higher $M$ results in less space between the different states, and hence demanding higher signal-to-noise ratios.

**PM Phase Modulation**

**PSK** Phase-shift keying conveys data by modulating the phase of a constant frequency reference signal (the carrier wave). The modulation is accomplished by varying the in-phase and quadrature-phase component at a precise time. PSK uses a finite number of phases, each assigned a unique pattern of binary digits (Fig. 4.14). The received signal is only required at an integer number of sampling periods, and hence the channel must fulfill the Nyquist ISI criterion (Section 4.2.1). That is why raised-cosine characteristics are often used in practice (Section 4.2.2).
CHAPTER 4. CONTINUOUS TIME-SIGNALS

Figure 4.14: The constellation diagram for BPSK, QPSK, PSK-8, and PSK-16 modulation, together with the Gray-codes that are associated with each of the symbols. Note that neighboring symbols only differ by 1 bit. The full lines represent the boundaries used to decide on the value of a received symbol.

**BPSK** Binary phase-shift keying uses two phases which are separated by 180°. This modulation outperforms all other PSKs when considering its resilience to noise. However, it is only able to modulate at 1 bit/symbol.

**QPSK** Quadrature phase-shift keying (also known as PSK-4, or QAM-4) uses four points on the constellation diagram, equispaced on a circle. With four phases, QPSK can encode 2 bit/symbol shown in the diagram with Gray coding to minimize the bit error rate (BER). Note that a single QPSK channel can be seen as two BPSK channels which are transmitted using orthogonal carriers: using a sine and cosine at $f_c$. 
**OQPSK** Offset QPSK offsets the timing of the odd and even bits by half a symbol-period, implying that the in-phase and quadrature-phase components never change at the same time. This limits the phase-shift to no more than 90° at a time very $T_s/2$, instead of 180° every $T_s$. This yields lower amplitude fluctuations than non-offset QPSK.

**$\pi/4$-QPSK** uses two QPSK constellations which are rotated by 45° with respect to one another. Usually, either the even or odd symbols are used to select points from one of the constellations and the other symbols select points from the other constellation. This reduces the phase-shifts from a maximum of 180° for QPSK to a maximum of 135°. This results in amplitude fluctuations for $\pi/4$-QPSK between OQPSK and QPSK.

![Figure 4.15: The constellation diagram QPSK, OQPSK, and $\pi/4$-QPSK, together with the Gray-codes that are associated with each of the symbols. The full lines represent the transition that happen during the complete symbol interval: QPSK and $\pi/4$-QPSK move from symbol 00 to 11 directly (without going through an intermediate state). The even and odd states are rotated with $\pi/4$ for the $\pi/4$-QPSK (different states are indicated with × and ◦), implying that the maximum phase transition is smaller than for the QPSK. The OPSK uses a time offset between the transition of the I and the Q path, resulting in a first transition from '00' to '01' (dotted line), followed by a second transition from '01' to '11' (dashed line) half a sample period later. This also results in smoother (phase) transitions than for the QPSK.](image-url)

**MPSK** M-ary PSK can have any number of phases to construct the PSK constellation, although one usually uses a power of 2 to allow an integer number of bits per symbol. With more than 8 phases, the error-rate becomes too high and there are better modulations available such as quadrature amplitude modulation (QAM).

**DPSK** Differential PSK conveys data not by the actual phase, but the phase difference between the sampling instances. Non-differential PSKs have a phase ambiguity if the constellation is rotated when passing through the
communications channel. This problem can be solved by encoding the data into the phase difference between 2 symbols in time (and hence the name: differential). Analysis shows that differential encoding approximately doubles the error rate compared to ordinary: the probability of making an error is twice as high as the information is encoded in the difference between two signals.

**FM Frequency Modulation**

**FSK** Frequency-shift keying transmits the digital information through a discrete set of frequency changes in the carrier signal. Hence, each symbol $x_i$ corresponds with a particular frequency $f_{x_i}$.

**BFSK** Binary FSK uses two discrete frequencies to transmit binary ‘1’ and ‘0’. The ‘1’ is called the mark frequency while the ‘0’ is called the space frequency.

The FM modulation index for a BFSK is given by

$$m_{FM} = \frac{\Delta f}{W} = \frac{\Delta f}{T_s} = \Delta fT_s$$

with $\Delta f = |f_{x_1} - f_{x_2}|$, $T_s$ the symbol period, and $W = 1/T_s$ an approximation of the bandwidth of a binary modulated waveform.

In the case of digital modulation, the carrier frequency $f_c$ is never transmitted: one of two frequencies ($f_c - \Delta f/2$ and $f_c + \Delta f/2$) is transmitted depending on the binary state of the modulation signal.

**CPM Continuous Phase Modulation**

**CPFSK** Continuous-phase FSK is a FSK which guarantees that the phase is continuous. Indeed, one can implement a FSK by using independently running oscillators, and by switching between them at the beginning of each symbol period. Switching between the independent oscillators will causing sudden discontinuities in the transmitted signal.

If the FSK transmitters uses only one oscillator (a VCO), then switching to a different frequency will preserve the continuity of the phase. The elimination of discontinuities in the phase (and therefore elimination of sudden changes in amplitude) reduces sideband power and hence the interference with neighboring channels.

**GFSK** Gaussian FSK applies a Gaussian filter before applying the signal to the CPFSK modulator. This makes the transition between the different data symbols (and hence phase jumps) smoother, resulting in a reduced sideband power. Gaussian filtering (also known as pulse shaping) is a standard way for reducing the sideband power/spectral width and hence reducing the interference with neighboring channels. This Gaussian pulse shaping comes at the cost of an increased inter symbol interference (ISI)
as the Gaussian filter does not satisfy the Nyquist ISI criterion.

GFSK is used in various standards such as DECT, Bluetooth, Cypress WirelessUSB, ...

**MSK** Minimum-Shift Keying (also known as minimum frequency-shift keying) is a particular spectrally efficient form of a CPFSK by using a frequency difference between the binary symbols equals half the symbols rate \(\Delta f = 1/(2T_s)\). Hence, the waveforms that represent a '0' and a '1' bit differ by exactly half a carrier period. As a result, the FM-modulation index \(m_{FM} = 0.5\). This is the smallest FSK modulation index that can be chosen such that the waveforms for '0' and a '1' are orthogonal.

**GMSK** Gaussian minimum-shift keying is a continuous-phase frequency-shift keying modulation which combines the standard minimum-shift keying (MSK) and Gaussian FSK (GFSK): the frequency difference between the binary symbols equals half the bit rate \(\Delta f = 1/(2T_s)\) (MSK) and the digital data stream is first shaped with a Gaussian filter before being applied to a frequency modulator (GFSK). This results in much a narrower phase shift angles than most MSK modulation systems and hence reduces the sideband power.

GMSK is most notably used in the Global System for Mobile Communications (GSM).

### 4.3.3 Nyquist-Shannon sampling theorem for passband signals

Consider a real-valued lowpass signal with a bandwidth \(W\). The Nyquist-Shannon sampling theorem for baseband signals (Theorem 6) states that the symbol rate can not exceed \(2W\) symbol/s of real-valued symbols.

In the passband case, a signal with passband bandwidth \(W\) can be converted to an equivalent baseband signal (using undersampling or a superheterodyne receiver), with upper cut-off frequency \(W/2\). This equivalent baseband signal is a complex-valued signal, containing two real-valued and orthogonal signals since sine and cosine are orthogonal functions. This explains why only \(W\) complex-valued symbols can be considered per time unit, instead of \(2W\) real-valued ones. Hence, double-sideband modulation schemes such as QAM, ASK, PSK or OFDM results in a maximum symbol rate of \(W\) symbol/s of complex-valued symbols.
Chapter 5

Shannon-Hartley theorem on channel capacity

Shannon’s theorem provides an upper bound of the data rate $R$ which must be smaller or equal to the channel capacity $C$ to enable a communication without any errors. The question to be answered in this chapter: how can we compute the channel capacity for realistic situations and how can we interpret the results? This chapter will assume that the channel is perturbing the signals with uncorrelated additive noise.

After a general introduction in Section 5.1, the theory will be developed to compute the channel capacity for Average White Gaussian Noise (AWGN) channels for independent continuous random variables (Section 5.2). This leads to the Shannon-Hartley theorem that determines the maximum information rate that can be transmitted over such AWGN channel. After studying a single channel (Section 5.2.1), the theory is extended towards multiple channels (=multiple variables) (Section 5.2.2). The results provides insight on how to optimally distribute the signal power for the case of independent variables (Section 5.2.3) and a general theory (Section 5.2.4) which used in Multiple-Input Multiple-Output (MIMO) communication (Section 5.2.5).

After developing the theory for independent continuous random variables in Section 5.2, the theorem is extended towards continuous-time communications signals with a specified bandwidth $W$ (Section 5.4). This use the Nyquist-Shannon sampling theorem of Section 4.1 which enables the perfect reconstruct a continuous-time signal $x(t)$ with a given bandwidth $W$ using a discrete-time sequence $x(n) = x(nT_s)$. The Nyquist-Shannon sampling theorem together with the Shannon-Hartley theorem for independent continuous random variables then results in the Shannon-Hartley theorem for the channel capacity for band limited continuous systems (Section 5.3). This is an important results since it links the channel capacity to the Signal-to-Noise-Ratio ($SNR$) and the bandwidth ($W$) of the signal.

The physical interpretation of the Shannon-Hartley theorem is studied in
Section 5.4. By introducing the spectral efficiency of a signal, it becomes possible to study Shannon’s limit of the channel capacity for extreme cases (Section 5.4). These limits provide the theoretical upper bound of the communication rate. This makes it possible to a priori know the maximum achievable communication rate and hence the gain one can obtain using e.g. more complex modulation schemes or channel coding.

5.1 General introduction

Consider an additive noise channel $y = x + n$ where $x$ represents the vector of a continuous input variables, $y$ the output variables, and $n$ the vector of zero-mean distributed noise which is independent of $x$.

The channel capacity is defined as the supremum of the mutual information over all possible inputs

$$C = \sup_{X} I(X;Y)$$

The mutual information can be expressed as

$$I(X;Y) = H(Y) - H(Y|X)$$
$$= H(Y) - H(X + N|X)$$
$$\overset{(a)}{=} H(Y) - H(X|X) - H(N|X)$$
$$\overset{(b)}{=} H(Y) - H(N)$$

where (a) is using the property that $H(X + N) = H(X) + H(N)$ if $X$ and $N$ are independent; and (b) the property that $H(X|X) = 0$ and $H(N|X) = H(N)$ since $N$ is independent of $X$.

The channel capacity can therefore be reformulated as

$$C = \sup_{X} I(X;Y)$$
$$= \sup_{X} (H(Y) - H(N))$$
$$\overset{(a)}{=} \sup_{X} (H(Y)) - H(N)$$

where (a) is due to the independence of the noise $N$ and the input $X$.

For the last step, we have to look for input distribution of $X$ which maximize the entropy of $Y$. Section 3.2 showed two important cases, namely

1. that a uniform distribution maximizes the entropy under a bounded amplitude constraint (Section 3.2.2), and

2. that Gaussian distribution maximizes the entropy under a fixed power constraint (Section 3.2.3).
5.1.1 Bounded amplitude constraint

Section 3.2.2 showed that $H(Y)$ is maximized under a bounded amplitude constraint $A$ when the output $Y$ is uniformly distributed over the complete amplitude range $[-A/2, +A/2]$. Note that $Y$ needs to be uniformly distributed for both continuous and discrete distributions of $Y$.

Consider now that the noise $N$ originates from an equally spaced quantization of $X$, resulting in $Y$. Then, the noise $N$ can be approximated a uniform distributed for the noise with a range $[-q/2, +q/2]$. The variable $q$ therefore represents for the quantization noise and has a variance of a uniform distribution $\sigma_q^2 = \frac{q^2}{12}$ and an entropy $H(N) = \log(q)$.

Hence, the channel capacity equals

$$C = \log(A) - \log(q)$$

$$= \log \left(\frac{A}{q}\right)$$

as $Y$ is uniformly distributed in $[-A/2, +A/2]$.

**Example 10.** Consider a channel with a capacity of $C = n \text{ bit/symbol}$. In that case, the quantization must fulfill

$$q = A/2^n.$$  

If the quantization noise $N$ is small compared to $X$, then it is reasonable to say that $X$ needs to be uniformly distributed to obtain the (optimal) uniform distribution of $Y$ $[-A/2, +A/2]$.

**Example 11.** A companding (= compressing + expanding) algorithm mitigates the detrimental that come with the limited dynamic range of the channel. This is the case when sampling audio signals using analog-to-digital converters with equally spaced quantization levels. When considering that the original (audio) signal is Gaussian distributed, then a nonlinear transformation is required to transform this Gaussian distribution into the (optimal) uniform distribution prior to the uniform quantization. This can be done using the cumulative distribution function for normal distributions as this function transforms a standard normal distribution $\mathcal{N}(0, 1)$ into a uniform distribution in the range $[0, 1]$.

The two most popular compander functions used for telecommunications are the A-law and $\mu$-law functions.

5.1.2 Fixed power constraint

Section 3.2.3 showed that $H(Y)$ is maximized under a fixed power constraint when the output $Y$ is zero-mean normally distributed with covariance $C_y$. Since
CHAPTER 5. SHANNON-HARTLEY THEOREM ON CHANNEL CAPACITY

104

• \( y = x + n \),
• \( y \) and \( n \) are both independent, and
• both \( y \) and \( n \) are zero-mean normal distributed with covariance matrices \( C_y \) and \( C_n \),

it is proven in the courses of statistics that \( x = y - n \) is also zero-mean normal distributed. Hence, the input distribution of \( X \) which maximizes the information rate is a zero-mean normal distribution with covariance \( C_x \). Since \( y = x + n \) and since \( x \) and \( n \) are independent, it is known that \( C_y = C_x + C_n \).

The channels are called “Gaussian channels” as both the signal and the noise are Gaussian distributed.

5.2 Shannon-Hartley theorem for independent continuous variables

The Shannon-Hartley theorem determines the maximum information rate that can be transmitted over an Additive White Gaussian Noise channel. This section develops the theory for independent continuous random variables, assuming an additive Gaussian noise channel \( y = x + n \) where \( x \) represents the vector with continuous input variables, \( y \) the output variables, and \( n \) the vector of zero-mean normally distributed noise, independent of \( x \) and with covariance matrix \( C_n \).

After a general introduction, the cases of a single Gaussian channel (Section 5.2.1) and independent parallel Gaussian channels (Section 5.2.2) are studied. A practical issue when using parallel channels is how to distribute the available transmit power optimally. This problem is solved in Section 5.2.3, leading to the so-called water-filling algorithm. Afterwards, the general multi-variable case is discussed in Section 5.2.4, followed by the usage of the latter for Multiple-Input Multiple-Output (MIMO) communication systems in Section 5.2.5.

5.2.1 Single Gaussian channel

**Theorem 8.** The mutual information for a Gaussian channel is maximized for the input \( X \) which is normal distributed with variance \( \sigma_x^2 \). The channel capacity equals

\[
C = \frac{1}{2} \log \left( 1 + \frac{\sigma_x^2}{\sigma_n^2} \right)
\]

or equivalently

\[
C = \frac{1}{2} \log (1 + \text{SNR})
\]

where \( \text{SNR} \) represents the Signal-to-Noise Ratio (= the ratio between the signal power \( \sigma_x^2 \) and the noise power \( \sigma_n^2 \)).
Proof. Section 3.2.3 provided an expression for the entropy for normal distributions

\[ H(N) = \frac{1}{2} \log (2\pi e \sigma_n^2) \]
\[ \sup_X (H(Y)) = \frac{1}{2} \log (2\pi e \sigma_y^2) \]

\[ = \frac{1}{2} \log (2\pi e (\sigma_x^2 + \sigma_n^2)) \]

Hence, an explicit expression for the channel capacity can be found and equals

\[ C = \sup_X (H(Y)) - H(N) \]
\[ = \frac{1}{2} \log \left( 1 + \frac{\sigma_x^2}{\sigma_n^2} \right). \]

\[ \square \]

5.2.2 Independent parallel Gaussian channels

Several independent parallel Gaussian channels are often used in communication systems. This implies that both the input \( x \) and the noise \( n \) are independent over the different channels. Additionally, each channel \( i \) has its own signal power constraint \( \sigma_x^2 \) and noise power \( \sigma_n^2 \).

Theorem 9. The channel capacity for independent parallel channels equals

\[ C = \sum_{i=1}^{n} \frac{1}{2} \log \left( 1 + \frac{\sigma_x^2}{\sigma_n^2} \right) \] (5.1)

Proof. Since we consider \( n \) independent Gaussian channels, the mutual information between the transmitted vector \( x^T = [x_1, \ldots, x_n] \) and the received vector \( y^T = [y_1, \ldots, y_n] \) can be written as

\[ I(x; y) = I(x_1, \ldots, x_n; y_1, \ldots, y_n) \]

\[ \leq \sum_{i=1}^{n} I(x_i; y_i) \] (a)

\[ \leq \sum_{i=1}^{n} \frac{1}{2} \log \left( 1 + \frac{\sigma_x^2}{\sigma_n^2} \right) \] (b)

where we have equality in (a) if the variables \( x_i \) are independent and equality in (b) if \( x_i \) and \( n_i \) are Gaussian distributed. \[ \square \]
5.2.3 Power distribution for independent parallel Gaussian channels

An important question is: how to distribute the signal powers of all the channels $\sigma^2_{x_i}$ to maximize the channel capacity for a fixed total power

$$P = \sum_{i=1}^{n} \sigma^2_{x_i}. \quad (5.2)$$

**Theorem 10.** Given $n$ independent parallel Gaussian channels with noise variance $\sigma^2_{n_i}$ for $i = 1, \ldots, n$ with a total transmitted power $P = \sum_{i=1}^{n} \sigma^2_{x_i}$. The channel capacity is given by

$$C = \sum_{i=1}^{n} \frac{1}{2} \log \left( 1 + \frac{\sigma^2_{x_i}}{\sigma^2_{n_i}} \right)$$

where

$$\sigma^2_{x_i} = (B - \sigma^2_{n_i})^+$$

with operator

$$(x)^+ = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases}$$

and $B$ such that $P = \sum_{i=1}^{n} \sigma^2_{x_i}$. 

**Proof.** The maximization of the channel capacity (5.1) must now consider the additional constraint (5.2). This can be done using a Lagrange multiplier and the maximization of the function

$$J = \sum_{i=1}^{n} \frac{1}{2} \log \left( 1 + \frac{\sigma^2_{x_i}}{\sigma^2_{n_i}} \right) + \lambda \left( \sum_{i=1}^{n} \sigma^2_{x_i} - P \right)$$

with respect to the signal powers $\sigma^2_{x_i}$ and the Lagrange multiplier $\lambda$. Setting the derivative with respect to $\sigma^2_{x_i}$ equal to zero makes

$$\frac{\partial J}{\partial \sigma^2_{x_i}} = \frac{1}{2 \ln 2} \frac{1}{\sigma^2_{x_i}} + \frac{1}{\sigma^2_{n_i}} + \lambda = 0$$

and hence

$$\sigma^2_{x_i} + \sigma^2_{n_i} = -\frac{1}{2 \ln 2} = B$$

Here, the constant $B$ is independent of the sub-channel $i$. This leads to a set of equations with $n + 1$ variables ($B$ and $\sigma^2_{x_i}$ for $i = 1, \ldots, n$) and $n + 1$ equations

$$\begin{cases} \sigma^2_{x_i} = B - \sigma^2_{n_i} & \forall i = 1, \ldots, n \\ P = \sum_{i=1}^{n} \sigma^2_{x_i} \end{cases}$$
There are \( n \) additional inequality constraints that need to be fulfilled: the signal power of all channels must be non-negative \( \sigma^2_{x_i} \geq 0 \).

This requires the use of the so-called Kuhn-Tucker method can be seen as a generalization of the Lagrangian multiplier method, and which is often used in non-linear optimization.

The Kuhn-Tucker conditions rewrites the first set of equations

\[
\sigma^2_{x_i} = (B - \sigma^2_{n_i})^+
\]

with

\[
(x)^+ = \begin{cases} 
  x & x \geq 0 \\
  0 & x < 0 
\end{cases}
\]

while the solution remains optimal.

This implies that some channels may have so much noise

\[ B < \sigma^2_{n_i} \]

that they channels becomes useless

\[ \sigma^2_{x_i} = 0. \]

This method is often referred to as the water-filling algorithm.

**Example 12.** Consider \( n = 4 \) independent parallel channels with the noise variances specified in Table 5.1 and assume that the total power is restricted to \( P = 6 \).

<table>
<thead>
<tr>
<th>( i )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma^2_{n_i} )</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>( \sigma^2_{x_i} )</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 5.1: Noise variances \( \sigma^2_{n_i} \) and the optimal power distribution \( \sigma^2_{x_i} \) to demonstrate the water-filling algorithm.

Summing up all 4 channels implies that

\[
4B = \sum_{i=1}^{4} \sigma^2_{x_i} + \sum_{i=1}^{4} \sigma^2_{n_i} = P + 15 = 21.
\]

Hence, \( B = 21/4 \) and \( \sigma^2_{x_3} = B - \sigma^2_{n_3} = -3/4 \). This is violating the constraint that the signal powers must be non-negative. Hence, channel 3 is too noisy and can’t be used.
A second iteration to find the optimal distribution follows the same procedure but now switches off channel 3. Hence, we have that

\[
3B = \sum_{i=\{1,2,4\}} \sigma_{x_i}^2 + \sum_{i=\{1,2,4\}} \sigma_{n_i}^2
= P + 9 = 15
\]

and therefore \( B = 5 \). This leads to the power distribution as shown in Table 5.1 with a channel capacity of

\[
C = \frac{1}{2} \log \left( 1 + \frac{3}{2} \right) + \frac{1}{2} \log \left( 1 + \frac{1}{4} \right) + \frac{1}{2} \log \left( 1 + \frac{2}{3} \right)
= 0.661 + 0.161 + 0.368 = 1.19 \text{ bits}
\]

The origin of the terminology “water filling” can be shown by using Fig. 5.1 which provides a graphical interpretation of the power allocation scheme. The noise level for each sub-channel act as surface of a landscape. Then the power is poured in the landscape as water and the surface will stay at the level \( B \). The depth of the water for each sub-channel is the amount of power used. If, as in sub-channel 3, the water level does not reach above the noise level, this sub-channel should not be used.

![Figure 5.1: Graphical representation of the water-filling algorithm.](image)

### 5.2.4 General multi-variable case

An explicit expression for the channel capacity can be found when the channels are dependent. Consider the entropy of a multi-variable normal distribution as
derived in Section 3.2.3
\[
\sup_{X} (H(Y)) = \frac{1}{2} \log ((2\pi)^n e \det (C_x + C_n))
\]
\[
H(N) = \frac{1}{2} \log ((2\pi)^n e \det (C_n)).
\]

Hence, the channel capacity equals
\[
C = \sup_{X} (H(Y)) - H(N)
\]
\[
= \frac{1}{2} \log \left( \frac{\det (C_x + C_n)}{\det (C_n)} \right).
\]

This expression can be rewritten using the following properties of determinants.

- If matrices \( A \) and \( B \) are both square matrices then
  \[
  \det (AB) = \det A \det B.
  \]

- If the matrix \( A \) is square and regular then
  \[
  \det (A^{-1}) = \frac{1}{\det A}.
  \]

- If the matrix \( A \) is non-negative definite, then \( A \) can be rewritten as the power of two matrices \( A^{1/2} \)
  \[
  A = A^{1/2} A^{1/2}.
  \]

- Sylvester’s determinant identity: if \( A \) and \( B \) are matrices of size \( m \times n \) and \( n \times m \) respectively, then
  \[
  \det(I_m + AB) = \det(I_n + BA)
  \]
  where \( I_k \) is the identity matrix of order \( k \).

Since \( C_n \) is a covariance matrix, it implies that \( C_n \) is non-negative definite and can be rewritten as
\[
C_n = C_n^{1/2} C_n^{1/2}.
\]

Hence, the channel capacity can also be written as
\[
C = \frac{1}{2} \log \left( \det \left( I_n + C_x C_n^{-1} \right) \right)
\]
\[
= \frac{1}{2} \log \left( \det \left( I_n + C_n^{-1/2} C_x C_n^{-1/2} \right) \right).
\]
CHAPTER 5. SHANNON-HARTLEY THEOREM ON CHANNEL CAPACITY

For independent parallel channels, the covariance matrices $C_x$ and $C_n$ are diagonal. Therefore,

$$\det (I_n + C_x C_n^{-1}) = \prod_{i=1}^{n} \left( 1 + \frac{\sigma_x^2}{\sigma_n^2} \right).$$

The log operator transforms the product in a summation leading to the same result as in Section 5.2.2, namely

$$C = \sum_{i=1}^{n} \frac{1}{2} \log \left( 1 + \frac{\sigma_x^2}{\sigma_n^2} \right).$$

5.2.5 Shannon-Hartley theorem for Multiple-Input Multiple-Output (MIMO) systems

MIMO communication makes use of the different channels in order to increase the channel capacity [6]. In this case, the received signal $y$ and transmitted signal $x$ contain respectively $m$ and $n$ continuous random variables that are related to each other using a linear transformation matrix $H$ and an independent additive noise source $n$.

$$y = Hx + n.$$  

The question posed in MIMO communication is “how to optimally chose $H$ in order to maximize the channel capacity”? If the optimal $H$ is known, then it is possible to engineer pre-/post-processing of the transmitted/received signal such that the cascade of the pre-processing, the channel characteristic, and the post-processing equals the optimal $H$. How to engineer this pre- and post-processing is part of other courses on wireless MIMO technologies.

The channel capacity is determined similarly as in the previous section, but now we have

$$\sup_X (H(Y)) = \frac{1}{2} \log ((2\pi)^n e \det (C_y))$$

$$= \frac{1}{2} \log ((2\pi)^n e \det (H C_x H^T + C_n))$$

and

$$H(N) = \frac{1}{2} \log ((2\pi)^n e \det (C_n)).$$

The channel capacity can therefore be written as

$$C = \sup_X (H(Y)) - H(N)$$

$$= \frac{1}{2} \log \left( \frac{\det (H C_x H^T + C_n)}{\det (C_n)} \right)$$

$$= \frac{1}{2} \log \left( \det (I_m + H C_x H^T C_n^{-1}) \right)$$

$$= \frac{1}{2} \log \left( \det (I_m + C_n^{-1/2} H C_x H^T C_n^{-1/2}) \right) \quad (5.3)$$
CHAPTER 5. SHANNON-HARTLEY THEOREM ON CHANNEL CAPACITY

To interpret this equation, consider the following case

- the input signals $x$ are independent and has a total power $P$ which is distribute uniformly over the $n$ different inputs
  
  \[ C_x = \sigma_x^2 I_n \]

  with $\sigma_x^2 = P/n$.

- the noise $n$ on the $m$ channel outputs are uncorrelated
  
  \[ C_n = \sigma_n^2 I_m. \]

Then (5.3) becomes

\[ C = \frac{1}{2} \log \left( \det \left( I_m + \frac{\sigma_x^2}{\sigma_n^2} HH^T \right) \right). \]

The matrix $HH^T$ is a $m \times m$ squared (non-negative) matrix that can be de-
composed in its eigenvalues and eigenvectors. This eigendecomposition makes it possible to rewrite

\[ HH^T = U \Sigma U^T \]

where $U$ is a unitary matrix ($\det U = 1$) and $\Sigma$ is a diagonal matrix with non-negative eigenvalues $\lambda_i$ on the diagonal. Hence,

\[ C = \frac{1}{2} \log \left( \det \left( I_m + \frac{\sigma_x^2}{\sigma_n^2} U \Sigma U^T \right) \right). \]

This equation can be transformed using the Sylvester’s determinant identity

\[ \det(I_m + AB) = \det(I_n + BA) \]

implying that

\[ C = \frac{1}{2} \log \left( \det \left( I_m + \frac{\sigma_x^2}{\sigma_n^2} \Sigma U^T U = I_m \right) \right) \]

\[ \stackrel{(a)}{=} \sum_{i=1}^{m} \frac{1}{2} \log \left( 1 + \frac{\sigma_x^2}{\sigma_n^2} \lambda_i \right) \]

where (a) uses the fact that $\Sigma$ is a diagonal matrix with $\lambda_i$ on its diagonal.

If the number of transmitted signals $n$ is smaller than the number of received ones, $m$, then at most eigenvalues $\lambda_i$ will be non-zero. Hence,

\[ C = \sum_{i=1}^{\min(n,m)} \frac{1}{2} \log \left( 1 + \frac{\sigma_x^2}{\sigma_n^2} \lambda_i \right) \]

It can easily be seen that the total capacity of a MIMO channel is made up from a sum of AWGN parallel channels as studied in Section 5.2.2. The number of branches depends on the number of non-zero eigenvalues of $\Sigma$ and is at most the minimum of the number of transmitted and received signals ($\min(n,m)$). The capacity of each individual channel $i$ depends on the eigenvalues of that particular channel $\lambda_i$. 
5.3 Shannon-Hartley theorem

The previous sections studied the concepts of information theory for continuous random variables that don’t depend on time. This section will move from these time independent random variables to random signal that depend on time. This time dependency can be both discrete in time (discrete-time signals) and continuous in time (continuous-time signals). The time dependency of the signals will be characterized by a signal bandwidth $W$.

This section will use the Nyquist-Shannon sampling theorem for band-limited systems to transfer the knowledge of channel capacity for discrete-time systems towards continuous-time systems Section 4.1. The band-limited Gaussian channel (defined in Section 4.2) will then be used in Section 5.3.1 to determine the Shannon-Hartley theorem. An extension of this theorem to independent parallel channels is explained in Section 5.3.2, followed by the optimal power distribution to be used for independent parallel channels (Section 5.3.3).

5.3.1 Shannon-Hartley theorem

The Shannon-Hartley theorem determines the maximum information rate that can be transmitted over a band-limited Gaussian channel with a specified bandwidth $W$.

**Theorem 11** (Shannon-Hartley theorem). Consider a band-limited Gaussian channel (Definition 20) with an input signal $X(t)$ and additive white noise $N(t)$, both band-limited up to $f_{\text{max}} = W$. Consider furthermore that the total power of $X(t)$ equals $P$, while the power spectral density of $N(t)$ equals $N_0/2$. The band-limited Gaussian channel $Y(t) = X(t) + N(t)$ then has the capacity of

$$C = W \log \left(1 + \frac{P}{N_0 W}\right)$$

(in bit/s) which is attained when $X(t)$ is zero-mean Gaussian distributed.

*Proof.* The study of the continuous-time channel and their signals $(X(t), N(t), Y(t))$ is converted into the study of a discrete-time channel $(X(kT_s), N(kT_s), Y(kT_s))$ using the Nyquist-Shannon sampling theorem (Theorem 6). As both representations (continuous/discrete-time) are equivalent, their channel capacities will be equal to each other.

The discrete-time band-limited noise $N(kT_s)$ has the property that the individual samples $N(kT_s)$ are independent, zero-mean normal distributed with variance

$$\sigma_N^2 = \frac{N_0}{2}$$

as was proven in Lemma 17. Remember that the total noise power for the bandwidth $W$ was equal to $N_0 W$.

To determine the channel capacity, we need to optimize over the signals input $X(kT_s)$ that maximizes the mutual information. As seen previously, this
mutual information will be maximized if all the random variables \( X(kT_s) \) are independent with respect to \( k \).

The fact that both \( X(kT_s) \) and \( N(kT_s) \) are independent over \( k \) makes that the channel capacity equals the channel capacity of independent parallel channel (Section 5.2.2). For each transmitted sample, the channel capacity therefore equals (Theorem 5.2)

\[
C = \frac{1}{2} \log \left( 1 + \frac{\sigma_X^2}{\sigma_N^2} \right) \text{bit/sample}
\]

where this channel capacity is attained when \( X(kT_s) \) is a zero-mean Gaussian distribution with variance \( \sigma_X^2 \).

As \( X(kT_s) \) are independent Gaussian distributed variables, it is easy to see that the total power of the signal \( P \) equals

\[
P = 2W\sigma_X^2
\]

since \( 2W \) samples are considered per second. Hence, each transmitted sample \( X(kT_s) \) has the variance of

\[
\sigma_X^2 = \frac{P}{2W}
\]

which gives the capacity per sample

\[
C = \frac{1}{2} \log \left( 1 + \frac{P}{N_0W} \right) \text{bit/sample.}
\]

With \( F_S = 2W \) independent samples per second, the achievable bit rate becomes

\[
C = W \log \left( 1 + \frac{P}{N_0W} \right) \text{bit/s.}
\]

5.3.2 Shannon-Hartley theorem for independent parallel channels

The channel is typically not constant over the entire band, but there are variations both in noise level and signal attenuation. A popular modulation scheme that uses independent parallel channels is OFDM (Orthogonal Frequency Division Multiplexing) modulation which is used in WLAN (IEEE 802.11), xDSL, DVB-T (digital TV) and the down link of LTE (Long Term Evolution). OFDM modulation divides the full bandwidth \( W \) into several sub-bands with a bandwidth \( W_\Delta \) which are independent of each other (using the orthogonality of the OFDM modulation). The bandwidth of the sub-bands are chosen such that both the channel response \(|H_i|\) of the \( i \)th channel and its noise spectral density \( N_{0,i} \) can be considered to be constant within each sub-band.
As the transmitted power $P_i$ of the $i^{th}$ sub-channel is now attenuated with the channel response, the channel capacity of the $i^{th}$ sub-channel becomes equal to
\[
C_i = W_\Delta \log \left(1 + \frac{|H_i|^2 P_i}{N_{0,i} W_\Delta}\right).
\]

As the different channels are independent (due to the orthogonality), one can compute the total channel capacity using
\[
C = \sum_i W_\Delta \log \left(1 + \frac{|H_i|^2 P_i}{N_{0,i} W_\Delta}\right).
\] (5.5)

### 5.3.3 Power distribution for independent parallel channels

Considering $n$ sub-channels with a bandwidth $W_\Delta$ where each of the sub-channels (represented by the index $i$) has

- a transmitted power $P_i$,
- an constant attenuation of the channel, represented by $|H_i|$, and
- a constant noise power spectral density, represented by $N_{0,i}$.

The aim is to determine the optimal power distribution $P_i$ to maximize the channel capacity (5.5) under the constraint that
\[
P = \sum_{i=1}^{n} P_i.
\]

This optimization problem can be solved similarly as in Section 5.2.3 using the Lagrangian multiplier. Since negative powers are not allowed ($P_i \geq 0$), it is necessary to use the Kuhn-Tucker argument to achieve the water-filling algorithm. This results into
\[
P_i = \left(B - \frac{N_{0,i} W_\Delta}{|H_i|^2}\right)^+
\]
with
\[
(x)^+ = \begin{cases} 
  x & x \geq 0 \\
  0 & x < 0
\end{cases}
\]
and where $B$ is chosen such that $P = \sum_{i=1}^{n} P_i$. 

5.4 Shannon’s limit of the channel capacity

This section will discuss one of the most important results from information theory: the fundamental limit (=limit of Shannon Theorem 4) that the information rate $R$ is bounded by the channel capacity $C$ ($R \leq C$). This channel capacity sets the requirements on the signal-to-noise ($SNR$) for a reliable communication using the Shannon-Hartley theorem (Theorem (11)). Hence

$$R \leq C = W \log \left( 1 + \frac{P}{N_0 W} \right)$$

and the maximum information rate therefore depends on the bandwidth $W$, the transmitted signal power $P$, and the noise power $N_0 W$. Equivalently, it is possible to consider the signal to noise ratio

$$SNR = \frac{P}{N_0 W}$$

and hence

$$R \leq C = W \log (1 + SNR).$$

This makes it possible to study the limits of the Shannon-Hartley theorem by considering various points of view, namely what if

- an infinite bandwidth $W$ is available, but the total transmit power $P$ is finite?
- the energy per bit (represented by $E_b$) is considered instead of the overall power of the signal $P$?
- a fixed number of information bits is used per transmitted samples?
- channel coding is introduced where $n$ samples contain $k$ information bits (e.g. introduced redundancy using a transducer to be resilient to communication errors)?

Due to all the above reasons, there are various limits that can be considered depending on the constraint used [7]. After introducing the spectral efficiency in Section 5.4.1, the following two limit cases will be considered:

**Band-limited Shannon limit** is determined assuming that the bandwidth of the channel is limited (Section 5.4.3)

**Ultimate Shannon limit** (also known as fundamental limit) is the lowest possible $E_b/N_0$ to achieve a communication, assuming that the bandwidth of the channel tends to infinity (Section 5.4.4).

After studying these limit cases, the difference between power and bandwidth-limited regimes will be explained (Section 5.4.5), followed by introducing the impact of forward error correction to move closer to Shannon’s limit (Section 5.4.6).
5.4.1 Spectral efficiency

**Definition 25** (Spectral efficiency). Spectral efficiency, or more specifically link spectral efficiency, is the information rate $R$ that can be transmitted over a given bandwidth $W$

$$\eta = \frac{R}{W}$$

which is measured in $(\text{bit/s})/\text{Hz}$ or (somewhat more ambiguous) bit/s/Hz.

Spectral efficiency is used to analyze the efficiency of a digital modulation method or line code, possibly combined with a forward error correction (FEC) code and possibly other physical layer overhead. In the latter case, a "bit" refers to a user data bit, FEC overhead is always excluded.

The signal to noise ratio ($SNR$) can be related to the spectral efficiency and the energy per bit:

$$SNR = \frac{P}{N_0W}$$

$$(a) \quad \frac{E_bR}{N_0W} = \frac{E_b}{N_0}$$

where (a) uses the property that the signal power $P$ and the average energy per bit $E_b$ are then related to each other by

$$P = E_bR$$

i.e. the energy per bit times the number of bits per second (=information rate).

**Example 13.** A communication link using 1 kHz bandwidth to transmit 1000 bit/s has a spectral efficiency of $\eta = 1$ (bit/s)/Hz.

**Example 14.** A V.92 modem for the telephone network can transfer 56000 bit/s downstream over an analog telephone network. The frequency range of an analog telephone is limited to between 300 Hz and 3400 Hz, corresponding to a bandwidth 3100 Hz. The spectral efficiency therefore equals $\eta = 18.1$ (bit/s)/Hz.

5.4.2 Upper bound for attainable spectral efficiency

An upper bound for the attainable spectral efficiency is given by Hartley’s law: if each symbol of the signal is drawn from an alphabet with $M$ alternative symbols, then each symbol represents $N = \log M$ bit where $N$ is the spectral efficiency measured in bit/symbol.
CHAPTER 5. SHANNON-HARTLEY THEOREM ON CHANNEL CAPACITY

Baseband transmission

Consider the case of baseband transmission (line coding or pulse-amplitude modulation) with a baseband bandwidth $W$. The Nyquist-Shannon sampling theorem (Theorem 6) then states that the symbol rate cannot exceed $2W$ symbol/s. Hence, the spectral efficiency is limited by $\eta \leq 2N$ (bit/s)/Hz in the baseband transmission case.

Example 15. The downlink of a V.92 modem uses a pulse-amplitude modulation with 1024 signal levels, resulting in $N = 10$ bit/symbol. Since the transmitted signal before passband filtering can be considered as baseband transmission, the spectral efficiency cannot exceed $\eta \leq 20$ (bit/s)/Hz over the full baseband channel (0 to 4 kHz).

Passband transmission

In the passband transmission case, a signal with passband bandwidth $W$ can be converted to an equivalent baseband signal (using undersampling or a superheterodyne receiver), with upper cut-off frequency $W/2$. This equivalent baseband signal is a complex-valued signal, containing 2 real-valued and orthogonal signals (sin and cosine are orthogonal functions). This explains why only $W$ complex valued symbols can be considered per time unit, instead of $2W$ real-valued ones. Hence, if double-sideband modulation schemes such as QAM, ASK, PSK or OFDM are used, this results in a maximum symbol rate of $W$ symbol/s. The spectral efficiency is therefore limited by $\eta \leq N$ (bit/s)/Hz in the passband transmission case.

Example 16. A QAM-16 modem has an alphabet size of $M = 16$ alternative symbols, with $N = 4$ bit/symbol. As a QAM is a form of double-sideband passband transmission, the spectral efficiency cannot exceed $\eta \leq 4$ (bit/s)/Hz.

<table>
<thead>
<tr>
<th>Modulation</th>
<th>LP/BP</th>
<th>$N$ (bit/symbol)</th>
<th>$\eta$ (bit/s)/Hz</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAM-2</td>
<td>LP</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>PAM-4</td>
<td>LP</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>PAM-8</td>
<td>LP</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>QAM-4</td>
<td>BP</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>QAM-16</td>
<td>BP</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>QAM-64</td>
<td>BP</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 5.2: Spectral efficiency of various low-pass (LP) and band-pass (BP) digital modulation schemes.

Forward error correction

If a forward error correction code is used, the spectral efficiency is reduced compared to the uncoded spectral efficiency. The reduction in spectral efficiency is given by the code rate $r$ of the forward error correction (FEC) code.
Example 17. Consider a forward error correction (FEC) code with a code rate
\( r = \frac{1}{2} \), meaning that the encoder input bit rate is half the encoder output rate. This will reduce the spectral efficiency with a factor two. The FEC usually reduces the bit-error rate, and typically enables operation at a lower signal to noise ratios (SNR).

5.4.3 Band-limited Shannon limit

The usage of the spectral efficiency makes it possible to rewrite the Shannon theorem (Theorem 4) and the Shannon-Hartley theorem (Theorem 11).

**Theorem 12** (Shannon limits using spectral efficiency). *It is possible to write Shannon’s limit in various ways*

- **Shannon limit on the information rate**
  \[ R \leq W \log(1 + SNR) \text{ bit/s} \]

- **Shannon limit on spectral efficiency**
  \[ \eta \leq \log(1 + SNR) \text{ bit/Hz} \]

- **Shannon limit on SNR for a given spectral efficiency**\( \eta \)
  \[ SNR \geq 2^\eta - 1 \]

- **Shannon limit on** \( E_b/N_0 \) *for a given spectral efficiency**\( \eta \)
  \[ \frac{E_b}{N_0} \geq \frac{2^\eta - 1}{\eta} \]

*Note that the two latter limits are lower bounds rather than upper bounds.*

**Proof.** The first three inequalities follow directly out of the Shannon Theorem 4
\[ R \leq C, \]
the Shannon-Hartley Theorem 11
\[ C = W \log(1 + SNR), \]
and the definition of the spectral efficiency
\[ \eta = \frac{R}{W}. \]

The last inequality uses the fact that
\[ SNR = \frac{E_b}{N_0} \eta \]
as shown in (5.6).
### Chapter 5. Shannon-Hartley Theorem on Channel Capacity

<table>
<thead>
<tr>
<th>( \eta )</th>
<th>( \text{SNR (dB)} )</th>
<th>( \frac{E_b}{N_0} ) (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( -\infty )</td>
<td>-1.59</td>
</tr>
<tr>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>4.77</td>
<td>1.76</td>
</tr>
<tr>
<td>3</td>
<td>8.45</td>
<td>3.68</td>
</tr>
<tr>
<td>4</td>
<td>11.76</td>
<td>5.74</td>
</tr>
<tr>
<td>5</td>
<td>14.91</td>
<td>7.92</td>
</tr>
<tr>
<td>6</td>
<td>17.99</td>
<td>10.21</td>
</tr>
<tr>
<td>7</td>
<td>21.04</td>
<td>12.59</td>
</tr>
<tr>
<td>8</td>
<td>24.07</td>
<td>15.03</td>
</tr>
</tbody>
</table>

Table 5.3: The required SNR and \( \frac{E_b}{N_0} \) for various spectral efficiencies \( \eta \).

#### 5.4.4 Ultimate Shannon limit

The ultimate limit considers the most general case with only one assumption, namely that the power is limited to \( P \). No assumptions are made on the bandwidth. As the function \( \frac{2^\eta - 1}{\eta} \) decreases monotonically with \( \eta \), the ultimate Shannon limit is reached when \( \eta \to 0 \), or equivalently the bandwidth \( W \to \infty \).

**Theorem 13** (Ultimate Shannon limit). A reliable communication needs that

\[
\frac{E_b}{N_0} \geq \ln 2 = 0.69 = -1.59 \text{ dB}
\]

with \( E_b \) the energy per bit and \( N_0 \) the noise power in a 1 Hz bandwidth.

**Proof.** If there is no constraint on \( W \), then it is favorable to use an as large as possible bandwidth \( W \to \infty \), or equivalently \( \eta \to 0 \) as \( P \) is limited. Hence,

\[
\frac{E_b}{N_0} \geq \lim_{\eta \to 0} \frac{2^\eta - 1}{\eta} \quad (a)
\]

\[
\geq \lim_{\eta \to 0} \frac{2^\eta \ln 2}{1}
\]

where (a) use L'Hôpital’s rule

\[
\lim_{x \to p} \frac{f(x)}{g(x)} = \lim_{x \to p} \frac{f'(x)}{g'(x)}
\]

and

\[
\frac{d}{dx} (a^x) = a^x \ln a.
\]

This leads to the final results that

\[
\frac{E_b}{N_0} \geq \ln 2 = 0.69 = -1.59 \text{ dB}
\]

where the dB function uses \( 10 \log_{10} \) as \( \frac{E_b}{N_0} \) represents a power ratio. \( \square \)
The Shannon limit is a hard limit to achieve reliable communication: if $E_b/N_0$ is less than this limit, then it is impossible to reach a error probability that tends to zero, independent of the system or channel coding used.

5.4.5 Power and bandwidth-limited regime

If the desired spectral efficiency is less than $\eta = 1 \text{ (bit/s) / Hz}$ (the so-called power-limited regime), then it can be shown that binary codes can be used on the AWGN channel with a cost in Shannon limit on SNR of less than 0.2 dB [7].

On the other hand, since for a binary coding scheme the discrete-time code rate is bounded by $R/2W \leq 1 \text{ bit/symbol}$, the spectral efficiency of a binary coding scheme is limited to $\eta \leq 2 \text{ (bit/s) / Hz}$. Hence, multilevel coding schemes must be used if the desired spectral efficiency is greater than 2 (bit/s) / Hz (the so-called bandwidth-limited regime). In practice, FEC coding schemes for the power-limited and bandwidth-limited regimes differ considerably [7].

5.4.6 Application of the Shannon limit: modulation schemes and forward error correction

The understanding Shannon’s limit is important when determining the optimal balance between the modulation schemes used and the FEC coding used to protect the data against errors. The study of channel coding is outside the scope of this chapter (and part of another course). This section will therefore make abstraction from the channel coding used and will only consider a coding rate $r$ which is defined as the ratio between the encoder input bit rate $k$ and the encoder output rate $n$

$$r = \frac{k}{n}.$$ 

Note that $r < 1$ if redundant data is added to the data stream

If the symbol rate approaches the Nyquist limit of $2W \text{ symbol/s}$, then the transmitted data rate can approach

$$R = \frac{k}{n} 2W \text{ symbol/s}.$$ 

Consider now that each symbol represents $N = \log M$ bit where $N$ is the spectral efficiency measured in bit/symbol. Shannon’s limit learns us that, using the ideal coding scheme, it is possible to approach the spectral efficiency of

$$\eta \leq 2N \frac{k}{n}$$

for baseband transmission and

$$\eta \leq N \frac{k}{n}$$

for bandpass transmission.
Example 18. Consider binary modulation without coding \((r = k/n = 1)\) using a PAM-2 (baseband with \(M = 2\)) or a QAM-4 (bandpass with \(M = 4\)) modulation with a bandwidth \(W\). The maximal spectral efficiency equals \(\eta = 2\) (bit/s)/Hz in both cases. For a FEC with a coding rate of \(r = 1/2\), the spectral efficiency equals \(\eta = 1\) (bit/s)/Hz.

Using an optimum modulation and detection, it is possible to determine the baseline performance curve of \(P_b\) versus \(E_b/N_0\) for uncoded transmission as shown in Fig. 5.2. For e.g. PAM-2, it is possible to determine the theoretical expressions for the probability of an error per bit \(P_b\) (the BER) assuming that the signals have an average energy per bit of \(E_b\) and additive white Gaussian noise with variance \(\sigma^2 = N_0\)

\[
P_b = Q\left(\sqrt{\frac{2E_b}{N_0}}\right)
\]

where

\[
Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} dt.
\]

Theoretical curves for the theoretical BER \((P_b\) versus \(E_b/N_0\)) for other digital modulation schemes (PAM-M, QAM-M, PSK-M, ...) can be found in the literature [11] (see e.g. Fig. 5.2).

Fig. 5.2 shows that an \(E_b/N_0 \geq 9.6\) dB is required in order to achieve a bit error probability of \(10^{-5}\) for uncoded PAM-2 digital communication.

![Figure 5.2: \(P_b\) versus \(E_b/N_0\) for uncoded PAM / QAM, compared to Shannon limit for various \(\eta\).](image-url)
CHAPTER 5. SHANNON-HARTLEY THEOREM ON CHANNEL CAPACITY

Gap to capacity

Shannon’s limit on $E_b/N_0$ for $\eta = 2$ equals $E_b/N_0 = 1.76$ dB. Hence, the gap between the uncoded PAM-2 and its spectral efficiency of $\eta = 2$ equals 7.8 dB. This is called “the gap to capacity”. For a PAM-8 baseband modulation, the maximal spectral efficiency equals $\eta = 4$ (bit/s)/Hz. Hence, The gap to capacity equals 12.2 dB to the Shannon’s limit with $\eta = 4$ (bit/s)/Hz.

If a coding scheme with unlimited bandwidth expansion were allowed using PAM-2, i.e. $\eta \to 0$, then a further gain of 3.35 dB would be achievable to reach the ultimate Shannon limit of $E_b/N_0 = -1.59$ dB. Hence, the gap to capacity to the ultimate Shannon limit equals 11.2 dB.

An intermediate (more realistic case) would be the usage of a FEC encoder on the PAM-2 with a code rate of e.g. $r = 1/2$. To determine the performance of the FEC encoder in combination with the binary modulation, one must use the $\eta = 2r = 1$(bit/s)/Hz Shannon limit, or equivalently $E_b/N_0 = 0$ dB.

Fig. 5.3 shows the impact of the FEC coding on the required $E_b/N_0$ for a convolutional code with a code rate $r = 1/2$ and a constraint length of 7. This figure shows that it is possible to reduce the required $E_b/N_0$ of 9.6 dB for a PAM-2 towards 6.5 dB and 4.2 dB when using respectively a hard- or a soft decoding algorithm. Hence, the gap to capacity now equals 6.5 dB and 4.2 dB for respectively a hard- or a soft decoding algorithm.

![Figure 5.3: $P_b$ versus $E_b/N_0$ for uncoded and coded PAM-2, compared to Shannon limit for various $\eta$. The FEC coding scheme used is a convolutional code with a code rate $r = 1/2$ and a constraint length of 7. Both a hard and a soft decoding Viterbi decoding algorithm are shown.](image-url)
Chapter 6

Optimal binary encoding

The goal of data compression is to represent a source with the fewest bits to optimally recover the source from the compressed data. Data compression can be broadly classified into lossless and lossy compression. Lossless compression aims for the minimum number of bits that perfectly reconstructs the source from the compressed data, i.e. lossless / without information loss. Lossy data compression the data are, on the other hand, subjected to a maximum tolerable distortion.

This chapter deals with binary lossless encoding of a discrete source $X$ with an alphabet with length $n$ into an optimal sequence of $N(x_i)$ bits / binary symbols representing $Y$. The length of the bit sequences $N(x_i)$ can be either fixed for all symbols $x_i$ of $X$ (fixed-length or block coding), or a variable length depending on $x_i$ (variable-length coding).

Binary encoding is a special case of the transducers introduced in Section 2.9 as it is a stateless transducer (no states $a(k)$). Imposing that the encoder is lossless, and therefore can neither add nor destroy any information, enables the usage of the data processing inequality of Theorem 3. This theorem demonstrates that the transducer is lossless (=non-singular) if the information rate at input $R(X)$ equals the information rate at the output $R(Y)$ of the transducer ($R(Y) = R(X)$)

To determine the information rates $R(X)$ and $R(Y)$, consider the results on Section 2.2.2 on the information rate for binary coders (2.4). There, it was shown that the information rate of the binary encoded message equals

$$R(X) = \frac{1}{\langle N \rangle T_s} H(X)$$

for a fixed sampling period $T_s$ with $\langle N \rangle$ the average length of the output symbols

$$\langle N \rangle = \sum_{i=1}^{n} p(x_i) N(x_i).$$

As $Y$ is a binary source enables to substitute $H(Y)$ with the binary entropy
(Section 2.1.5) resulting in an information rate of $Y$ of

$$R(Y) = \frac{1}{T_s} H(Y) = \frac{1}{T_s} H_b(p)$$

with $p$ the probability of success $P[X = 1]$.

The data processing equality for non-singular transducers ($R(X) = R(Y)$ of Theorem 3) results in

$$R(X) = \frac{1}{\langle N \rangle T_s} H(X) = R(Y) = \frac{1}{T_s} H_b(p).$$

Hence,

$$H(X) = \langle N \rangle H_b(p).$$

It is known that $H_b(p) \leq 1$ bit and that $H_b(p)$ attains its maximum when both a 0 and a 1 are equally probable ($H_b(1/2) = 1$). This implies that the average length of the output symbol $\langle N \rangle$ must be larger or equal to the entropy of the source

$$H(X) \leq \langle N \rangle,$$

and that the equality is attained when both a ’0’ and a ’1’ are equally probable and when the bits are independent.

**Definition 26** (Efficiency of binary encoding). Efficiency of binary encoding is expressed as the ratio of the information rate over the maximum information rate

$$e_X = \frac{H(X)}{\langle N \rangle}.$$ 

This efficiency is bounded $0 \leq e_X \leq 1$ and attains its maximum $e_X = 1$ at the maximum information rate. Hence, an optimal binary encoding has $e_X = 1$.

Section 2.4.1 already provided two examples (one for fixed and one for variable length coding) and showed that the example on variable length coding was optimal. This chapter aims to determine the algorithms for (close to) optimal binary encoders.

The chapter is organized as follows. Section 6.1 will define the concept of uniquely decodable codes which is required to uniquely decode the message at the receiver side. Section 6.2 introduces the prefix codes, while Section 6.3 derives the Kraft’s inequality which gives the condition for uniquely decodable prefix codes. Section 6.4 describes three well known binary prefix codes, namely Shannon coding (Section 6.4.1), Shannon-Fano coding (Section 6.4.2), and Huffman coding (Section 6.4.3). Last but not least, the Shannon coding is used to prove Shannon’s fundamental source coding theorem which provides both an upper and a lower bound for the average code length.

### 6.1 Uniquely decodable codes

**Definition 27** (Uniquely decodable codes). A code is uniquely decodable if all possible output sequences of the source can be encoded in an unambiguous way.
In other words, a code is uniquely decodable if for every output sequence of the source with a finite alphabet length, the set of assembled codewords is different from the assembled codewords of every other output sequence that yields the same length. Two sets of codewords with the same length have to be different from one another if they were formed by different output sequences of the source.

Example 19. Consider the 5 codes defined in Table 6.1. The entropy $H(X)$ equals $1.75 \text{ bit/symbol}$ and is bounded between 0 and $\log 4 = 2$.

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$p(x_i)$</th>
<th>code 1</th>
<th>code 2</th>
<th>code 3</th>
<th>code 4</th>
<th>code 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$1/2$</td>
<td>00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>$1/4$</td>
<td>01</td>
<td>1</td>
<td>01</td>
<td>01</td>
<td>10</td>
</tr>
<tr>
<td>C</td>
<td>$1/8$</td>
<td>10</td>
<td>10</td>
<td>011</td>
<td>011</td>
<td>110</td>
</tr>
<tr>
<td>D</td>
<td>$1/8$</td>
<td>11</td>
<td>11</td>
<td>0111</td>
<td>111</td>
<td>111</td>
</tr>
<tr>
<td>$\langle N \rangle$</td>
<td></td>
<td>2.0</td>
<td>1.25</td>
<td>1.875</td>
<td>1.75</td>
<td>1.75</td>
</tr>
<tr>
<td>$\epsilon_X$</td>
<td></td>
<td>88%</td>
<td>140%</td>
<td>93%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>$K$</td>
<td></td>
<td>1.0</td>
<td>1.5</td>
<td>0.9375</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 6.1: Examples of fixed and variable-length codes for a source with $H(X) = 1.75$.

**Code 1** is has a fixed code length (also known as a block code) with $N(x_i) = 2$ for all $i$. The code is uniquely decodable as the length of each symbol is fixed. The code sequence "000110" can only be encoded as "ABC".

**Code 2** is a variable-length code which is **not uniquely decodable**. Code sequence "10011" could be interpreted as "BAABB" or "CABB" or "CAD",... Hence, this code destroys source information and is therefore unusable. This can also be seen by the fact the efficiency $\epsilon_X$ is larger than 100%.

**Code 3** is a so called comma code that uses the binary symbol "0" to indicate the beginning of each codeword (auto interpunction). This code is **uniquely decodable** as the start of a new code is always started with a "0". The code sequence "01110100101" can only be decoded as "DBABB". Its efficiency (of $\epsilon_X = 93\%$) is better than code 1 ($\epsilon_X = 88\%$).

**Code 4** is **uniquely decodable**, but the decoding can’t be done instantaneously as the ambiguity can only be resolved by considering future bits. Consider the coding of the code sequence "01110". Although the first symbol is an "A" (="0"), we first need to move up to the fourth bit before ensuring that the first two codes are "AD". Such a decoding delay is highly undesirable.

**Code 5** is quite similar to code 4, but this is an example of an instantaneous code as it can be decoded without considering future bits. It belongs to the class of prefix codes that will be studied in the next section. This code is uniquely decodable and is optimal since the coding efficiency equals 1. Note that the probability for a "0" and a "1" are equal in this code.

It is extremely important to be able to verify whether or not a code is uniquely decodable. This can be done using Kraft’s inequality. Although the
Kraft’s inequality can be proven in a general setting, this course will limit the proof to prefix codes.

6.2 Prefix codes

A prefix code is typically a variable-length code that uses the so-called "prefix condition" to distinguish the different code: there is no whole codeword that is a prefix of any other codeword. For example, a code with codewords \{01, 111\} has the prefix property; a code consisting of \{01, 011\} does not, because "01" is a prefix of "01" and also of "011".

**Definition 28 (Prefix condition).** The prefix condition requires that for a given codeword of length \(k\) having elements \((y_1, y_2, \ldots, y_k)\), there is no shorter codeword of length \(l < k\) with the elements \((y_1, y_2, \ldots, y_l)\).

In other words, there is no codeword of length \(l < k\) that is identical to the first \(l\) elements of another codeword of length \(k\). This property makes the codewords uniquely and instantaneously decodable.

**Lemma 18 (Prefix codes are uniquely decodable).** Given a sequence of prefix codes which satisfy the prefix condition (Definition (28)), a receiver can identify each word without requiring a special marker between the words.

There are uniquely decodable codes that are not prefix codes: the reverse of a prefix code is still uniquely decodable. An example of such suffix code is code 4 in Example 19.

Prefix codes can also be represented using tree graphs, which boil down to binary tree in the case of binary codes. The interpretation of these binary tree will allow us to proof Kraft’s inequality using a rather straightforward reasoning in Section 6.3.

Consider the equivalent binary tree shown in Fig. 6.2 for the binary prefix code \{00, 10, 110, 111\}. This binary tree is draw up the maximum length of all codes, which equals three. All valid prefix codes are shown with a black background. Once the codes are chosen (e.g. 00), then all codes that decedent from that code (i.e. 000 and 001) violate the prefix condition. Hence, they can’t belong to the set of valid prefix codes and are therefore displayed in gray. It can, however, be that some valid prefix codes are not considered as shown in Fig. 6.2. This is indicated with the squares to indicate where the code is incomplete.

**Example 20.** Consider the binary prefix code \{0, 10, 110, 111\} with \(N(x_i) = \{1, 2, 3, 3\}\) (code 5 in Example 19) which fulfills the prefix condition of Definition 28. The binary tree in Fig. 6.1 shows that there are no other valid prefix codes still available in this tree. Hence, it is an example of a **complete** binary prefix code.
Example 21. Consider the binary prefix code \{01, 10, 110, 111\} with $N(x_i) = \{2, 2, 3, 3\}$ that fulfills the prefix condition of Definition 28. Fig. 6.2 shows that prefix code code has some **redundancy** since the sub-tree that starts at 01 could be used to specify either one (\{01\}) or two (\{010, 011\}) additional prefix codes.

Example 22. Consider the binary prefix code \{00, 01, 10, 110, 111\} with $N(x_i) = \{2, 2, 3, 3\}$ that fulfills the prefix condition of Definition 28. This code is the result of a Huffman coding described in Section 6.4.3. Fig. 6.3 shows once more
that this is a complete binary prefix code.

Figure 6.3: Binary prefix encode for code \{00, 01, 10, 110, 111\}.

The next section will provide the condition, the Kraft’s inequality, to determine the possible sets of codeword lengths for uniquely decodable codes.

### 6.3 Kraft-McMillan inequality


Assume a discrete source \(X\) with an alphabet of \(n\) symbols and assume that each code \(x_i\) of \(X\) is encoded into \(y_i\). Furthermore, assume that each \(y_i\) consists out of a set of \(N_i\) symbols from an alphabet that contains \(m\) symbols. Such code is uniquely decodable if it satisfies Kraft’s inequality.

**Theorem 14** (Kraft’s inequality). Kraft’s inequality limits the lengths of codewords in a prefix code: the prefix code is a uniquely decodable over an alphabet of size \(m\) with codeword lengths \(N(x_i), \ldots, N(x_n)\) if

\[
K = \sum_{i=1}^{n} m^{-N(x_i)} \leq 1.
\]

with \(K\) the Kraft number. Conversely, for a given set of natural numbers \(N(x_1), \ldots, N(x_n)\) satisfying the above inequality, there exists a uniquely decodable code over an alphabet of size \(m\) with those codeword lengths.

The Kraft’s inequality for binary encoding \((m = 2)\) is

\[
K = \sum_{i=1}^{n} 2^{-N(x_i)} \leq 1.
\]
The following statements also hold:
1. Kraft’s inequality holds with **strict inequality** \( K < 1 \) if the code is **uniquely decodable** and has redundancy / is **incomplete**.
2. Kraft’s inequality holds with **equality** \( K = 1 \) if the code is **uniquely decodable** and has no redundancy / **complete**, and
3. Kraft’s inequality does not hold \( (K > 1) \) if the code is **not uniquely decodable**.

Proof. Consider the case of prefix codes. The proof can be extended to uniquely decodable encoding [14], but this is outside the scope of this course.

The proof uses the binary tree interpretation and counts the number of leaf nodes (nodes at the lowest level) that are covered by the prefix codes. The proof will be illustrated using Fig. 6.1.

The codes are first sorted such that \( N(x_1) \leq N(x_2) \leq \ldots \leq N(x_n) \). We can interpret all possible codes up to a length \( N(x_n) \) (the maximum length) in terms of leaf nodes of an \( m \)-ary tree of depth \( N(x_n) \). The total number of leaf nodes in this tree equals \( m^{N(x_n)} \).

To verify whether the code is uniquely decodable, we need to determine whether there remain enough available symbols after ruling out these prefixes. This is done in an iterative way by choosing a word of length \( N(x_i) \) arbitrarily, and then ruling out all words/nodes with that prefix.

First, choose any node from the full tree at depth \( N(x_1) \). This node corresponds to the first word of our new code (code 0 for \( x_1 \) in the example). Since we are building a prefix code, all the descendants leaf nodes of this node (i.e., all codes with length \( N(x_n) \) that have this first word as a prefix) become excluded in the code (codes \{000, 001, 010, 011\} in the example). As we consider the descendants up to the maximal depth \( N(x_n) \), there are \( m^{(N(x_n) - N(x_1))} \) descendant leaf nodes of that node that are removed from consideration (\( 2^2 \) in the example).

The next iteration picks a (surviving) node at depth \( N_2 \) and removes further \( m^{(N(x_n) - N(x_2))} \) leaf nodes, and so on. After \( n \) iterations, we have removed a total of

\[
\sum_{i=1}^{n} m^{(N(x_n) - N(x_i))}
\]

leaf nodes.

The question is whether we removed more leaf nodes than available \( m^{N(x_n)} \), or \( 2^3 \) in our example) in the process of building the code. Hence, if

\[
\sum_{i=1}^{n} m^{(N(x_n) - N(x_i))} \leq m^{N(x_n)}
\]

or equivalently

\[
\sum_{i=1}^{n} m^{-N(x_i)} \leq 1
\]

then there are more than enough leaf nodes to construct the code. \( \Box \)
Based on the proof, one can easily find an interpretation of the Kraft number for binary encoding

\[
K = \sum_{i=1}^{n} 2^{-N(x_i)}
\]

\[
= \frac{1}{2^{N(x_n)}} \sum_{i=1}^{n} 2^{(N(x_n) - N(x_i))}
\]

namely, it is the ratio between all the leaf nodes considered by the prefix code

\[
\sum_{i=1}^{n} 2^{(N(x_n) - N(x_i))}
\]

and the total number of leaf nodes in a binary tree of order \(N(x_n)\)

\[
2^{N(x_n)}.
\]

Table 6.1 shows the Kraft number for the example codes. From the Kraft number, it can be concluded that

- codes 1, 4, and 5 are complete and uniquely decodable,
- code 2 is not uniquely decodable, and
- code 3 is uniquely decodable, but has redundancy/is incomplete.

### 6.4 Binary prefix coding techniques

This section discusses three important binary prefix coding techniques, namely

**Shannon coding** (Section 6.4.1) which is sub-optimal but is used to proof Shannon’s fundamental source coding theorem (Section 6.5),

**Shannon-Fano coding** (Section 6.4.2) which has a better compression efficiency compared to the original Shannon coding,

**Huffman coding** (Section 6.4.3) which can be proven to be an optimal source coding technique.

#### 6.4.1 Shannon coding

Shannon coding is a lossless data compression technique for constructing a prefix code. It was the first of its kind and was used as the basis to proof Shannon’s source coding theorem in [1] (see also Theorem 15 and 16).
Algorithm 1 (Shannon coding). Shannon coding performs the following steps:

1. Sort the symbols $x_i$ from the most probable to the least probable $p(x_i) \geq p(x_{i+1})$.
2. Assign the length of the codewords to $N(x_i) = \lceil -\log p(x_i) \rceil$ with operator $\lceil x \rceil$ the ceiling function.
3. Assign the codewords of $y_i$ from the binary expansion of the cumulative probabilities

$$
\sum_{k=1}^{i-1} p(x_k).
$$

Shannon coding satisfies Kraft’s inequality by construction (see proof of Theorem 15). Hence, Shannon coding will always lead to a uniquely decodable prefix code.

Example 23. Table 6.2 gives the complete overview of a Shannon encoding. The entropy of the source is $H(X) = 2.0639$, the average code length $\langle N \rangle = 2.55$, and the Kraft number $K = 0.72$. Hence, the code is incomplete ($K < 1$) and its efficiency equals $e_X = 81\%$.

<table>
<thead>
<tr>
<th>$i$</th>
<th>$p(x_i)$</th>
<th>$-\log(x_i)$</th>
<th>$N(x_i)$</th>
<th>$\sum_{k=1}^{i-1} p(x_k)$</th>
<th>$y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.35</td>
<td>1.5146</td>
<td>2</td>
<td>0</td>
<td>00</td>
</tr>
<tr>
<td>2</td>
<td>0.30</td>
<td>1.7370</td>
<td>2</td>
<td>0.35</td>
<td>01</td>
</tr>
<tr>
<td>3</td>
<td>0.20</td>
<td>2.3219</td>
<td>3</td>
<td>0.65</td>
<td>101</td>
</tr>
<tr>
<td>4</td>
<td>0.10</td>
<td>3.3219</td>
<td>4</td>
<td>0.85</td>
<td>1101</td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>4.3219</td>
<td>5</td>
<td>0.95</td>
<td>11110</td>
</tr>
</tbody>
</table>

Table 6.2: Example of Shannon coding.

A comparative overview of the performance of the various encoding techniques on this example is shown in Table 6.3. Shannon coding is suboptimal as it does not achieve the lowest expected code length like Huffman coding. Furthermore, the expected code length of Shannon coding is never better (potentially equal to) the length of Shannon-Fano coding.

<table>
<thead>
<tr>
<th></th>
<th>$N(x_1)$</th>
<th>$\langle N \rangle$</th>
<th>$N(x_5)$</th>
<th>$e_X$</th>
<th>$K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shannon</td>
<td>2</td>
<td>2.55</td>
<td>5</td>
<td>81%</td>
<td>0.72</td>
</tr>
<tr>
<td>Shannon-Fano</td>
<td>2</td>
<td>2.15</td>
<td>3</td>
<td>96%</td>
<td>1.00</td>
</tr>
<tr>
<td>Huffman (1)</td>
<td>2</td>
<td>2.15</td>
<td>3</td>
<td>96%</td>
<td>1.00</td>
</tr>
<tr>
<td>Huffman (2)</td>
<td>1</td>
<td>2.15</td>
<td>4</td>
<td>96%</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 6.3: Performance comparison of the binary coding techniques for $p(x_i) = \{0.35, 0.30, 0.20, 0.10, 0.05\}$ ($H(X) = 2.0639$). The Shannon-Fano and Huffman (1) coding have, in this example, the same performance as they generate in identical codes.
6.4.2 Shannon-Fano coding

Shannon coding is a lossless data compression technique for constructing a prefix code and is named after Claude Shannon [1] and Robert Fano [15].

Algorithm 2 (Shannon–Fano coding). Shannon–Fano coding [15] performs the following steps:

1. Sort the symbols $x_i$ from the most probable to the least probable $p(x_i) \geq p(x_{i+1})$ and put them in a list $L = \{x_1, x_2, \ldots, x_n\}$.
2. Divide $L = \{x_k, \ldots, x_l\}$ into two lists $L_1 = \{x_k, \ldots, x_m\}$ and $L_2 = \{x_{m+1}, \ldots, x_l\}$ where $m$ is chosen such that the total probability of $L_1$ is as close as possible to the total probability of $L_2$. Hence, choose $m$ that minimizes

$$\left| \left( \sum_{i=k}^{m} p(x_i) \right) - \left( \sum_{i=m+1}^{l} p(x_i) \right) \right|.$$  

3. The symbols in $L_1$ are assigned a binary digit 0, the symbols of $L_2$ a digit 1. Hence, all codes of $L_1$ will start with 0; the codes of $L_2$ with 1.
4. Recursively apply steps 2 and 3 for both $L_1$ and $L_2$ until each symbol has become a corresponding code leaf on the tree.

Example 24. Table 6.4 gives the complete overview of a Shannon–Fano encoding. The algorithm starts with $L = \{x_1, x_2, x_3, x_4, x_5\}$. In the first iteration, the list $L$ is split up in $L_1 = \{x_1, x_2\}$ and $L_2 = \{x_3, x_4, x_5\}$. The second iteration first processes $L_1$ and (obviously) divides this into two singletons $\{x_1\}$ and $\{x_2\}$. Then, the list $L_2 = \{x_3, x_4, x_5\}$ is broken up into the lists $\{x_3\}$ and $\{x_4, x_5\}$.

<table>
<thead>
<tr>
<th>i</th>
<th>$p(x_i)$</th>
<th>iteration</th>
<th>$y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.35</td>
<td>0 0</td>
<td>00</td>
</tr>
<tr>
<td>2</td>
<td>0.30</td>
<td>0 1</td>
<td>01</td>
</tr>
<tr>
<td>3</td>
<td>0.20</td>
<td>1 0</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>0.10</td>
<td>1 1 0</td>
<td>110</td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>1 1 1</td>
<td>111</td>
</tr>
</tbody>
</table>

Table 6.4: Example of Shannon–Fano coding.

The entropy of the source is $H(X) = 2.0639$, the average code length $\langle N \rangle = 2.15$, and the Kraft number $K = 1.00$. Hence, the code is complete ($K = 1.00$) and its efficiency equals $e_X = 96\%$.

Shannon–Fano coding is suboptimal in the sense that it does not always achieve the lowest possible expected codeword length.

6.4.3 Huffman coding

The Shannon and Shannon–Fano algorithms don’t always generate an optimal code [16]. The Huffman code is a lossless compression technique that always
produces an optimal code. While the Shannon-Fano algorithm is created from the root to the leaves, the Huffman algorithm works in the opposite direction, namely from the leaves to the root.

The Huffman algorithm works by creating a binary tree of nodes (see e.g. Fig. 6.4). A node can be either a leaf node or an internal node. Initially, all nodes are leaf nodes which contain the symbol itself $x_i$, the probability $p_i$, and optionally a link to a parent node. The latter makes it easy to read the code starting from a leaf node. Internal nodes contain 1. a probability, 2. links to two child nodes, and 3. optionally a link to a parent node. A finished tree has up to $n$ leaf nodes and $n - 1$ internal nodes.

The simplest construction algorithm uses a priority queue where the node with lowest probability is given highest priority.

**Algorithm 3** (Huffman coding). Huffman coding performs the following steps:

1. Create a leaf node for each symbol and add it to the priority queue.
2. Select and remove the two nodes of lowest probability from the queue.
3. Create a new internal node with the two selected nodes as children and put the probability of this new internal node equal to the sum of the probabilities of the two child nodes.
4. Add the new node to the queue.
5. Repeat steps 2 to 4 until there is only one node left in the queue: the root node.
6. Generate the binary codes starting from the root node.

**Example 25.** Fig. 6.4 shows the graph that is generated for the case $p(x_i) = \{0.35, 0.30, 0.20, 0.10, 0.05\}$. The algorithm starts with a priority queue that contains the leaf nodes $\{x_1, \ldots, x_5\}$ (on the left of Fig. 6.4).

A first iteration selects the nodes with the lowest priority ($x_4$ and $x_5$) and generates a new (internal) node $x'_6$ with a probability $p(x'_6) = 0.15$. The priority queue is then updated by removing $x_4, x_5$ and adding $x'_6$.

The second iteration uses the priority queue $\{x_1, x_2, x_3, x'_6\}$, selects those with the lowest priority ($x_3$ and $x'_6$), and generates a new (internal) node $x'_7$ with a probability $p(x'_7) = 0.35$.

The third iteration starts with the priority queue $\{x_1, x'_7, x_2\}$. Note that $x_2$ is no longer at the second position as $p(x_1) = p(x'_7) = 0.35 \geq p(x_2) = 0.03$. The algorithm then has two options to combine two nodes with the lowest priority, namely by combining $x_1$ and $x_2$, or combining $x'_7$ and $x_2$. The former solution generates (after one addition iteration) the tree in Fig. 6.4, while the latter generates Fig. 6.5.

A Huffman code is in general not unique, but that they are optimal as can be seen in overview Table 6.3.
Figure 6.4: Huffman code tree for $p(x_i) = \{0.35, 0.30, 0.20, 0.10, 0.05\}$ (Huffman (1) in Table 6.3).

Figure 6.5: Alternative Huffman code tree for $p(x_i) = \{0.35, 0.30, 0.20, 0.10, 0.05\}$ (Huffman (2) in Table 6.3).

Table 6.5: The two possible solutions for Huffman coding for (1) Fig. 6.4 and (2) Fig. 6.5. The first solution (1) generates (for this example) exactly the same codes as the Shannon-Fano coding.
6.5 Shannon’s fundamental source coding theorem

Shannon’s source coding theorem (or noiseless coding theorem) establishes the limits to possible data compression. It shows that (in the limit, as the length of a stream of independent data tends to infinity) it is impossible to compress the data such that the code rate (average number of bits per symbol) is less than the Shannon entropy of the source. However it is possible to get the code rate arbitrarily close to the Shannon entropy, with negligible probability of loss.

The proof of this theorem is divided into two parts. First, the bounds are determined (Theorem 15) when using the Shannon coding (Section 6.4.1) to code the source. Afterwards, the result is generalized to a multiple of symbols (Theorem 16). The latter theorem shows that it is possible to get the code rate arbitrarily close to the Shannon entropy when the symbol lengths tend to infinity.

**Theorem 15** (Shannon’s fundamental source coding theorem - single symbol). Consider a discrete source $X$ with entropy $H(X)$. Shannon’s fundamental source coding theorem states that the average code length $\langle N \rangle$ to binary encode $(m = 2)$ a single symbols of $X$ satisfies

$$H(X) \leq \langle N \rangle < H(X) + 1.$$  \hspace{1cm} (6.1)

*Proof.* Assume a source $X$ with alphabet length $n$ that is binary encoded with an average code length $\langle N \rangle$. Consider the the Shannon coding of Section 6.4.1, a prefix code with the lengths $N(x_i)$ a natural number that equals the ceiling of the self-information

$$N(x_i) = \lceil - \log p(x_i) \rceil$$

where operator $\lceil x \rceil$ represents the ceiling function that outputs the least integer greater than or equal to $x$. This implies that $N(x_i)$ is bounded by

$$- \log p(x_i) \leq N(x_i) < - \log p(x_i) + 1.$$  \hspace{1cm} (6.2)

Inequality $- \log p(x_i) \leq N(x_i)$ in (6.2) can be used to show that

$$2^{-N(x_i)} \leq p(x_i)$$  \hspace{1cm} (6.3)

where the equality sign holds if $N(x_i) = - \log p(x_i)$ is a natural number. The chosen prefix code therefore satisfies the Kraft’s inequality (Theorem 14) since

$$\sum_{i=1}^{n} 2^{-N(x_i)} \leq \sum_{i=1}^{n} p(x_i) = 1.$$  

Kraft’s inequality then states that the chosen prefix code (the Shannon coding) must exists and that it is uniquely decodable.
a. Proof of $H(X) \leq \langle N \rangle$

Consider

$$H(X) - \langle N \rangle \overset{(a)}{=} - \sum_{i=1}^{n} p(x_i) \log p(x_i) - \sum_{i=1}^{n} p(x_i) N(x_i)$$

$$= \sum_{i=1}^{n} p(x_i) \log \left( \frac{2^{-N(x_i)}}{p(x_i)} \right) \overset{(b)}{=}$$

with (a) the definition of $H(X)$ and $\langle N \rangle$, and (b) using $x = \log (2^x)$. The argument in the log operator is always smaller or equal to than 1 using (6.3). Hence, one can us the inequality $\ln 2 \log x = \ln x \leq x - 1$ for $x \leq 1$ which results in

$$H(X) - \langle N \rangle \leq \frac{1}{\ln 2} \sum_{i=1}^{n} p(x_i) \left( \frac{2^{-N(x_i)}}{p(x_i)} - 1 \right)$$

$$\leq \frac{1}{\ln 2} \sum_{i=1}^{n} \left( 2^{-N(x_i)} - p(x_i) \right)$$

$$\overset{(a)}{=} \frac{1}{\ln 2} (K - 1)$$

where (a) uses the Kraft number and $\sum p(x_i) = 1$. Hence,

$$H(X) \leq \langle N \rangle + \frac{1}{\ln 2} (K - 1)$$

$$\overset{(a)}{=} \langle N \rangle .$$

where (a) uses the Kraft’s inequality for uniquely decodable encoding ($K \leq 1$).

b. Proof of $\langle N \rangle < H(X) + 1$

Inequality $N(x_i) < - \log p(x_i) + 1$ in (6.2) can be used to determine an upper bound for the the average code length

$$\langle N \rangle = \sum_{i=1}^{n} p(x_i) N(x_i)$$

$$< \sum_{i=1}^{n} p(x_i) (- \log p(x_i) + 1) = H(X) + 1. $$

The two parts of the above proof results into

$$H(X) \leq \langle N \rangle < H(X) + 1.$$ 

The optimal source coding will be obtained if the limit reaches the equality $H(X) = \langle N \rangle$. The efficiency $e_X$ is then 100%. The code is then as compact as possible, making the redundancy minimal.
Example 26. Consider a discrete source $X$ with $n = 3$ \((x_1, x_2, x_3)\) and $p(x_i) = \{0.45, 0.35, 0.20\}$. The Huffman code for this example is shown in Table 6.6 with the tree representation shown in Fig. 6.6. The performance of this code is $H(X) = 1.513$ bit/symbol, $\langle N \rangle = 1.55$ bit/symbol, and an efficiency of $e_X = 97.6\%$. Note that inequality (6.1) of Theorem 15 has been satisfied since

$$H(X) = 1.513 \leq \langle N \rangle = 1.55 < H(X) + 1 = 2.513.$$  

<table>
<thead>
<tr>
<th>$i$</th>
<th>$p(x_i)$</th>
<th>$y_i$</th>
<th>$N(x_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.45</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0.35</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0.20</td>
<td>11</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 6.6: Huffman coding for $p(x_i) = \{0.45, 0.35, 0.20\}$.

![Huffman code tree for $p(x_i) = \{0.45, 0.35, 0.20\}$](image)

Figure 6.6: Huffman code tree for $p(x_i) = \{0.45, 0.35, 0.20\}$.

Theorem 16 (Shannon’s source coding theorem - general [1]). Consider a block of independent symbols of a discrete source $X$ with entropy $H(X)$. Shannon’s fundamental source coding theorem states that the average code length $\langle N \rangle$ (per symbol) to binary encode \((m = 2)\) this block satisfies

$$H(X) \leq \langle N \rangle < H(X) + \varepsilon$$  \hspace{1cm} (6.4)

where $\varepsilon$ is a positive number. This $\varepsilon$ can be chosen arbitrarily small when the number of symbols in the block tends to infinity.

Proof. Consider $k$ independent symbols of $X$ in a block before binary encoding the block. This block of $k$ symbols will have an entropy of $kH(X)$. Applying Theorem 15 implies that a prefix code exists for the block which satisfies

$$kH(X) \leq k \langle N \rangle < kH(X) + 1.$$  

Dividing all terms by $k$ results into

$$H(X) \leq \langle N \rangle < H(X) + \varepsilon$$

with $\varepsilon = 1/k$.

Observe that $\varepsilon$ tends to zero if the block length tends to infinity ($k \to \infty$).
It is possible to extend the above results from binary to \( m \)-ary encoding, leading to

\[
\frac{H(X)}{\log m} \leq \langle N \rangle < \frac{H(X)}{\log m} + 1
\]

for (6.1) and

\[
\frac{H(X)}{\log m} \leq \langle N \rangle < \frac{H(X)}{\log m} + \varepsilon
\]

for (6.4).

**Example 27.** Consider the discrete source \( X \) of Example 26 and assume that the realizations are independent and combined pairwise. Hence, one can consider a new set of 9 symbols \( x_i x_j \) with a probability \( p(x_i)p(x_j) \) as shown in Table 6.7. The Huffman tree representation is illustrated in Fig. 6.7.

<table>
<thead>
<tr>
<th>( i,j )</th>
<th>( p(x_i)p(x_j) )</th>
<th>( y_{i,j} )</th>
<th>( N(x_i x_j) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 x_1 )</td>
<td>0.2025</td>
<td>00</td>
<td>2</td>
</tr>
<tr>
<td>( x_1 x_2 )</td>
<td>0.1575</td>
<td>100</td>
<td>3</td>
</tr>
<tr>
<td>( x_2 x_1 )</td>
<td>0.1575</td>
<td>111</td>
<td>3</td>
</tr>
<tr>
<td>( x_2 x_2 )</td>
<td>0.1225</td>
<td>110</td>
<td>3</td>
</tr>
<tr>
<td>( x_1 x_3 )</td>
<td>0.09</td>
<td>010</td>
<td>3</td>
</tr>
<tr>
<td>( x_3 x_1 )</td>
<td>0.09</td>
<td>1010</td>
<td>4</td>
</tr>
<tr>
<td>( x_2 x_3 )</td>
<td>0.07</td>
<td>1011</td>
<td>4</td>
</tr>
<tr>
<td>( x_3 x_2 )</td>
<td>0.07</td>
<td>0110</td>
<td>4</td>
</tr>
<tr>
<td>( x_3 x_3 )</td>
<td>0.04</td>
<td>0111</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 6.7: Huffman coding for paired symbols \( x_i x_j \) with \( p_i = \{0.45, 0.35, 0.20\} \) for a single symbol.
Table 6.7 can be used to compute the entropy of a symbol pair $2H(X) = 3.026 \text{ bit/symbol pair}$, the average code length $\langle N \rangle = 3.0675 \text{ bit/symbol}$, and an efficiency of $e_X = 98.6\%$. Note that inequality (6.4) of Theorem 16 is satisfied

$$H(X) = 1.513 \leq \langle N \rangle / 2 = 1.534 < H(X) + 1/2 = 2.013$$

and that the coding efficiency increased from $e_X = 97.6\%$ for a single symbol to $e_X = 98.6\%$ for a pair of symbols.
Chapter 7

Basic concepts in signal theory

The previous chapters studied the general communication model Fig. 1.1 from information point of view. Until now, the focus was on the overall communication (from source to destination), making abstraction of the properties of the signals that are generated by the transmitter, the noise which perturbs the transmitted signal, and the received signal at the input of the receiver.

The main issue with the signals in a communication system is that both the transmitted signal and the perturbing noise are often non-deterministic. The transmitted signal is a stochastic signal as it carries information (and hence entropy / randomness). Perturbing noise is often a stochastic signal by definition (e.g. thermal noise).

Not all signals in the communication systems are stochastic. The perturbing noise can also contain a deterministic component (e.g. a sinewave generated by a third party) and the transmitter / receiver might use deterministic signals to convert the stream of messages into a signal.

The aim of this chapter is to introduce the important concepts to characterize signals. After classifying the types of signals (Section 7.1), important concepts for stochastic signals are introduced in Section 7.2. The statistical properties of stochastic signals are described from two points of view:

1. the average over the ensemble of possible realizations of the signal (Section 7.3), and
2. the average over time, as the signal is a random variable that depends on time, (Section 7.4).

Practically, one can only observe one (or a few) realization(s) of the received message. The concepts of stationarity (Section 7.3.1) and ergodicity (Section 7.5) introduce the necessary conditions to be able to study the ensemble of all possible realizations from a single (infinitely long) observation.
Although the real world uses real-valued signals, all the derivations will be written for (possibly) complex valued signals. The motivation is that bandpass signals are typically represented using their complex-valued equivalent base-band/lowpass equivalent (Section 4.3).

7.1 Classification of signals

7.1.1 Deterministic signals

The values of a deterministic signal can be obtained at any time instance using a mathematical model. In the simplest case this can be a formula of which time $t$ is one of the parameters. Hence, no statistical analysis is required.

As a deterministic signal is exactly known, there is nothing stochastic in the signal and hence its entropy equals zero. Deterministic signals are therefore unable to transport any information.

Example 28. Consider the signal $s(t) = A_0 \cos(\omega_0 t + \varphi_0)$ with $A_0, \omega_0, \varphi_0$ constant. When $A_0, \omega_0,$ and $\varphi_0$ are known, it is possible to compute the value the signal $s(t)$ for any freely chosen time instant $t$. Hence, there is no advantage to transmit such signal over the channel to transport information as it does not carry any.

7.1.2 Stochastic signals

The value of a stochastic signal over time cannot be computed from a deterministic model. This means that the behavior of the signal is (partially) unpredictable. The properties of the signals can, however, be described using statistical methods. Stochastic signals are therefore treated as statistical quantities over both time, and the realization (e.g. the message) of the signal.

It should be noted that the average value of a stochastic signal can’t be used to convey information. It is therefore custom in signal theory to study zero-mean stochastic signals.

7.1.3 Periodic signals

Periodic signals is an important subclass of signals that satisfy $x(t) = x(t + T) \quad \forall t$ where $T$ represents the period of the periodic signal. They form the basis for harmonic analysis and are highly important in theoretical studies. Note that $x(t)$ can be both deterministic, or a realization of a stochastic process that is repeated every period $T$. 
Example 29 (Fourier analysis of square wave). An ideal square wave with an amplitude of $\pm A$, period $T$, and a duty-cycle of 50% can be represented as a sum of sinusoidal waves

$$x(t) = \frac{4A}{\pi} \sum_{k=1}^{\infty} \frac{1}{2k-1} \sin \left( 2\pi (2k-1) \frac{t}{T} \right).$$

Example 30 (PRBS). A Pseudo-Random Binary Sequence is a binary sequence (having two levels, 0 and 1) which has properties (the auto-correlation of the sequence) similar to noise, although the sequence is generated using a deterministic algorithm using a linear-feedback shift register which comprises binary shift registers and EXOR operations [11, 19]. An example of such linear-feedback shift register (LFSR) is shown in Fig. 7.1. Although its behavior is similar to a stochastic signal (see its auto-correlation), it remains a deterministic sequence with a fixed periodicity. Hence, a PRBS is a periodic signal.

![Figure 7.1](image)

Figure 7.1: A Pseudo-Random Binary Sequence can be generated using a linear-feedback shift register and results in a noise-like binary sequence. The maximum-length shift register sequences repeats itself every $n = 2^m - 1$ cycles.

Most PRBS sequences use the maximum-length shift register sequences, or m-sequence for short. They are generated by an $m$-stage shift register with linear feedback. They have a periodicity of $n = 2^m - 1$ which contain $2^{m-1} - 1$ zeros and $2^{m-1}$ ones. While outputting all these $2^m - 1$ samples, the shift register runs over all possible $2^m - 1$ non-zero states. An all zero state can’t be used since this state result in a 0 which would be shifted in (and hence keeping the shift register in an all zero state).

### 7.2 Properties of stochastic signals

A stochastic signal is a process that develops in time and where (at least for part) a probabilistic behavior can be noted. Mathematically, a stochastic signal generated by a process will depend on two variables

$$X(k, t) = X^{(k)}(t)$$

(7.1)
where $k$ represents the values generated by the stochastic space and $t$ the time.

At any arbitrarily time instance $t = t_i$, the value of the signal $X^{(k)}(t_i)$ is a stochastic variable generated by the stochastic space. Such attribution of values is called the realization of the process. Time instances can be either discrete for time instances $t_1, t_2, \ldots, t_n$ or continuous for arbitrarily values of $t$.

The set of all possible values / realizations of $k$ will be called the ensemble of the signals.

**Example 31.** Assume the transmission of a message code $x_k$ and assume that this corresponds to transmitting the signal $X^{(k)}(t)$. The signal $X^{(k)}(t)$ is then the realization of the stochastic process. The set of all signals over all the possible code realizations of $k$ equals the ensemble.

To study the statistical properties of stochastic signals one must obviously rely on statistical techniques starting from the classical statistical concepts for a stochastic variable $X$:

**Probability** $P[X \leq x]$

**Cumulative distribution function**

$$F_X(x) = P[X \leq x]$$

**Probability density function**

$$f_X(x) = \frac{\partial F_X(x)}{\partial x}$$

**Expected value**

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$

**Variance**

$$\sigma_{XX}^2 = E((X - E(X))^2) = E(X^2) - \mu_X^2$$

**Covariance**

$$\sigma_{XY}^2 = E((X - E(X))(Y - E(Y))) = E(XY) - \mu_X \mu_Y$$

**Raw second-order moment**

$$\rho_{XY}^2 = E(XY) = \sigma_{XY}^2 + \mu_X \mu_Y$$

**Independent**

$$f_{XY}(x, y) = f_X(x)f_Y(y)$$
CHAPTER 7. BASIC CONCEPTS IN SIGNAL THEORY

Uncorrelated

\[ \sigma_{XY}^2 = 0 \]

**Definition 29.** The probability that the value of a stochastic variable \( X(t) \) at the time instance \( t_1 \) is smaller than a constant value \( x_1 \) equals \( P[X(t_1) \leq x_1] \).

This probability will depend both on the chosen time instance \( t_1 \) and on the value \( x_1 \). This will be indicated by specifying the two variables \( x_1; t_1 \) as argument.

**Definition 30** (cumulative distribution function (cdf)). The cumulative distribution function

\[ F_X(x_1; t_1) = P[X(t_1) \leq x_1] \]

(7.2)

determines the probability \( P[X(t_1) \leq x_1] \) for all possible values of \( x_1 \).

This cumulative distribution function is then used to define the probability density function.

**Definition 31** (probability density function (pdf)). The probability density function is obtained by taking the (partial) derivative of the cumulative distribution function

\[ f_X(x_1; t_1) = \frac{\partial F_X(x_1; t_1)}{\partial x_1}. \]

These definitions can easily be extended towards multiple variables

**Definition 32.** The multivariate cumulative distribution function of \( n \) random variables equals

\[ F_{X_1 \ldots X_n}(x_1 \ldots x_n; t_1 \ldots t_n) = P[X(t_1) \leq x_1 \cap \ldots \cap X(t_n) \leq x_n] \]

(7.3)

i.e. the probability that \( X(t_1) \leq x_1, \text{ and } X(t_2) \leq x_2, \ldots, \text{ and } X(t_n) \leq x_n \).

The multivariate probability density function (pdf) of \( n \) random variables is obtained by taking the (partial) derivative of the cumulative distribution function with respect to each random variable \( x_1, \ldots, x_n \)

\[ f_{X_1 \ldots X_n}(x_1 \ldots x_n; t_1 \ldots t_n) = \frac{\partial^n F_{X_1 \ldots X_n}(x_1 \ldots x_n; t_1 \ldots t_n)}{\partial x_1 \partial x_2 \ldots \partial x_n}. \]

(7.4)

If \( X \) and \( Y \) are independent variables, then

\[ f_{XY}(x_1, y_2; t_1, t_2) = f_X(x_1; t_1)f_Y(y_2; t_2). \]

### 7.3 Ensemble averages

Ensemble averages are simply the statistical moments (first order, second order or mixed moments) of the stochastic process. These statistical moments are defined with respect to the expected value over all the realizations of \( X(t) \). This expected value will be represented by the operator \( E(\cdot) \).
In order to study complex-valued signals, we need to be able to take the expected value for complex-valued quantities. When considering the expected value over complex-value quantities \( z = x + jy \) requires the integration over both the real and imaginary part to obtain the expected value

\[
E(g(z)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x + jy) f_Z(x + jy) \, dx \, dy
\]

\[
= \int_{\mathbb{C}} g(z) f_Z(z) \, dS_z
\]

where \( dS_z \) represents an elementary surface \( dx \, dy \) in the complex plane, and where \( \int_{\mathbb{C}} dS_z \) represents the integration over the complete complex plane.

**Definition 33** (Ensemble averages).

1. The (ensemble) **mean** value is the first order moment

\[
\mu_X(t_1) = E(X(t_1)) = \int_{\mathbb{C}} x_1 f_X(x_1; t_1) \, dS_{x_1}
\]

for \( x_1 \in \mathbb{C} \)

\[
\mu_X(t_1) = E(X(t_1)) = \int_{-\infty}^{\infty} x_1 f_X(x_1; t_1) \, dx_1
\]

for \( x_1 \in \mathbb{R} \).

2. The (ensemble) **quadratic moment**

\[
E(|X(t_1)|^2) = E(X(t_1)X^*(t_1)) = \int_{\mathbb{C}} x_1 x_1^* f_X(x_1; t_1) \, dS_{x_1}
\]

for \( x_1 \in \mathbb{C} \)

\[
E(|X(t_1)|^2) = \int_{-\infty}^{\infty} x_1^2 f_X(x_1; t_1) \, dx_1
\]

for \( x_1 \in \mathbb{R} \).

3. The (ensemble) **auto-correlation**

\[
\rho_{XX}(t_1, t_2) = E(X(t_1)X^*(t_2)) = \int_{\mathbb{C}} \int_{\mathbb{C}} x_1 x_2^* f_{XX}(x_1, x_2; t_1, t_2) \, dS_{x_1} \, dS_{x_2}
\]

for \( x_1, x_2 \in \mathbb{C} \)

\[
\rho_{XX}(t_1, t_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 f_{XX}(x_1, x_2; t_1, t_2) \, dx_1 \, dx_2
\]

for \( x_1, x_2 \in \mathbb{R} \).
4. The (ensemble) cross-correlation between two different processes (signals) $X$ and $Y$

$$\rho_{XY}(t_1, t_2) = E(X(t_1)Y^*(t_2))$$

$$= \int_{\mathbb{C}} \int_{\mathbb{C}} x_1 y_2^* f_{XY}(x_1, y_2; t_1, t_2) dS_{x_1} dS_{y_2}$$

for $x_1, y_2 \in \mathbb{C}$

$$\rho_{XY}(t_1, t_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 y_2 f_{XY}(x_1, y_2; t_1, t_2) dx_1 dy_2$$

for $x_1, y_2 \in \mathbb{R}$.
If $X$ and $Y$ are independent variables, then

$$\rho_{XY}(t_1, t_2) = \mu_X(t_1)\mu_Y^*(t_2)$$

5. The (ensemble) variance

$$\sigma_X^2(t_1) = E \left( |X(t_1) - E(X(t_1))|^2 \right)$$

$$= E(X(t_1)X^*(t_1)) - |E(X(t_1))|^2$$

$$= \rho_{XX}(t_1, t_1) - |\mu_X(t_1)|^2$$

Note that

$$\rho_{XY}(t_1, t_2) = \rho_{YX}^*(t_2, t_1)$$

and that $\rho_{XY}(t_1, t_2)$ is real-valued for real-valued signals $X(t)$ and $Y(t)$.

### 7.3.1 Stationary signals

The stationarity of the signal expresses how the stochastic properties of the process change over time. Two main classes can be distinguished: the strict stationary (=strong stationary) and weak stationary (wide-sense stationary, abbreviated as WSS). The former demands that all (also higher order) pdf’s are invariant with respect to a time shift, while the latter only demands the pdf’s up to order two to be invariant to a time shift.

Stationarity of a process can be expressed mathematically by considering a time shift $\tau$ on the pdf’s of that stochastic processes

$$f_{X,\ldots,X}(x_1, \ldots, x_n; t_1, \ldots, t_n) = f_{X,\ldots,X}(x_1, \ldots, x_n; t_1 + \tau, \ldots, t_n + \tau) \quad (7.5)$$

**Definition 34** (Strictly or strong stationary process). A stochastic process is strictly or strongly stationary process if the pdf’s satisfy (7.5) for all possible time shifts $\tau$ and this for all orders $n = 1, \ldots, \infty$. 

Note that the \( n^{th} \) order pdf’s of stationary processes can be rewritten as function of the time difference by substituting \( \tau = -t_1 \) in (7.5)

\[
f_{X \ldots X}(x_1, \ldots, x_n; t_1, \ldots, t_n) = f_{X \ldots X}(x_1, \ldots, x_n; 0, t_2 - t_1, \ldots, t_n - t_1). \tag{7.6}
\]

or using \( \tau = -t_2 \)

\[
f_{X \ldots X}(x_1, \ldots, x_n; t_1, \ldots, t_n) = f_{X \ldots X}(x_1, \ldots, x_n; t_1 - t_2, 0, \ldots, t_n - t_2).
\]

A less stringent condition is also used in practice which only considers the pdf’s up the second order.

**Definition 35** (Weakly or wide-sense stationary (WSS) process). A stochastic process is weakly or wide-sense stationary process if the pdf’s satisfy (7.5) for all possible time shifts \( \tau \) and this for orders \( n = 1, 2 \).

Hence, WSS processes satisfy

\[
f_X(x_1; t_1) = f_X(x_1)
\]

and

\[
f_{XX}(x_1, x_2; t_1, t_2) = f_{XX}(x_1, x_2; t_1 - t_2, 0).
\]

Another type of stationarity which is often encountered is cyclostationarity.

**Definition 36** (Cyclostationary process). A stochastic process is strictly or strongly stationary process with periodicity \( T \) if the pdf’s satisfy (7.5) for all shifts \( \tau = T \). Strictly or strong cyclostationarity demands this for all \( n = 1, \ldots, \infty \), while weakly or wide-sense cyclostationarity only demands this for \( n = 1, 2 \).

**Example 32.** Consider an electronic device (e.g. a transistor in an amplifier) that generates

- **thermal noise**: independent Gaussian noise which depends on the temperature of the device, and
- **shot noise**: noise that originates from the discrete nature of electric charge and hence depends on the current through the device.

Consider now the following cases:

1. Both the temperature and the current through the device are constant. Then, both the thermal noise and the shot noise will have pdf’s which are invariant over time. Hence, the generated noise is stationary.

2. The temperature changes in an arbitrary way (e.g. temperature variations induced by the stochastic behavior of the ambient temperature). The variance of the thermal noise will then change over time and hence the thermal noise is now non-stationary.

3. Constant temperature but with a sinusoidal current running through the device. This is the case in e.g. frequency converters / mixers to convert signals from one frequency band to another. As the pdf’s of the shot noise change with the current, and since the current is a periodic signal, it can be concluded that the shot noise is now a cyclostationary process.
7.3.2 Properties for wide-sense stationary signals

The following properties are valid for wide-sense stationary processes $X(t)$.

**Theorem 17.** The mean value and the variance over the ensemble of a wide-sense stationary (WSS) process $X(t)$ are constant, i.e.

\[
\mu_X(t_1) = \mu_X \\
\sigma_X^2(t_1) = \sigma_X^2
\]

for all $t_1$.

**Proof.** The mean value equals

\[
\mu_X(t_1) = \int_C x_1 f_X(x_1; t_1) dS_{x_1} = \int_C x_1 f_X(x_1; t_1) dS_{x_1} \forall t_1 = \mu_X = \text{Cte}
\]

The variance is expressed as

\[
\sigma_X^2(t_1) = E\left(\left|X(t_1) - E(X(t_1))\right|^2\right) = \int_C |x_1 - \mu_X|^2 f_X(x_1; t_1) dS_{x_1} = \int_C |x_1 - \mu_X|^2 f_X(x_1) dS_{x_1} \forall t_1 = \sigma_X^2 = \text{Cte}
\]

**Theorem 18.** The auto-correlation over the ensemble of a wide-sense stationary (WSS) process $X(t)$ only depends on the time difference $\tau = t_1 - t_2$

\[
\rho_{XX}(t_1, t_2) = \rho_{XX}(\tau) = \rho_{XX}^*(-\tau).
\]

If $X(t)$ is real-valued, then $\rho_{XX}(\tau) = \rho_{XX}^*(-\tau)$ is an even, real-valued function.

**Proof.** The auto-correlation satisfies

\[
\rho_{XX}(t_1, t_2) = E(X(t_1)X^*(t_2)) = \int_C \int_C x_1 x_2^* f_{XX}(x_1, x_2; t_1, t_2) dS_{x_1} dS_{x_2} \equiv a \int_C \int_C x_1 x_2^* f_{XX}(x_1, x_2; t_1 - t_2, 0) dS_{x_1} dS_{x_2} = E(X(+\tau)X^*(0)), \quad \tau = t_1 - t_2 = \rho_{XX}(\tau)
\]
where (a) uses the stationary property. Furthermore,
\[
\rho_{XX}(\tau) = E(X(\tau)X^*(0)) = E(X(0)X^*(-\tau)) = (E(X(-\tau)X^*(0)))^* = \rho^*_X(-\tau).
\]

If \(X\) is real-valued, then \(\rho_{XX}(\tau)\) is an even, real-valued function with \(\rho_{XX}(\tau) = \rho_{XX}(-\tau)\).

**Theorem 19.** The cross-correlation over the ensemble of wide-sense stationary (WSS) processes \(X(t)\) and \(Y(t)\) only depends on the time difference \(\tau = t_1 - t_2\)
\[
\rho_{XY}(t_1, t_2) = \rho_{XY}(\tau) = \rho^*_Y(-\tau)
\]
or equivalently
\[
\rho_{XY}(-\tau) = \rho^*_Y(\tau).
\]
If \(X(t)\) and \(Y(t)\) are real-valued, then \(\rho_{XY}(\tau) = \rho_{YX}(-\tau)\).

**Proof.** This can be proven starting from the definition of the cross-correlation \(\rho_{XY}(t_1, t_2)\) with \(\tau = t_1 - t_2\)
\[
\rho_{XY}(\tau) = E(X(+\tau)Y^*(0)) = E(Y^*(-\tau)X(0)) = (E(Y(-\tau)X^*(0)))^* = \rho^*_Y(-\tau).
\]

**Theorem 20.** If the wide-sense stationary (WSS) process \(X(t)\) is purely stochastic process then its asymptotic value equals
\[
\lim_{\tau \to \infty} \rho_{XX}(\tau) = |\mu_X|^2
\]
and its variance is
\[
\sigma^2_X = \rho_{XX}(0) - |\mu_X|^2
\]

**Proof.** Consider two stochastic variables \(X(t_1)\) and \(X(t_2)\) at two different time instances \(t_1\) and \(t_2\) that are \(\tau = t_1 - t_2\) separated from each other. If the signal is purely stochastic, and hence has no deterministic component, and if the interval \(\tau\) tends to infinity, then the dependencies of the stochastic variables \(X(t_1)\) and \(X(t_2)\) become less and less pronounced. In the limit, one can assume for \(\tau \to \infty\) that they become stochastic independent. Hence,
\[
\lim_{\tau \to \infty} \rho_{XX}(\tau) = \lim_{\tau \to \infty} E(X(t + \tau)X^*(t)) = \lim_{\tau \to \infty} E(X(t + \tau))E(X^*(t)) = |\mu_X|^2.
\]
The variance of the (stationary) signal equals to
\[
\sigma_X^2 = \sigma_X^2(t_1) = E \left( |X(t_1) - E(X(t_1))|^2 \right) = E (X(t_1)X^*(t_1)) - |E(X(t_1))|^2 = \rho_{XX}(0) - |\mu_X|^2.
\]

Hence, the asymptotic value of the auto-correlation \(\rho_{XX}(\tau)\) for \(\tau \to \infty\) of a noise signal can be used to estimate its mean \(\mu_X\). Its variance can be determined by the difference of the auto-correlation at \(\tau = 0\) (\(\rho_{XX}(0)\)) and at infinity (\(\rho_{XX}(\infty)\)) as illustrated in Fig. 7.2.

![Figure 7.2: Example of an auto-correlation with \(\rho_{XX}(\tau) = \text{sinc} \tau = \sin(\pi \tau)/(\pi \tau)\) and \(\mu_X = 0.5\).](image.png)

**Theorem 21.** The continuous auto-correlation function reaches its peak at the origin at \(\tau = 0\)
\[
|\Re(\rho_{XX}(\tau))| \leq \rho_{XX}(0) \quad \forall \tau.
\]
For real-valued processes this becomes
\[
|\rho_{XX}(\tau)| \leq \rho_{XX}(0) \quad \forall \tau.
\]
Proof. Consider the expected value of $|X(t + \tau) \pm X(t)|^2$. As this norm is always non-negative, it is easy to see that

$$E \left( |X(t + \tau) \pm X(t)|^2 \right) \geq 0$$

and hence

$$E \left( |X(t + \tau) \pm X(t)|^2 \right) = E \left( |X(t + \tau)|^2 \right) + E \left( |X(t)|^2 \right)$$

$$\pm E (X(t + \tau)X^*(t)) \pm E (X(t)X^*(t + \tau))$$

$$= 2\rho_{XX}(0) \pm (\rho_{XX}(\tau) + \rho_{XX}^*(\tau))$$

$$= 2\rho_{XX}(0) \pm 2\Re(\rho_{XX}(\tau)) \geq 0.$$

This inequality can only be satisfied if

$$|\Re(\rho_{XX}(\tau))| \leq \rho_{XX}(0)$$

for all values of $\tau$.

Real-valued processes have a real-valued auto-correlation, implying that

$$|\rho_{XX}(\tau)| \leq \rho_{XX}(0) \ \forall \tau.$$

This property is also illustrated in Fig. 7.2.

**Definition 37** (Uncorrelated processes). Two processes $X(t)$ and $Y(t)$ are called uncorrelated if $\rho_{XY}(t_1, t_2) = \mu_X(t_1)\mu_Y^*(t_2) \ \forall t_1, t_2$.

If two processes are independent, then they are also uncorrelated. It should be stressed that the opposite is not always true: uncorrelated processes are not necessarily independent!

**Definition 38** (Orthogonal processes). Two processes $X(t)$ and $Y(t)$ are called orthogonal if $\rho_{XY}(t_1, t_2) = E(X(t_1)Y^*(t_2)) = 0 \ \forall t_1, t_2$.

**Lemma 19.** Two processes $X(t)$ and $Y(t)$ are orthogonal if $X(t)$ and $Y(t)$ are uncorrelated and if one of the processes is zero-mean wide-sense stationary.

**Proof.** Uncorrelated processes $X(t)$ and $Y(t)$ implies that

$$\rho_{XY}(t_1, t_2) - \mu_X(t_1)\mu_Y^*(t_2) = 0.$$ 

Suppose that $X(t)$ is WSS and zero mean, then $\mu_X(t) = 0$ for all $t$. Hence,

$$\rho_{XY}(t_1, t_2) = \mu_X(t_1)\mu_Y^*(t_2) = 0$$

for all $t_1$ and $t_2$. 

Theorem 22. The auto-correlation of the sum \( Z(t) = X(t) + Y(t) \) of two orthogonal signals \( X(t) \) and \( Y(t) \) (the cross-correlation \( \rho_{XY}(\tau) = 0 \) for all \( \tau \)) equals the sum of the auto-correlations of each signal separately
\[
\rho_{ZZ}(\tau) = \rho_{XX}(\tau) + \rho_{YY}(\tau).
\]

Proof. This follows directly out of the definition of the auto-correlation
\[
\rho_{ZZ}(\tau) = E((X(t+\tau) + Y(t+\tau))(X^*(t) + Y^*(t)))
\]
\[
= E(X(t+\tau)X^*(t)) + E(X(t+\tau)Y^*(t))
\]
\[
+ E(Y(t+\tau)X^*(t)) + E(Y(t+\tau)Y^*(t))
\]
\[
= \rho_{XX}(\tau) + \rho_{XY}(\tau) + \rho_{YX}(\tau) + \rho_{YY}(\tau)
\]
\[
= \rho_{XX}(\tau) + \rho_{YY}(\tau).
\]

Theorem 23. The auto-correlation of the product \( Z(t) = X(t)Y(t) \) of two independent signals \( X(t) \) and \( Y(t) \) (independent for all \( \tau \)) equals the product of the auto-correlations of each signal separately
\[
\rho_{ZZ}(\tau) = \rho_{XX}(\tau)\rho_{YY}(\tau).
\]

Proof. This follows directly out of the definition of the auto-correlation
\[
\rho_{ZZ}(\tau) = E(X(t+\tau)Y(t+\tau)(X^*(t)Y^*(t))]
\]
\[
\overset{(a)}{=} E(X(t+\tau)X^*(t))E(Y(t+\tau)Y^*(t))
\]
\[
= \rho_{XX}(\tau)\rho_{YY}(\tau)
\]

where (a) uses the fact that \( E(XY) = E(X)E(Y) \) if \( X \) and \( Y \) are independent random variables.

7.3.3 Auto-correlation after linear time-invariant (LTI) filtering

Linear system theory learns use that a Linear Time-Invariant (LTI) system can relates it’s output \( y(t) \) with the input \( x(t) \) through the convolution with the impulse response \( h(t) \) of the LTI system
\[
y(t) = \int_{-\infty}^{\infty} h(u)x(t-u)du = \int_{-\infty}^{\infty} h(t-u)x(u)du.
\]

In physical systems, the impulse response is causal (\( h(t) = 0 \) for \( t < 0 \)) and hence
\[
y(t) = \int_{0}^{\infty} h(u)x(t-u)du = \int_{-\infty}^{t} h(t-u)x(u)du.
\]
Theorem 24 (Correlation after LTI filtering). If $X(t)$ is a WSS signal passing through a LTI system with impulse response $h(t)$, then

$$
\mu_Y = \mu_X \int_{-\infty}^{\infty} h(t) \, dt
$$

$$
\rho_{YX}(\tau) = h(\tau) * \rho_{XX}(\tau)
$$

$$
\rho_{XY}(\tau) = h^*(-\tau) * \rho_{XX}(\tau)
$$

$$
\rho_{YY}(\tau) = h(\tau) * h^*(-\tau) * \rho_{XX}(\tau)
$$

with $*$ the convolution product.

Proof. The mean value

$$
\mu_Y = E(Y(t)) = \int_{-\infty}^{\infty} h(u) E(X(t-u)) \, du
$$

$$
= \mu_X \int_{-\infty}^{\infty} h(u) \, du
$$

The cross-correlation between $X(t)$ and $Y(t)$

$$
\rho_{YX}(\tau) = E(Y(t+\tau)X^*(t))
$$

$$
= E \left( \left( \int_{-\infty}^{\infty} h(u) X(t+\tau-u) \, du \right) X^*(t) \right)
$$

$$
= \int_{-\infty}^{\infty} h(u) E(X(t+\tau-u)X^*(t)) \, du
$$

$$
= \int_{-\infty}^{\infty} h(u) \rho_{XX}(\tau-u) \, du
$$

$$
= h(\tau) * \rho_{XX}(\tau)
$$

and

$$
\rho_{XY}(\tau) = \rho_{YX}(-\tau)
$$

$$
= h^*(-\tau) * \rho_{XX}^*(-\tau)
$$

$$
= h^*(-\tau) * \rho_{XX}(\tau)
$$

The auto-correlation of $Y(t)$

$$
\rho_{YY}(\tau) = E(Y(t+\tau)Y^*(t))
$$

$$
= E \left( \left( \int_{-\infty}^{\infty} h(u) X(t+\tau-u) \, du \right) Y^*(t) \right)
$$

$$
= \int_{-\infty}^{\infty} h(u) E(X(t+\tau-u) Y^*(t)) \, du
$$

$$
= \int_{-\infty}^{\infty} h(u) \rho_{XY}(\tau-u) \, du
$$

$$
= h(\tau) * \rho_{XY}(\tau)
$$

$$
= h(\tau) * h^*(-\tau) * \rho_{XX}(\tau).
$$
The auto-correlation of a periodic signal with period $T$ satisfying
\[ X(t) = X(t + T) \quad \forall t \]
is itself periodic with the same period
\[ \rho_{XX}(\tau) = \rho_{XX}(\tau + T). \]

Proof. The proof is quite trivial since
\[ \rho_{XX}(\tau) = E(X(t + \tau)X^*(t)) \]
\[ = E(X(t + \tau + T)X^*(t)) \]
\[ = \rho_{XX}(\tau + T) \]
where (a) uses the property that $X(t) = X(t + T)$ for all $t$.

Example 33 (Square wave with duty cycle). Consider a square wave with period $T$, duty cycle $D$, and amplitude level $A$ for the high value and 0 for the low ones. This square wave can be written as
\[ X(t - t_0) = \begin{cases} A & 0 \leq t < DT \\ 0 & DT \leq t < T \end{cases} \]
with the periodic constraint that $X(t) = X(t + T)$. The start time $t_0$ is a random variable which is uniformly distributed between 0 and $T$. Note that $t_0$ represent the time instance of the rising edge of the square wave.

The computation of the auto-correlation can be limited to $\tau$ in the range between 0 and $T$ as it is known that $\rho_{XX}(\tau)$ is periodic with periodicity $T$ ($\rho_{XX}(\tau) = \rho_{XX}(\tau + T)$).

For $|\tau| < DT$, the high value (non-zero) of the signals $X(t + \tau - t_0)$ and $X(t - t_0)$ overlap in a time window $DT - |\tau|$.
\[ \rho_{XX}(\tau) = 1 \int_{0}^{DT-|\tau|} X(t + \tau - t_0)X(t - t_0)dt_0. \]

For $|\tau| > DT$, the signals overlap with at least 1 zero-values signal. Hence, the integration over the time with a shift $\tau$ equals
\[ \rho_{XX}(\tau) = \begin{cases} A^2(D - |\tau|) & |\tau| < DT \\ 0 & DT \leq |\tau| \leq T - DT \end{cases} \]
and $\rho_{XX}(\tau) = \rho_{XX}(\tau + kT)$ for all integers $k$. 
Fig. 7.3 provides a graphical representation of the considered square wave and Fig. 7.4 shows the corresponding auto-correlation. Note that the periodicity of the square wave signals equals the periodicity of the auto-correlation function.

Figure 7.3: Illustration of the square wave with amplitude $A = 2$ and a duty cycle of $D = 25\%$. The upper signal is the original signal $X(t)$, the lower the signal $X(t + \tau)$ as used in the computation of the auto-correlation with $\tau = 0.125T$. 
Example 34. A PRBS (introduced Example 30) is a binary periodic signal with period \( n \) such that \( x(t_i) = x(t_{i+n}) \). To easily characterize this sequence, consider the bipolar sequence \( \{-1, 1\} \) given by

\[
y(t_i) = 2x(t_i) - 1.
\]

The auto-correlation function then equals [11]

\[
\rho_{YY}(j) = \sum_{i=1}^{n} y(t_{i+j})y(t_i)
\]

for \( 0 \leq j < n \) and has a periodicity of \( n \). Ideally, a pseudo-random sequence should have an auto-correlation function which satisfies

\[
\rho_{YY}(j) = \begin{cases} 
1, & j = 0, \\
0, & 0 < j < n.
\end{cases}
\]

For a PRBS, the periodic auto-correlation function equals

\[
\rho_{YY}(j) = \begin{cases} 
n, & j = 0, \\
-1, & 0 < j < n.
\end{cases}
\]

which approximates the ideal pseudo-random sequence for large \( n \). Fig. 7.5 shows this auto-correlation function for \( m = 5 \), resulting in a periodic auto-correlation with period \( n = 2^5 - 1 = 31 \).
Figure 7.5: Auto-correlation for a maximum length PRBS signal with \( k = 5 \) and hence \( n = 31 \).

### 7.3.5 Modulation of WSS process

**Theorem 26 (Amplitude Modulation).** Consider the amplitude modulation of a real-valued WSS process \( X(t) \)

\[
Y(t) = X(t) \cos(2\pi f_c t + \phi)
\]

with a fixed frequency \( f_c \) and with an initial phase \( \phi \) which is uniformly distributed between 0 and \( 2\pi \) and independent of \( X(t) \). The amplitude modulated signal \( Y(t) \) is zero-mean WSS with

\[
\rho_{YY}(\tau) = \frac{1}{2} \rho_{XX}(\tau) \cos(2\pi f_c \tau).
\]

**Proof.** The mean value of \( Y(t) \) equals zero for all \( t \) since

\[
E(Y(t)) = E(X(t) \cos(2\pi f_c t + \phi)) = E(X(t)) E(\cos(2\pi f_c t + \phi)) = 0.
\]
The correlation function equals
\[ \rho_{YY}(\tau) = E(X(t + \tau)X(t)) \cos(2\pi f_c(t + \tau) + \phi) \cos(2\pi f_c t + \phi) \]
\[ = E(X(t + \tau)X(t)) E(\cos(2\pi f_c(t + \tau) + \phi) \cos(2\pi f_c t + \phi)) \]
\[ = \frac{1}{2} \rho_{XX}(\tau) E(\cos(2\pi f_c t + 2\phi) + \cos(2\pi f_c t + 2\phi)) \]
\[ = \frac{1}{2} \rho_{XX}(\tau) \cos(2\pi f_c \tau) + \frac{1}{2} \rho_{XX}(\tau) \int_{0}^{2\pi} \cos(2\pi f_c (2t + \tau) + 2\phi) \frac{d\phi}{2\pi} \]
\[ = \frac{1}{2} \rho_{XX}(\tau) \cos(2\pi f_c \tau) \]
showing that \( Y(t) \) is wide-sense stationary.

Note that the auto-correlation has the same period \( 1/f_c \) as the original sine wave, and that the power of the signal equals \( \rho_{YY}(0) = \rho_{XX}(0)/2 \).

**Theorem 27 (IQ Modulation).** Consider the IQ modulation of a two real-valued mutual WSS processes \( I(t) \) and \( Q(t) \)
\[ Y(t) = I(t) \cos(2\pi f_c t) - Q(t) \sin(2\pi f_c t) \]
with a fixed frequency \( f_c \). \( Y(t) \) is zero-mean WSS if and only if \( I(t) \) and \( Q(t) \) are zero mean
\[ E(I(t)) = E(Q(t)) = 0 \]
and
\[ \rho_{II}(\tau) = \rho_{QQ}(\tau) \]
\[ \rho_{IQ}(\tau) = -\rho_{QI}(\tau). \]
The auto-correlation of \( Y(t) \) is then given
\[ \rho_{YY}(\tau) = \rho_{II}(\tau) \cos(2\pi f_c \tau) + \rho_{IQ}(\tau) \sin(2\pi f_c \tau). \]

**Proof.** The mean value of \( Y(t) \) equals zero for all \( t \) since
\[ E(Y(t)) = E(I(t) \cos(2\pi f_c t) - Q(t) \sin(2\pi f_c t)) \]
\[ = E(I(t)) \cos(2\pi f_c t) - E(Q(t)) \sin(2\pi f_c t) \]
\[ = 0. \]
The correlation function equals

$$\rho_{YY}(\tau) = E \left( (I(t + \tau) \cos (2\pi f_c(t + \tau)) - Q(t + \tau) \sin (2\pi f_c(t + \tau))) 
(I(t) \cos (2\pi f_c t) - Q(t) \sin (2\pi f_c t)) \right)$$

$$= \frac{1}{2} E \left( I(t + \tau) I(t) (\cos (2\pi f_c (2t + \tau)) + \cos (2\pi f_c \tau)) \right)$$

$$+ \frac{1}{2} E \left( (Q(t + \tau) I(t)) (\cos (2\pi f_c (2t + \tau)) + \cos (2\pi f_c \tau)) \right)$$

$$+ \frac{1}{2} E \left( (Q(t + \tau) Q(t)) (\cos (2\pi f_c (2t + \tau)) + \cos (2\pi f_c \tau)) \right)$$

$$= \frac{1}{2} (\rho_{II}(\tau) + \rho_{QQ}(\tau)) \cos (2\pi f_c \tau) + \frac{1}{2} (\rho_{IQ}(\tau) - \rho_{QI}(\tau)) \sin (2\pi f_c \tau)$$

$$+ \frac{1}{2} (\rho_{II}(\tau) - \rho_{QQ}(\tau)) \cos (2\pi f_c (2t + \tau))$$

$$+ \frac{1}{2} (-\rho_{IQ}(\tau) - \rho_{QI}(\tau)) \sin (2\pi f_c (2t + \tau)).$$

In order to make $\rho_{YY}(\tau)$ independent over time $t$, we need that $\rho_{II}(\tau) = \rho_{QQ}(\tau)$ and $\rho_{IQ}(\tau) = -\rho_{QI}(\tau)$. When satisfying these conditions, we obtain

$$\rho_{YY}(\tau) = \rho_{II}(\tau) \cos (2\pi f_c \tau) + \rho_{IQ}(\tau) \sin (2\pi f_c \tau).$$

\[\Box\]

### 7.4 Time averages or realization means

Often one does not have access to all the realizations $k$ that are produced by the stochastic space. The number of realizations is often limited, sometimes even limited to a single realization.

Assume to have access to at least one realization. The question to be answered is: when do these limited number of realizations provide sufficient information to capture the stochastic properties of the stochastic process? We therefore first define moments for a single realization $k$ (represented by $X^{(k)}(t)$) using a time average.

**Definition 39 (Time averages / realization means).**

1. Time average (realization means) is the mean value taken over time instead of averaging over the realization. The time average is represented by $\mathcal{E}(\cdot)$ to emphasize the difference with the expected value over the realization $E(\cdot)$. The time average of the realization $X^{(k)}(t)$ equals

$$\mathcal{E}(X^{(k)}(t_0 + t)) = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} X^{(k)}(t_0 + t) dt \quad (7.7)$$
CHAPTER 7. BASIC CONCEPTS IN SIGNAL THEORY

2. The quadratic mean over time for the realization \( X^{(k)}(t) \) is represented by

\[
\mathcal{E} \left( \left| X^{(k)}(t_0 + t) \right|^2 \right) = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{+T/2} \left| X^{(k)}(t_0 + t) \right|^2 dt
\]

\[
= \mathcal{E} \left( \left| X^{(k)}(t') \right|^2 \right) \quad \forall t'
\]

This mean value is independent of the origin of the time axis \( t_0 \) and represents the power \( P_X \) of the signal.

3. The correlation function (also known as auto-correlation) of the realization \( X^{(k)}(t) \) is given by

\[
R^{(k)}_{XX}(t_1, t_2) = \mathcal{E}(X^{(k)}(t_1 + t)X^{(k)*}(t_2 + t))
\]

\[
= \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{+T/2} [X^{(k)}(t_1 + t)X^{(k)*}(t_2 + t)] dt
\]

\[
= R^{(k)}_{XX}(t_1 - t_2) = R^{(k)*}_{XX}(t_2 - t_1)
\]

This value is independent of the origin of the time axis, \( t_0 \), and only depends on the chosen time interval \( \tau = t_1 - t_2 \). Furthermore, the auto-correlation function satisfies \( R^{(k)}_{XX}(\tau) = R^{(k)*}_{XX}(-\tau) \) and is an even function of \( \tau \) for real-valued signals \( X(t) \): \( R^{(k)}_{XX}(\tau) = R^{(k)}_{XX}(-\tau) \).

4. The cross-correlation function is the generalization of the auto-correlation function from a single towards two signals \( X \) and \( Y \)

\[
R^{(k)}_{XY}(t_1, t_2) = \mathcal{E}(X^{(k)}(t_1 + t)Y^{(k)*}(t_2 + t))
\]

\[
= \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{+T/2} [X^{(k)}(t_1 + t)Y^{(k)*}(t_2 + t)] dt
\]

\[
= R^{(k)}_{XY}(t_1 - t_2)
\]

\[
= R^{(k)*}_{YX}(t_2 - t_1)
\]

The cross-correlation function satisfies \( R^{(k)}_{XY}(\tau) = R^{(k)*}_{YX}(-\tau) \). For real-valued functions \( X(t) \) and \( Y(t) \) this equation becomes \( R^{(k)}_{XY}(\tau) = R^{(k)}_{YX}(-\tau) \).
Note that these time averages still have a stochastic behavior as they depend on the realization of \(X^{(k)}(t)\). Hence, it is non-trivial that the time averages and the ensemble averages correspond to each other. If they do respond, then we say that the process is ergodic.

### 7.5 Ergodicity

Section 7.3 defined the statistical moments when considering the ensemble of possible signals. These statistical moments lead to a non-stochastic result (such as the mean value). Section 7.4 considered the time averages of different orders for a single realization \(X^{(k)}(t)\). These time averages depend on the considered realization and are therefore stochastic variables. Hence, it is non-trivial that these ensemble averages are equal to their corresponding time averages.

**Definition 40 (Mean-ergodic).** A process is ergodic in the mean if the time average
\[
\mathcal{E}(X^{(k)}(t)) = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} X^{(k)}(t) dt
\]
equals (in mean squared sense for \(T \to \infty\)) the ensemble mean \(E(X(t))\) for all realizations
\[
E(X(t)) = \mathcal{E}(X^{(k)}(t)), \quad \forall k
\]

A similar definition can be given for the higher order moments.

**Definition 41 (Correlation-ergodic).** A process is ergodic in the correlation if the time correlation
\[
R_{XX}^{(k)}(\tau) = \mathcal{E}(X^{(k)}(t + \tau) X^{(k)*}(t))
\]
equals (in mean squared sense for \(T \to \infty\)) the ensemble correlation \(\rho_{XX}(\tau)\) for all realizations
\[
\rho_{XX}(\tau) = R_{XX}^{(k)}(\tau), \quad \forall k
\]

It is tempting to think that strict stationarity and ergodicity are equivalent, but this is only the case under some conditions on the statistics of \(X\) as given in the theorems below.

**Theorem 28 (Mean-ergodic).** A real-valued process is ergodic in the mean
\[
\lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} X^{(k)}(t) dt = \mu_X
\]
if and only if
\[
\lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} \rho_{XX}(\tau) d\tau = \mu_X^2
\]  
(7.8)
Proof. see [17] □

This theorem can be interpreted as follows: Consider a process that is weakly stationary (stationary up to the second order). This implies that \( \mu_X(t) = \mu_X \) and \( \rho_{XX}(t_2 + \tau, t_2) = \rho_{XX}(\tau) \). Equation (7.8) then states that the auto-correlation needs to tend to \( \mu_X^2 \) sufficiently fast. This is equivalent to asking that the variance over the ensemble \( \sigma_X^2(\tau) = \rho_{XX}(\tau) - \mu_X^2 \) tends to zero sufficiently fast.

**Theorem 29** (Correlation-ergodic). A real-valued process is ergodic in the correlation

\[
\lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} X^{(k)}(t + \tau)X^{(k)}(t)dt = \rho_{XX}(\tau)
\]

for a given \( \tau \) if and only if

\[
\lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} \left( 1 - \frac{|\lambda|}{T} \right) \left( E \left( X^{(k)}(t + \lambda + \tau)X^{(k)}(t + \lambda)X^{(k)}(t) \right) - \rho_{XX}^2(\tau) \right) d\lambda = 0
\]

Proof. see [17] □

Ergodicity of the correlation hence puts a conditions on the statistical moments up to the fourth order.

**Example 35** (Stationary, but non-ergodic process). Consider \( Y(t) = X(t) + Z \) with \( X(t) \) a weakly stationary, mean-ergodic process with ensemble mean \( \mu_X \), and \( Z \) a stochastic variable with mean value \( E(Z) = \zeta_0 \). This process is weakly stationary but non-ergodic in the mean.

Proof. The ensemble mean equals

\[
E(Y(t)) = E(X(t) + Z) = E(X(t)) + E(Z) = \mu_X + \zeta_0
\]

while the realization mean equals

\[
E(Y^{(k)}(t)) = E(X^{(k)}(t) + Z) = E(X^{(k)}(t)) + Z = \mu_X + Z
\]

since \( X \) is ergodic in the mean. The ensemble mean and the realization mean can only be equal to each other if \( Z = E(Z) = Cte \). This can only be the case if \( Z \) is not an stochastic variable. □

**Example 36** (Non-stationary ergodic process). We will now give an example of a non-stationary process \( Y(t) \) that is ergodic, namely

\[
Y(t) = X(t) \cos(2\pi f_c t)
\]

(7.9)
where $X(t)$ is a stationary ergodic process with average $E(X(t)) = 0$. Compared to Theorem 26, note the absence of a random initial phase $\phi$ in the argument of the cosine.

**Proof.** Process (7.9) is not a WSS process.

We therefore need to consider the average of $E(Y(t))$

$$E(Y(t)) = E(X(t) \cos(2\pi f_c t))$$

$$= E(X(t)) \cos(2\pi f_c t)$$

$$= 0$$

which equals zeros for all values of $t$. Hence, the process is stationary in the mean.

When considering the correlation

$$E(Y(t + \tau)Y(t)) = E(X(t + \tau) \cos(2\pi f_c (t + \tau))X(t) \cos(2\pi f_c t))$$

$$= E(X(t + \tau)X(t)) \cos(2\pi f_c (t + \tau)) \cos(2\pi f_c t)$$

$$= \frac{\rho_{XX}(\tau)}{2} [\cos(2\pi f_c \tau) + \cos(2\pi f_c (2t + \tau))]$$

we see that $E(Y(t + \tau)Y(t))$ depends on the time $t$. Hence, the process $Y(t)$ is not stationary in the second order moment, and hence not WSS. Note that the signal is cycle-stationary due to the periodicity of $\cos(2\pi f_c (2t + \tau))$ over the time variable $t$.

**Process (7.9) is a mean-ergodic process.**

To verify the ergodity in the mean value, we need to compute the time-average of a realization $Y^{(k)}(t)$

$$E(Y^{(k)}(t)) = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} X^{(k)}(t) \cos(2\pi f_c t) dt$$

$$= 0$$

We can therefore conclude that $E(Y^{(k)}(t)) = E(Y(t)) = 0$ and hence that the process is mean-ergodic, although it is not weakly stationary. □
Chapter 8

Spectral analysis of stochastic processes

An important technique in the study of deterministic signals consists in using harmonic functions (i.e. sums of sine and cosine functions) to perform a spectral analysis. The computation of the spectrum of signals is performed using the Fourier integral. The direct application of the Fourier transform on stochastic signals is generally not possible, since the $f(t)$ is not always integrable, i.e

$$
\int_{-\infty}^{\infty} |f(t)|^2 dt < \infty
$$

If this condition is not satisfied, then the Fourier integral does not exist.

This problem has already been encountered earlier when determining the Fourier spectrum of periodic signals. The issue was solved for periodic signals by using the periodicity and by introducing the Dirac delta functions that can be loosely though of as a function

$$
\delta(t) = \begin{cases} 
\infty & t = 0 \\
0 & t \neq 0 
\end{cases}
$$

with the condition that

$$
\int_{-\infty}^{\infty} \delta(t)dt = 1.
$$

Another point of view is that

$$
\int_{-\infty}^{+\infty} f(t)\delta(t)dt = f(0).
$$

The Fourier spectrum for (wide-sense stationary) stochastic signals will also be introduced using an indirect method. The method used will determine the Fourier spectrum of a stochastic signal $X(t)$ using a non-stochastic quantity that characterized this stochastic signal: its auto-correlation function $\rho_{XX}(\tau)$.  

164
8.1 Power spectra using Wiener-Kinchin theorem

The Wiener-Kinchin theorem enables the computation of the power spectral density of a wide-sense stationary stochastic signal \( X(t) \) using the auto-correlation function \( \rho_{XX}(\tau) \). This auto-correlation characterizes the stochastic process using a non-stochastic measure by considering the (second order) moment over the ensemble over all possible realizations of \( X^{(k)}(t) \).

The power spectral density (PSD) \( \Sigma_X(f) \) of a WSS stochastic process \( X(t) \) is defined as the Fourier transform of the ensemble auto-correlation \( \rho_{XX}(\tau) \)

\[
\Sigma_{XX}(f) \triangleq \mathcal{F}(\rho_{XX}(\tau))
\]

where \( \mathcal{F}(\cdot) \) represents the Fourier transform

\[
X(f) = \mathcal{F}(X(t)) = \int_{-\infty}^{\infty} X(t)e^{-j2\pi ft} dt
\]

and \( \mathcal{F}^{-1}(\cdot) \) the inverse Fourier transform

\[
X(t) = \mathcal{F}^{-1}(X(f)) = \int_{-\infty}^{\infty} X(f)e^{j2\pi ft} df.
\]

Note that the Fourier transform \( X(f) = \mathcal{F}(X(t)) \) of a real-valued process \( X(t) \) satisfies

\[
X(f) = X^*(-f).
\]

**Definition 42 (Wiener-Kinchin theorem).** The power spectrum (also known as spectral density) \( \Sigma_X(f) \) of WSS stochastic process \( X(t) \) equals the Fourier transform of its auto-correlation

\[
\Sigma_{XX}(f) \triangleq \mathcal{F}(\rho_{XX}(\tau))
\]

where

\[
\rho_{XX}(\tau) = 2 \int_{0}^{+\infty} \rho_{XX}(\tau) \cos(2\pi f \tau) d\tau
\]

with

\[
\rho_{XX}(\tau) = 2 \int_{0}^{+\infty} \Sigma_{XX}(f) \cos(2\pi f \tau) df.
\]

**Theorem 30.** The Wiener-Kinchin theorem for real-valued WSS stochastic processes \( X(t) \) equals

\[
\Sigma_{XX}(f) = 2 \int_{0}^{+\infty} \rho_{XX}(\tau) \cos(2\pi f \tau) d\tau \tag{8.1}
\]

\[
\rho_{XX}(\tau) = 2 \int_{0}^{+\infty} \Sigma_{XX}(f) \cos(2\pi f \tau) df \tag{8.2}
\]

with

\[
\rho_{XX}(\tau) = \rho_{XX}(\tau) \in \mathbb{R}
\]

\[
\Sigma_{XX}(f) = \Sigma_{XX}(-f) \in \mathbb{R}.
\]
This makes it possible to reduce the integration bounds from $(-\infty, +\infty)$ to $[0, +\infty)$. This is done by replacing the integrant with the double of its real part.

$$
\Sigma_{XX}(f) = \int_{-\infty}^{\infty} \rho_{XX}(\tau)e^{-j2\pi f \tau} d\tau \\
= \int_{-\infty}^{0} \rho_{XX}(\tau)e^{-j2\pi f \tau} d\tau + \int_{0}^{\infty} \rho_{XX}(\tau)e^{-j2\pi f \tau} d\tau \\
\overset{(a)}{=} \int_{0}^{\infty} \rho_{XX}(\tau')e^{j2\pi f \tau'} d\tau' + \int_{0}^{\infty} \rho_{XX}(\tau)e^{-j2\pi f \tau} d\tau, \quad \tau' = -\tau \\
= 2 \int_{0}^{\infty} \rho_{XX}(\tau) \cos(2\pi f \tau) d\tau
$$

where (a) uses $\rho_{XX}(\tau) = \rho_{XX}(-\tau)$ for real-valued processes by Theorem 18.

The integration after multiplying with $\cos(2\pi f \tau)$ demonstrates that the power spectra of real-valued processes are real-valued functions satisfying $\Sigma_{XX}(f) = \Sigma_{XX}(-f)$.

Furthermore, it can be shown that the integrated power over frequency equals the auto-correlation in $\rho_{XX}(0)$

$$
\int_{-\infty}^{\infty} \Sigma_{XX}(f) df = \left. \int_{-\infty}^{\infty} \Sigma_{XX}(f)e^{-j2\pi f \tau} df \right|_{\tau=0} = \rho_{XX}(\tau)|_{\tau=0} = E(|X(t)|^2) \geq 0
$$

implying that the total area of $\Sigma_{XX}(f)$ equals the average power of the process $X(t)$.

**Definition 43.** The cross-power spectrum $\Sigma_{XY}(f)$ of two WSS processes $X(t)$ and $Y(t)$ is the Fourier transform of their cross-correlation $\rho_{XY}(\tau)$

$$
\Sigma_{XY}(f) \triangleq \mathcal{F}(\rho_{XY}(\tau)) \\
\rho_{XY}(\tau) \triangleq \mathcal{F}^{-1}(\Sigma_{XY}(f)).
$$

with

$$
\rho_{XY}(\tau) = \rho_{YX}(-\tau) \\
\Sigma_{XY}(f) = \Sigma_{YX}(-f).
$$
The cross power spectra $\Sigma_{XY}(f)$ and $\Sigma_{YX}(f)$ are related by

$$\Sigma_{XY}(f) = \int_{-\infty}^{\infty} \rho_{XY}(\tau)e^{-j2\pi f \tau} d\tau$$

(a) uses $\rho_{XY}(\tau) = \rho_{YX}^*(\tau)$, (b) considers the complex conjugate of the integral, and (c) makes a change in variables $\tau' = -\tau$.

Lemma 20. The cross-power spectra for real-valued processes $X(t)$ and $Y(t)$ satisfy

$$\Sigma_{XY}(f) = \Sigma_{XY}^*(-f).$$

Proof. Starting from the cross-power spectrum, we find

$$\Sigma_{XY}(f) = \int_{-\infty}^{\infty} \rho_{XY}(\tau)e^{-j2\pi f \tau} d\tau$$

(a) uses that the cross-correlation of real-valued processes is also real-valued.

8.2 Properties of power spectra

Theorem 31. If the WSS processes $X(t)$ and $Y(t)$ are orthogonal ($\rho_{XY}(\tau) = 0 \forall \tau$), then

$$\Sigma_{XY}(f) = 0$$

and their sum $Z(t) = X(t) + Y(t)$ satisfies

$$\rho_{ZZ}(\tau) = \rho_{XX}(\tau) + \rho_{YY}(\tau)$$

and

$$\Sigma_{ZZ}(f) = \Sigma_{XX}(f) + \Sigma_{YY}(f).$$
Proof. This theorem is simple to prove. The cross-correlation equals zero and hence its Fourier transform as well. Furthermore, Theorem 22 showed that the auto-correlation of the sum \( Z(t) = X(t) + Y(t) \) satisfies

\[
\rho_{ZZ}(\tau) = \rho_{XX}(\tau) + \rho_{YY}(\tau).
\]

As the spectrum equals the Fourier transform of the auto-correlation, and since the Fourier transform is a linear operator, it immediately follows that

\[
\Sigma_{ZZ}(f) = \Sigma_{XX}(f) + \Sigma_{YY}(f).
\]

A list of properties between auto-correlation and power spectra is shown in Table 8.1. These properties can be deduced directly from the properties the Fourier transform.
### Table 8.1: Different auto-correlation functions and their corresponding power spectra.

<table>
<thead>
<tr>
<th>$X(t)$</th>
<th>$\rho_{XX}(\tau)$</th>
<th>$\Sigma_{XX}(f)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$aX(t)$</td>
<td>$</td>
<td>a</td>
</tr>
<tr>
<td>$X(t)e^{\pm j2\pi f_c t}$</td>
<td>$\rho_{XX}(\tau)e^{\pm j2\pi f_c \tau}$</td>
<td>$\Sigma_{XX}(f \mp f_c)$</td>
</tr>
<tr>
<td>$X(t)\cos(2\pi f_c t)$</td>
<td>$\rho_{XX}(\tau)\cos(2\pi f_c \tau)$</td>
<td>$\frac{1}{2}[\Sigma_{XX}(f - f_c) + \Sigma_{XX}(f + f_c)]$</td>
</tr>
<tr>
<td>$X(t)\sin(2\pi f_c t)$</td>
<td>$\rho_{XX}(\tau)\sin(2\pi f_c \tau)$</td>
<td>$\frac{1}{2j}[\Sigma_{XX}(f - f_c) - \Sigma_{XX}(f + f_c)]$</td>
</tr>
</tbody>
</table>

| $X(t) \in \mathbb{R}$ | $\rho_{XX}(\tau) = \rho_{XX}(-\tau) \in \mathbb{R}$ | $\Sigma_{XX}(f) = \Sigma_{XX}(-f) \in \mathbb{R}$ |
| $\rho_{XY}(\tau) = \rho_{YX}(-\tau)$ | $\Sigma_{XY}(f) = \Sigma_{YX}(-f)$ |

| $X(t), Y(t) \in \mathbb{R}$ | $\rho_{XY}(\tau) = \rho_{YX}(-\tau) \in \mathbb{R}$ | $\Sigma_{XY}(f) = \Sigma_{YX}(-f)$ |
| $\text{rect}(a\tau)$ | $\frac{1}{|a|}\text{sinc}(f/a)$ |
| $\text{sinc}(a\tau)$ | $\frac{1}{|a|}\text{rect}(f/a)$ |
| $\text{tri}(a\tau)$ | $\frac{1}{|a|}\text{sinc}^2(f/a)$ |
| $\text{sinc}^2(a\tau)$ | $\frac{1}{|a|}\text{tri}(f/a)$ |
| $e^{-a|\tau|}$ ($a > 0$) | $\frac{2a}{a^2 + (2\pi)^2}$ |
| $e^{-a\tau^2}$ ($a > 0$) | $\sqrt{\frac{a}{\pi}}e^{-\left(\frac{\pi\tau^2}{a}\right)}$ |

Table 8.1 uses the following notations:

- the Dirac-comb
  \[ \delta(\tau) = \sum_{k=-\infty}^{\infty} \delta(t - kT) \]

- the convolutional product
  \[ (f * g)(t) \triangleq \int_{-\infty}^{\infty} f(u)g(t-u)du = (g * f)(t) \]
CHAPTER 8. SPECTRAL ANALYSIS OF STOCHASTIC PROCESSES

- the normalized sinc function
  \[ \text{sinc}(x) = \frac{\sin(\pi x)}{\pi x} \]
- the rectangular function
  \[ \text{rect}(x) = \begin{cases} 
  1 & |x| < \frac{1}{2} \\
  0 & |x| \geq \frac{1}{2} 
\end{cases} \]
- the triangular function
  \[ \text{tri}(x) = \begin{cases} 
  1 - |x| & |x| < 1 \\
  0 & |x| \geq 1 
\end{cases} \]

The tri(x) function can be seen as a convolution product
\[
\text{tri}(\tau) = \text{rect}(\tau) \ast \text{rect}(\tau) = \int_{-\infty}^{\infty} \text{rect}(\tau - u)\text{rect}(u)du.
\]

Hence,
\[
\text{tri}(a\tau) = \int_{-\infty}^{\infty} \text{rect}(a\tau - u)\text{rect}(u)du
\]
\[
= |a| \int_{-\infty}^{\infty} \text{rect}(\tau - u')\text{rect}(u')du', \text{ with } u' = u/|a|
\]
\[
= |a|\text{rect}(\tau) \ast \text{rect}(\tau)
\]

and therefore
\[
\mathcal{F}(\text{tri}(a\tau)) = |a| (\mathcal{F}(\text{rect}(\tau)))^2
\]
\[
= \frac{1}{|a|} \text{sinc}^2(f/a).
\]

8.3 Power spectrum after linear time-invariant (LTI) filtering

The time-domain response of a LTI system with impulse response \( h(t) \) when excited by an input \( X(t) \) equals
\[
Y(t) = \int_{-\infty}^{\infty} h(u)X(t-u)du = \int_{-\infty}^{\infty} h(t-u)X(u)du.
\]

This convolution can be transformed in a multiplication in the frequency domain using the Fourier transform
\[
\mathcal{F}(Y(t)) = \mathcal{F}(h(t)) \mathcal{F}(X(t))
\]
\[
= H(f) \mathcal{F}(X(t))
\]
where \( H(f) = \mathcal{F}(h(t)) \) represent the system’s frequency response.
CHAPTER 8. SPECTRAL ANALYSIS OF STOCHASTIC PROCESSES 171

Theorem 32. If $X(t)$ is a WSS signal passing through a LTI system with impulse response $h(t)$, then

$$
egin{align*}
\mu_Y &= H(0) \mu_X \\
\Sigma_{YX}(f) &= H(f) \Sigma_{XX}(f) \\
\Sigma_{XY}(f) &= H^*(f) \Sigma_{XX}(f) \\
\Sigma_{YY}(f) &= |H(f)|^2 \Sigma_{XX}(f).
\end{align*}
$$

Proof. All the equations can easily be derived by combining the properties of the Fourier transform, the results on the auto- and cross-correlation in Theorem 24, and the Wiener-Kinchin theorem to determine the (cross) power spectra (Definition 42).

The mean value can be expressed as

$$
\begin{align*}
\mu_Y &= E \left( \int_{-\infty}^{\infty} h(u)X(t-u)du \right) \\
&= \int_{-\infty}^{\infty} h(u) E(X(t-u)) du \\
&\overset{(a)}{=} \int_{-\infty}^{\infty} h(u) \mu_X du \\
&= \mu_X \int_{-\infty}^{\infty} h(u) du \\
&= \mu_X H(0)
\end{align*}
$$

where (a) uses the assumption that $X(t)$ is WSS and hence has a constant mean value.

The cross-power density between $X(t)$ and $Y(t)$

$$
\begin{align*}
\Sigma_{YX}(f) &= \mathcal{F}(\rho_{YX}(\tau)) \\
&= \mathcal{F}(h(\tau) * \rho_{XX}(\tau)) \\
&= \mathcal{F}(h(\tau)) \mathcal{F}(\rho_{XX}(\tau)) \\
&= H(f) \Sigma_{XX}(f)
\end{align*}
$$

and

$$
\begin{align*}
\Sigma_{XY}(f) &= \mathcal{F}(\rho_{XY}(\tau)) \\
&= \mathcal{F}(\rho_{XX}(\tau) * h^*(-\tau)) \\
&= \mathcal{F}(\rho_{XX}(\tau)) \mathcal{F}(h^*(-\tau)) \\
&= H^*(f) \Sigma_{XX}(f)
\end{align*}
$$
since
\[
\mathcal{F}(h^*(-\tau)) = \int_{-\infty}^{\infty} h^*(-\tau)e^{-j2\pi f\tau} d\tau
\]
\[
= \left( \int_{-\infty}^{\infty} h(-\tau)e^{j2\pi f\tau} d\tau \right)^* \]
\[
= \left( \int_{-\infty}^{\infty} h(\tau')e^{j2\pi f\tau'} d\tau' \right)^*, \quad \tau' = -\tau
\]
\[
= H^*(f).
\]

The power spectral density of \(Y(t)\) equals
\[
\Sigma_{YY}(f) = \mathcal{F}(\rho_{YY}(\tau))
\]
\[
= \mathcal{F}(h(\tau) * h^*(-\tau)) * \rho_{XX}(\tau)
\]
\[
= \mathcal{F}(h(\tau))\mathcal{F}(h^*(-\tau))\mathcal{F}(\rho_{XX}(\tau))
\]
\[
= H(f)H^*(f)\Sigma_{XX}(f)
\]
\[
= |H(f)|^2\Sigma_{XX}(f).
\]

8.4 Power spectrum of periodic auto-correlation

A periodic signal \(f(t)\) satisfying
\[
f(t) = f(t + kT)
\]
for integer \(k\) can be written as the convolution of the windowed function
\[
g(t) = \begin{cases} 
  f(t), & -T/2 < t < T/2, \\
  f(t)/2 & |t| = T/2 \\
  0, & \text{otherwise.}
\end{cases}
\]

with the Dirac-comb since
\[
f(t) = \mathcal{F}(\tau(t) * g(t))
\]
\[
= \sum_{k=-\infty}^{\infty} g(t - kT).
\]

Theorem 33. A WSS process with a periodic auto-correlation \(\rho_{XX}(\tau) = \rho_{XX}(\tau + kT)\) for integer \(k\) has a power spectrum with discrete frequency lines and can be computed using
\[
\Sigma_{XX}(f) = \frac{1}{T} \sum_{k=-\infty}^{\infty} \Sigma_{YY}(f - \frac{k}{T})\delta(f - \frac{k}{T})
\]
with $\Sigma_{YY}(f) = \mathcal{F}(\rho_{YY}(\tau))$ and

$$\rho_{YY}(\tau) = \begin{cases} 
\rho_{XX}(\tau), & -T/2 < \tau < T/2, \\
\rho_{XX}(\tau)/2 & |\tau| = T/2 \\
0, & \text{otherwise.}
\end{cases}$$

Proof. The periodic auto-correlation $\rho_{XX}(\tau)$ can be written as the convolution of a Dirac-comb with the auto-correlation $\rho_{YY}(\tau)$ which is non-zero in $[-T/2, T/2]$

$$\rho_{XX}(\tau) = \sum_{k=-\infty}^{\infty} \rho_{YY}(\tau - kT) = \Lambda_{\frac{T}{2}}(\tau) * \rho_{YY}(\tau).$$

The Fourier transform of a convolution and a Dirac-comb (Table 8.1) then imply that

$$\Sigma_{XX}(f) = \mathcal{F}(\rho_{XX}(\tau)) = \mathcal{F}(\Lambda_{\frac{T}{2}}(\tau) * \rho_{YY}(\tau)) = \mathcal{F}(\Lambda_{\frac{T}{2}}(\tau)) \mathcal{F}(\rho_{YY}(\tau)) = \int_{-\infty}^{\infty} \frac{1}{T} \Lambda_{\frac{T}{2}}(f) \Sigma_{YY}(f) = \int_{-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \Sigma_{YY}(f - \frac{k}{T}) \delta(f - \frac{k}{T}).$$

8.5 Periodogram

The Wiener-Kinchin theorem starts from the auto-correlation over the ensemble of the signals to determine the power spectra. Similarly to the previous chapter, it is possible to define content of the signal power using the time average for a realization $X^{(k)}(t)$.

As the time averaging is done over a time window $T$, it is (from notation point of view) interesting to introduce the signal’s properties of a windowed version of $X^{(k)}(t)$. This is done by multiplying the the original signal $X^{(k)}(t)$ with the window function

$$w_T(t) = \begin{cases} 
1 & |t| \leq \frac{T}{2} \\
0 & |t| > \frac{T}{2}
\end{cases}$$

to obtain

$$X_T^{(k)}(t) = X^{(k)}(t)w_T(t).$$
The correlation over time of the windowed signal $X_T^{(k)}(t)$ is a stochastic variable which equals

$$R_{XX,T}^{(k)}(\tau) = \frac{1}{T} \int_{-T/2}^{T/2} X_T^{(k)}(t)X_T^{(k)*}(t + \tau) \, dt.$$  

We will now link this correlation with the spectral representation of the signal.

### 8.5.1 Power representation in the frequency domain

The Fourier integral remains finite for the windowed signal $X_T^{(k)}(t)$. Hence, it is possible to compute its spectrum. Note that the introduction of the rectangular window will, however, introduce spectral leakage as already studied in other courses. The Fourier transform of $X_T^{(k)}(t)$ equals

$$X_T^{(k)}(f) = \mathcal{F}(X_T^{(k)}(t))$$

$$= \int_{-\infty}^{\infty} X_T^{(k)}(t)e^{-j2\pi ft} \, dt$$

$$= \int_{-T/2}^{T/2} X_T^{(k)}(t)e^{-j2\pi ft} \, dt$$

where $X_T^{(k)}(f)$ is still a stochastic variable. This Fourier transform can be used to define the periodogram

**Definition 44** (Periodogram). The periodogram $S_{XX,T}^{(k)}(f)$ is the spectral density of the signal realization $X^{(k)}(t)$ over the time window $T$

$$S_{XX,T}^{(k)}(f) = \frac{1}{T}|X_T^{(k)}(f)|^2$$

with $X_T^{(k)}(f) = \mathcal{F}(X_T^{(k)}(t))$.

The periodogram is a stochastic variable that depends on the realization $X^{(k)}(t)$ and contains no phase information due to the norm $|\cdot|^2$.

The following relation exists between the periodogram and the auto-correlation over time.

**Theorem 34.** The periodogram equals the Fourier transform of the auto-correlation

$$S_{XX,T}^{(k)}(f) = \mathcal{F}(R_{XX,T}^{(k)}(\tau)).$$
CHAPTER 8. SPECTRAL ANALYSIS OF STOCHASTIC PROCESSES

Proof. Starting from the right hand side of the equality
\[ \mathcal{F}\left(R_{XX,T}^{(k)}(\tau)\right) = \int_{-\infty}^{\infty} R_{XX,T}^{(k)}(\tau)e^{-j2\pi f \tau} d\tau \]
\[ = \int_{-\tau/2}^{\tau/2} R_{XX,T}^{(k)}(\tau)e^{-j2\pi f \tau} d\tau \]
\[ = \frac{1}{T} \int_{-\tau/2}^{\tau/2} \left( \int_{-\tau/2}^{\tau/2} [X_T^{(k)}(t+\tau)X_T^{(k)*}(t)] dt \right) e^{-j2\pi f (t+\tau)} e^{j2\pi ft} dt d\tau \]
\[ = \frac{1}{T} \int_{-\tau/2}^{\tau/2} \int_{-\tau/2}^{\tau/2} \left[X_T^{(k)}(u)X_T^{(k)*}(t)\right] e^{-j2\pi f u} e^{j2\pi ft} du d\tau \]
\[ = \frac{1}{T} \int_{-\tau/2}^{\tau/2} X_T^{(k)}(u) e^{-j2\pi f u} du \int_{-\tau/2}^{\tau/2} X_T^{(k)*}(t) e^{j2\pi ft} dt \]
\[ = \frac{1}{T} \int \mathcal{F}(X_T^{(k)}(t)) \left(\mathcal{F}(X_T^{(k)}(t))\right)^* \]
\[ = \frac{1}{T} |X_T^{(k)}(f)|^2 \]
\[ = S_X^{(k)}(f). \]

This periodogram will converge asymptotically, for \( T \to \infty \), towards estimate of the power spectrum of the realization \( X^{(k)}(t) \). The question that still remains is: under which conditions will the periodogram of a single realization \( X^{(k)}(t) \) converge to the power spectrum \( \Sigma_{XX}(f) \) (= a form of ergodicity)??

8.5.2 Ergodicity in the spectrum

Theorem 35 (Ergodicity in the power spectral density). If the WSS process \( X(t) \) satisfies
\[ \lim_{T \to \infty} \int_{-\tau/2}^{\tau/2} |\tau \rho_{XX}(\tau)| d\tau < \infty \quad (8.3) \]
then the expected value of the periodogram \( S_{XX,T}^{(k)}(f) \) converges asymptotically (in mean square sense) for \( T \to \infty \) towards the power spectral density of the signal
\[ \lim_{T \to \infty} \mathbb{E} \left( S_{XX,T}^{(k)}(f) \right) = \Sigma_{XX}(f). \]

Proof. See [17].

The condition (8.3) basically demands that the auto-correlation of the signal decreases faster than linearly towards zero.
CHAPTER 8. SPECTRAL ANALYSIS OF STOCHASTIC PROCESSES 176

8.6 Applications and examples

This section illustrates the item discussed in the previous section using various examples.

8.6.1 Stationary white noise

A stationary white noise process generates a random signal with a constant power spectral density. Being uncorrelated in time does not restrict the considered distribution of the signal, as long as they are zero mean. This includes binary noise, Gaussian, Poisson, Cauchy, ... noise.

Definition 45. White noise has a flat power spectrum

$$\Sigma_{XX}(f) = \sigma_X^2 \forall f$$

or equivalently a Dirac function as auto-correlation

$$\rho_{XX}(\tau) = \sigma_X^2 \delta(\tau).$$

This can easily be seen by considering

$$\Sigma_{XX}(f) = \int_{-\infty}^{\infty} \sigma_X^2 \delta(\tau)e^{-j2\pi f \tau} d\tau = \sigma_X^2.$$

Also note that the mean value must be equal to zero since

$$\mu_X^2 = \lim_{\tau \to \infty} \rho_{XX}(\tau) = 0.$$

The term white noise originates from white light, since white light contains all spectral frequencies. Note that light that appears to be white generally does not have a flat power spectral density over the visible band.

An infinite-bandwidth white noise signal is a purely theoretical concept as the overall power of the signal would tend to infinity. Practically, the bandwidth is always limited by the mechanism of noise generation, by the transmission medium, or by the finite observation capabilities. Hence, a "white noise" random signal is considered when the observed spectrum is flat in the frequency band of interest.

8.6.2 Stationary Gaussian noise

Assume a wide-sense stationary Gaussian process. Hence, the mean value $\mu_X$ and variance $\sigma_X^2$ are constant over time. The first order probability density function (pdf) of a stationary Gaussian signal satisfies

$$f_X(x_1; t_1) = f_X(x_1) = \frac{1}{\sqrt{2\pi \sigma_X^2}} \exp\left(-\frac{(x_1 - \mu_X)^2}{2\sigma_X^2}\right)$$

Note that this first order pdf is independent of the time due to the stationarity.
The second order pdf of a stationary Gaussian signal only depends on the time difference $\tau = t_1 - t_2$ using the **Pearson’s correlation coefficient** which depends on $\tau$

$$\rho(\tau) = \frac{E((X(t+\tau) - \mu_X) (X(t) - \mu_X))}{\sqrt{E((X(t + \tau) - \mu_X)^2) E((X(t) - \mu_X)^2)}} \quad \forall t$$

**Important note:** the symbol $\rho(\tau)$ (without subscript $XX$) represents the Pearson’s correlation coefficient of two random variables $X(t_1)$ and $X(t_2)$. This coefficient measures the linear correlation between two variables and is bounded $-1 \leq \rho(\tau) \leq 1$ where 1 is total positive linear correlation, 0 is no linear correlation, and $-1$ is total negative linear correlation.

The probability density function of the $n$-dimensional zero-mean Gaussian distribution is given by

$$f(x_1, \ldots, x_n) = \frac{1}{\sqrt{(2\pi)^n \det C}} e^{\left(-\frac{1}{2}x^T C^{-1} x\right)}.$$ 

The second order can then be computed using

$$C(\tau) = \begin{bmatrix} \sigma_X^2 & \rho(\tau) \sigma_X^2 \\ \rho(\tau) \sigma_X^2 & \sigma_X^2 \end{bmatrix}$$

and

$$x = \begin{bmatrix} x_1 - \mu_X \\ x_2 - \mu_X \end{bmatrix}.$$ 

Hence,

$$f_{XX}(x_1, x_2; \tau)$$

$$= \frac{1}{2\pi\sigma_X^2 \sqrt{1 - \rho^2(\tau)}} \exp\left(-\frac{1}{2\sigma_X^2} \begin{bmatrix} x_1 - \mu_X \\ x_2 - \mu_X \end{bmatrix}^T \begin{bmatrix} 1 & \rho(\tau) \\ \rho(\tau) & 1 \end{bmatrix}^{-1} \begin{bmatrix} x_1 - \mu_X \\ x_2 - \mu_X \end{bmatrix}\right)$$

$$= \frac{1}{2\pi\sigma_X^2 \sqrt{1 - \rho^2(\tau)}} \exp\left(-\frac{(x_1 - \mu_X)^2 - 2\rho(\tau)(x_1 - \mu_X)(x_2 - \mu_X) + (x_2 - \mu_X)^2}{2\sigma_X^2 (1 - \rho^2(\tau))}\right).$$

**Theorem 36.** Weakly stationary Gaussian processes implies strong stationarity.

**Proof.** It is known that the probability density function of a $n$-dimensional Gaussian distribution

$$f_X \ldots X^{(x_1, \ldots, x_n; t_1, \ldots, t_n)}$$
is completely characterized by its first and second order moments, namely the mean values \( \mu_X(t_i) \) and the covariance matrix given by \( \sigma^2_X(t_i, t_j) \). If the Gaussian process is weakly stationary, then the mean \( \mu_X \) is constant and the covariance matrix only depends on the time difference \( C_X(t, t + \tau) = C_X(\tau) \).

Hence,

\[
f_X \ldots X(x_1, \ldots, x_n; t_1, \ldots, t_n) = f_X \ldots X(x_1, \ldots, x_n; \tau)
\]

for all \( n \), implying strong stationarity.

**Theorem 37.** Uncorrelated Gaussian processes are independent.

*Proof.* If there is no correlation over time, then Pearson’s correlation coefficient equals zero at all time

\[
\rho(\tau) = 0 \quad \forall \tau
\]

and the second order pdf reduces to

\[
f_{XX}(x_1, x_2; \tau) = \frac{1}{2\pi \sigma^2_X} \exp \left( -\frac{(x_1 - \mu_X)^2 + (x_2 - \mu_X)^2}{2\sigma^2_X} \right)
\]

\[
= \frac{1}{\sqrt{2\pi \sigma^2_X}} \exp \left( -\frac{(x_1 - \mu_X)^2}{2\sigma^2_X} \right) \frac{1}{\sqrt{2\pi \sigma^2_X}} \exp \left( -\frac{(x_2 - \mu_X)^2}{2\sigma^2_X} \right)
\]

\[
= f_X(x_1)f_X(x_2)
\]

This can be generalized to a diagonal co-variance matrix \( C_X(\tau) \) (as \( \rho(\tau) = 0 \forall \tau \)) with time-independent variances \( (\sigma^2_X) \) on the diagonal. The co-variance matrix \( C_X(\tau) \) is therefore constant for all \( \tau \). Hence, the higher order pdfs can also be written as a product of the (time independent) first order pdfs \( f_X(x_i) \)

\[
\underline{f_X \ldots X}(x_1, \ldots, x_n; t_1, \ldots, t_n) = \prod_{i=1}^{n} f_X(x_i)
\]

for all \( n \), implying that uncorrelated Gaussian random variables are also independent.

Note that uncorrelated variables are in general not independent!

White Gaussian noise is often used in the context of signal theory. It combines the properties of a white noise process (Section 8.6.1) and the WSS Gaussian noise process.

**Definition 46 (White Gaussian noise).** A white Gaussian noise process follows a zero-mean Gaussian distribution

\[
f_X(x_1) = \frac{1}{\sqrt{2\pi \sigma^2_X}} \exp \left( -\frac{x_1^2}{2\sigma^2_X} \right)
\]
with variance $\sigma_X^2$. The process is uncorrelated over the time

$$\rho_{XX}(\tau) = \sigma_X^2 \delta(\tau)$$

and has a flat power spectrum

$$\Sigma_{XX}(f) = \sigma_X^2 \quad \forall f.$$ 

If can easily be seen using the above theorems that white Gaussian noise is strongly stationary and independence over time. Furthermore, white Gaussian noise is ergodic in the mean value (Theorem 28) since

$$\lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} \sigma_X^2 \delta(\tau)d\tau = \lim_{T \to \infty} \frac{1}{T} \sigma_X^2 = 0 = \mu_x^2$$

and ergodic in the power spectral density (Theorem 35) as

$$\lim_{T \to \infty} \int_{-T/2}^{T/2} |\tau\rho_{XX}(\tau)|d\tau = \lim_{T \to \infty} \int_{-T/2}^{T/2} |\tau|\sigma_X^2 \delta(\tau)d\tau = \sigma_X^2 |\tau|_{\tau=0} = 0 < \infty.$$ 

### 8.6.3 Filtered white noise

Consider a white noise source that is wide-sense stationary and that is filtered by an ideal filter.

**Definition 47** (Ideal low-pass filtered white noise). A wide-sense stationary process $X(t)$ is an ideal low-pass filtered white noise with a bandwidth $W$ if its power spectrum satisfies

$$\Sigma_{XX}(f) = \begin{cases} 
\frac{N_0}{2} = Cte & |f| < W \\
0 & |f| \geq W 
\end{cases}$$

$$= \frac{N_0}{2} \text{rect}(f/(2W)).$$

The auto-correlation equals

$$\rho_{XX}(\tau) = \mathcal{F}^{-1}(\Sigma_{XX}(f))$$

$$= \frac{N_0}{2} \mathcal{F}^{-1}(\text{rect}(f/(2W)))$$

$$= \frac{N_0}{2} 2W \text{sinc}(2W\tau)$$

$$= N_0 \frac{\sin(2\pi W\tau)}{2\pi \tau}.$$
An example of the auto-correlation of an ideal low-pass filter white noise process is shown in Fig. 8.1.

\[
E(X(t + nT)X(t)) = \rho_{XX}(nT) = N_0 \frac{\sin(2\pi W n T)}{2\pi n T}\]

If, additionally, the process \(X(t)\) is Gaussian distributed, then the combination of uncorrelated with the Gaussian distributions implies that the samples of the Gaussian noise at \(X(nT)\) are independent (Theorem 37). This property was used to prove the Shannon-Hartley theorem (Theorem 11).

8.6.4 Binary Synchronous Non-Return-to-Zero (NRZ) Signal

The binary synchronous Non-Return-to-Zero signal is the easiest and simplest line coding to transmit binary coded information in a bit serial manner. The
signal takes two distinct values, $a_0$ and $a_1$, depending on the data bit, 0 or 1, that is to be transmitted. The transitions from $a_0$ to $a_1$ (and back) are considered to be instantaneous. The data bit are transmitted at a fixed symbol rate $1/T_s$ where $T_s$ represent the period of a single bit. Furthermore, consider that all data bits are stochastic independent and that ones and zeros are equally probable $P[x = 0] = P[x = 1] = 1/2$, i.e. the optimal choice from information point of view.

The ensemble average of the NRZ signal can easily be computed since the signal has only two states

$$
\mu_X = E(X(t)) = a_0 P[x = 0] + a_1 P[x = 1] = \frac{1}{2}(a_0 + a_1).
$$

Observe that the mean value of the signal equals zero if $a_1 = -a_0$. This type of NRZ signal is also known as bipolar NRZ. An unipolar NRZ signal, on the other hand, uses $a_0 = 0$ and hence has a mean value of $a_1/2$. An example of a unipolar NRZ signal can be found in Fig. 8.2.

![Figure 8.2](image.png)

Figure 8.2: Example of unipolar Non-Return-to-Zero (NRZ) signal with $P[x = 0] = P[x = 1] = 1/2$, and $a_0 = 0$ and $a_1 = 2$.

The quadratic mean of the NRZ equals

$$
E(X^2(t)) = a_0^2 P[x = 0] + a_1^2 P[x = 1] = \frac{1}{2}(a_0^2 + a_1^2)
$$
and hence
\[
\sigma_X^2 = E(X^2(t)) - \mu_X^2
= \frac{1}{2} (a_0^2 + a_1^2) - \frac{1}{4} (a_0 + a_1)^2
= \frac{1}{4} (a_0 - a_1)^2.
\]

The auto-correlation is computed using
\[
\rho_{XX}(\tau) = E(X(t+\tau)X(t)).
\]

This expected value can be decomposed in two ranges, \(|\tau| > T_s\) and \(|\tau| \leq T_s\).

**Auto-correlation for \(|\tau| > T_s\)**

Consider that the time shift \(\tau\) is larger than the period of a single bit. Since all bits are independent and since \(|\tau| > T_s\), it is known that the stochastic variables \(X(t+\tau)\) and \(X(t)\) are independent from each other for all possible values of the time \(t\). Hence,
\[
E(X(t+\tau)X(t)) \overset{(a)}{=} E(X(t+\tau))E(X(t)) \text{ for } |\tau| > T_s
= \mu_X^2.
\]

where (a) is due to the independence of \(X(t+\tau)\) and \(X(t)\)

**Auto-correlation for \(|\tau| \leq T_s\)**

The study of the auto-correlation is a bit more complex as \(X(t+\tau)\) and \(X(t)\) can be either identical to each other (if \(X(t+\tau)\) and \(X(t)\) are representing the same data bit), or independent from each other (if \(X(t+\tau)\) is the value of the data bit that follow / proceeds \(X(t)\)). We therefore first study the auto-correlations for both cases at given time instance \(t_0\): \(\rho_{XX}(\tau|t_0)\). Afterwards, the auto-correlation of the NRZ signal \(\rho_{XX}(\tau)\) will be computed using the distribution of the time instance \(t_0\).

**Step 1: The auto-correlation for \(|\tau| \leq T_s\) at a time instance \(t_0\)**

The auto-correlation at a time instance \(t_0\), \(\rho_{XX}(\tau|t_0)\) can be expressed as \(E(X(t_0+\tau)X(t_0))\) and can take two values

1. If \(X(t_0+\tau)\) and \(X(t_0)\) represent the same data bit, then they their realizations are always equal \((X(t_0+\tau) = X(t_0)\) since they contain the same information) and hence
\[
\rho_{XX}(\tau|\text{same info at } t_0 \text{ and } t_0 + \tau) = E(X(t_0+\tau)X(t_0))
= E(X^2(t_0))
= \sigma_X^2 + \mu_X^2.
\]
2. If \(X(t_0 + \tau)\) and \(X(t_0)\) represent different (=independent) data bits, then their realizations are stochastic independent and hence
\[
\rho_{XX}(\tau|\text{indep. info at } t_0 \text{ and } t_0 + \tau) = E(X(t_0 + \tau)X(t_0)) \\
= E(X(t_0 + \tau))E(X(t_0)) \\
= \mu_X^2.
\]

**Step 2: The auto-correlation for \(|\tau| \leq T_s\) considering the distribution of \(t_0\)**

To consider the distribution of \(t_0\) on the auto-correlation, we first need to determine its distribution. As all the symbols have a period \(T_s\), one can assume that a random time instance \(t_0\) is uniformly distributed within the period \([0, T_s]\)
\[
f_{T_0}(t_0) = \begin{cases} 
\frac{1}{T} & t_0 \in [0, T_s] \\
0 & t_0 \notin [0, T_s].
\end{cases}
\]

To determine the auto-correlation \(\rho_{XX}(\tau)\), one needs to consider that
\[
\rho_{XX}(\tau) = \rho_{XX}(\tau|\text{same info at } t_0 \text{ and } t_0 + \tau) \\
\times P[\text{same info at } t_0 \text{ and } t_0 + \tau] \\
+ \rho_{XX}(\tau|\text{indep. info at } t_0 \text{ and } t_0 + \tau) \\
\times P[\text{indep. info at } t_0 \text{ and } t_0 + \tau]
\]

The probability that the signal at \(t_0\) and \(t_0 + \tau\) represent the same information, \(P[\text{same info at } t_0 \text{ and } t_0 + \tau]\), depends on when \(t_0\) actually starts and when the signal’s transition takes place. This probability equals
\[
P[\text{same info at } t_0 \text{ and } t_0 + \tau] = 1 - \frac{|\tau|}{T_s}
\]
while \(P[\text{indep. info at } t_0 \text{ and } t_0 + \tau]\) equals is complement
\[
P[\text{indep. info at } t_0 \text{ and } t_0 + \tau] = \frac{|\tau|}{T_s}.
\]

This results into
\[
\rho_{XX}(\tau) = (\sigma_X^2 + \mu_X^2) \left(1 - \frac{|\tau|}{T_s}\right) + \mu_{XX}^2 \frac{|\tau|}{T_s} \\
= \sigma_X^2 \left(1 - \frac{|\tau|}{T_s}\right) + \mu_{XX}^2 \\
= \sigma_X^2 \text{tri}(\frac{\tau}{T_s}) + \mu_X^2.
\]

Fig. 8.3 shows the auto-correlation for the sequence that corresponds to the NRZ as shown in Fig. 8.2.
CHAPTER 8. SPECTRAL ANALYSIS OF STOCHASTIC PROCESSES

Figure 8.3: The auto-correlation of the non-return to zero signal of Fig. 8.2 with $a = 0$ and $a_1 = 2$. Note the triangle for $|\tau| \leq T_s$ and the constant value for $\tau > T_s$.

**Power spectrum of a NRZ signal**

The power spectrum of a NRZ signal can now be computed using the Wiener-Kinchin theorem

\[
\Sigma_{XX}^{NRZ}(f) = \mathcal{F}(\rho_{XX}(\tau)) = \mathcal{F} \left( \sigma_X^2 \text{tri} \left( \frac{T}{T_s} \right) + \mu_X^2 \right) = T_s \sigma_X^2 \text{sinc}^2(fT_s) + \mu_X^2 \delta(f) = T_s \sigma_X^2 \frac{(\sin(\pi f T_s))^2}{(\pi f T_s)^2} + \mu_X^2 \delta(f).
\]

Fig. 8.4 shows the power spectral density of the NRZ on a linear frequency axis. Note that the power spectral density equals zeros for all frequencies $f = k/T_s$ for all non-zero integer $k$ and that the DC value at $f = 0$ contains a Dirac function. This Dirac contribution equals $\mu_X^2$ and vanishes if the average value of the NRZ signal equals zero, i.e. in the case of a bipolar binary signal.

Fig. 8.5 shows the same spectrum on the logarithmic axis (excluding the DC value). It is possible to proof that the PSD contains $91\%$ of its power in the frequency range $[-1/T_s, 1/T_s]$. As a rule of the thumb one can state that the bandwidth of a NRZ line encoder yields the same value as the rate of the clock, e.g. a $100 \text{Mb/s}$ transmission will require a bandwidth of approximately
100 MHz. Furthermore, note that the first side lobe (the maximum between the frequencies $1/T_s$ and $2/T_s$) is $-13.3 \text{ dB}$. This implies that a relatively large amount of energy is still present in the frequency band above $1/T_s$. This may cause interference with the neighboring frequency bands.

![Figure 8.4: The power spectral density of the non-return to zero signal of Fig. 8.2 with $a_0 = 0$ and $a_1 = 2$. The dashed line at $f = 0$, terminated with a 'o', represents the Dirac $\mu_X^2 \delta(f)$.](image-url)
Interpretation using filtered Dirac pulses.

An alternative interpretation / derivation of the power spectrum of a NRZ is using a series of Dirac pulses in time that carry the information

\[ X(t) = \sum_{n=-\infty}^{\infty} X_n \delta(t - nT_s) \]

which are filtered afterwards using a zero-order-hold impulse response with

\[ h_{ZOH}(t) = \begin{cases} 1 & 0 \leq t < T_s \\ 0 & \text{elsewhere} \end{cases} \]

in order to generate

\[ Y(t) = h_{ZOH}(t) \ast X(t). \]

If the data symbols \( X_n \) are zero-mean and independent, then it is known that

\[ E(X_n X_k) = E(|X_n|^2) \delta_{nk} \]

since \( X_n \) and \( X_k \) are independent.
Hence,

\[
\rho_{XX}(\tau) = E \left( \sum_{n=-\infty}^{\infty} X_n \delta(t + \tau - nT_s) \sum_{k=-\infty}^{\infty} X_k \delta(t - kT_s) \right)
\]

\[
= E \left( \sum_{n=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} X_n X_k \delta(t + \tau - nT_s) \delta(t - kT_s) \right)
\]

\[
= \sum_{n=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} E(X_n X_k) \delta(t + \tau - nT_s) \delta(t - kT_s)
\]

\[
= E(|X_n|^2) \sum_{n=-\infty}^{\infty} \delta(t + \tau - nT_s) \delta(t - nT_s)
\]

\[
= E(|X_n|^2) \delta(\tau).
\]

The relation between \(E(|X_n|^2)\) and \(\sigma_X^2\) is determined using the same method as before: we start with a time reference \(t_0\) and take the average / integrate over the (uniform) distribution of \(t_0\) in \([0, T_s]\)

\[
\sigma_X^2 = E \left( \sum_{n=-\infty}^{\infty} |X_n|^2 \delta(t_0 - nT_s) \right)
\]

\[
= \int_0^{T_s} E(|X_n|^2) dt
\]

\[
= T_s E(|X_n|^2).
\]

This shows that \(X(t)\) is a white noise process (Definition 45) with

\[
\Sigma_{XX}(f) = \frac{\sigma_X^2}{T_s} \forall f.
\]

The impact of the zero-order-hold filter on \(X(t)\) is given by Theorem 32 which states that

\[
\Sigma_{YY}(f) = |H(f)|^2 \Sigma_{XX}(f)
\]

with \(H(f) = \mathcal{F}(h_{ZOH}(t))\), implying that (see Table 8.1)

\[
|H(f)|^2 = |T_s \text{sinc} (fT_s)|^2.
\]

This leads to the same result as derived above

\[
\Sigma_{YY}(f) = \frac{\sigma_X^2}{T_s} |T_s \text{sinc} (fT_s)|^2
\]

\[
= T_s \sigma_X^2 \text{sinc}^2 (fT_s).
\]
Ergodicity of a NRZ signal

The NRZ signal is ergodic in the mean value (Theorem 28) since

$$\lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} \rho_{XX}(\tau) d\tau = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} \left( \sigma_X^2 \text{tri} \left( \frac{T}{T_s} \right) + \mu_X^2 \right) d\tau$$

$$= \mu_X^2.$$  

The NRZ signal is ergodic in the power spectral density (Theorem 35) only for a zero-mean process $\mu_X = 0$

$$\lim_{T \to \infty} \int_{-T/2}^{T/2} |\tau \rho_{XX}(\tau)| d\tau = \lim_{T \to \infty} \int_{-T/2}^{T/2} |\tau| \left( \sigma_X^2 \text{tri} \left( \frac{T}{T_s} \right) + \mu_X^2 \right) d\tau$$

$$= 2\sigma_X^2 \int_0^{T_s} \left( 1 - \frac{\tau}{T_s} \right) d\tau$$

$$= 2\sigma_X^2 \left[ \tau - \frac{\tau^2}{2T_s} \right]_0^{T_s}$$

$$= T_s \sigma_X^2 < \infty.$$  

8.6.5 Random Telegraph Signal

The Random Telegraph Signal (RTS) is a memoryless continuous-time stochastic process that shows two distinct values, 1 and -1. The transition between the two states are driven by a Poisson process. The name Random Telegraph Signal originates from the fact that one of its first application was the power spectrum of morse code over the telegraph. The RTS signals are used nowadays to model burst noise, also known as popcorn noise, impulse noise, bi-stable noise, and random telegraph signal (RTS) noise. An (important) application is flicker or $1/f$ noise that is present in microelectronics devices.

Poisson distribution

The RTS starts with a Poisson distribution which expresses the probability of a given number of events occurring in a given time window, assuming that the event rate is constant and independently of the time since the last event. It is possible to derive a Poisson distribution starting from different sets of assumptions (e.g. by conducting a large number of Bernoulli experiments). We will consider standalone assumptions for the Poisson distribution that don’t rely on another (binomial) distribution.

**Definition 48 (Poisson distribution).** The Poisson distribution gives the probability that $k$ events occur in a time window $t$

$$p_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$
For such Poisson process, the number of events $n(\Delta t)$ in a time window $\Delta t$ satisfy the following conditions:

1. the probability of the occurrence of one incident is proportional with the width of the time window $\Delta t$

$$P[n(\Delta t) = 1] = \lambda \Delta t + o(\Delta t)$$

where $\lambda$ is a constant value representing the average number of events per time unit, and where $o(\Delta t)$ converges sufficiently fast to zero such that $
\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$.

2. the probability of multiple incidents in a time window $\Delta t$ is negligible for $\Delta t$ tending to zero

$$P[n(\Delta t) > 1] = o(\Delta t)$$

with $\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$.

3. the number of incidents $n(\Delta t)$ of two non-overlapping time windows ($\Delta t_1$ and $\Delta t_2$ with $\Delta t_1 \cap \Delta t_2 = \emptyset$) are independent random variables (i.e. memory less)

$$P[n(\Delta t_1) = n_1 \cap n(\Delta t_2) = n_2] = P[n(\Delta t_1) = n_1]P[n(\Delta t_2) = n_2]$$

**Random Telegraph Signal**

**Definition 49** (Random Telegraph Signal). A Random Telegraph Signal (RTS) $X(t)$ is a random process which takes has two states with output values, $a$ and $-a$, with equal probability

$$P[X(t) = a] = P[X(t) = -a] = 1/2.$$ 

The transition from one state to the other is triggered by a Poisson process which is characterized by the average number of events per time-unit $\lambda$. A typical random telegraph signal is shown in Fig. 8.6.
The main value and the variance of the random telegraph signal are simple to compute as $X(t)$ has only a discrete number of outcome

\[
E(X(t)) = aP[X(t) = a] - aP[X(t) = -a] = 0
\]
\[
E(X^2(t)) = a^2P[X(t) = a] + a^2P[X(t) = -a] = a^2.
\]

To study the power spectrum generated a RTS, it is necessary to compute the ensemble auto-correlation.

**Theorem 38** (Auto-correlation Random Telegraph signal). The auto-correlation of WSS random telegraph signal equals

\[
\rho_{XX}(\tau) = a^2e^{-2\lambda|\tau|}
\]

where $\lambda$ represents the average number of events per time-unit.

**Proof.** A Poisson distribution with an average number of events per time-unit $\lambda$ gives the probability

\[
p_k(t) = \frac{(\lambda t)^k}{k!}e^{-\lambda t}
\]

that $k$ events take place in the window $t$.

If an even number of events take place in a time interval $\tau \geq 0$, then the outcome of $X(t + \tau)$ equals $X(t)$. An odd number of events will result into $X(t + \tau) = -X(t)$.

Figure 8.6: A realization of a random telegraph signal with both $a$ and $\lambda$ equal to one.
CHAPTER 8. SPECTRAL ANALYSIS OF STOCHASTIC PROCESSES 191

The probability of an even number of events in a window $\tau$ can be computed by

$$P[X(t + \tau) = X(t)] = P^{\text{even}}(\tau) = \sum_{k=0}^{\infty} p_{2k}(\tau) = e^{-\lambda\tau} \sum_{k=0}^{\infty} \frac{(\lambda\tau)^{2k}}{(2k)!}.$$  

The probability of an odd number of events in a window $\tau$ equals

$$P[X(t + \tau) = -X(t)] = P^{\text{odd}}(\tau) = P[\text{odd number of events in } t] = \sum_{k=0}^{\infty} p_{2k+1}(\tau) = e^{-\lambda\tau} \sum_{k=0}^{\infty} \frac{(\lambda\tau)^{2k+1}}{(2k+1)!}.$$  

First, consider the auto-correlation for $t_1 \geq t_2$

$$E(X(t_1)X(t_2)) = a^2 P[X(t_1) = a \cap X(t_2) = a] + a^2 P[X(t_1) = -a \cap X(t_2) = -a] - a^2 P[X(t_1) = a \cap X(t_2) = -a] - a^2 P[X(t_1) = -a \cap X(t_2) = a]$$

The joint probabilities can be rewritten using a conditional probability, e.g.

$$P[X(t_1) = a \cap X(t_2) = a] = P[X(t_1) = a]P[X(t_2) = a] = \frac{1}{2} P[X(t_1) = X(t_2)] = \frac{1}{2} P^{\text{even}}(t_1 - t_2)$$

This implies that

$$E(X(t_1)X(t_2)) = a^2 P^{\text{even}}(t_1 - t_2) - a^2 P^{\text{odd}}(t_1 - t_2) = a^2 e^{-\lambda(t_1 - t_2)} \left( \sum_{k=0}^{\infty} \frac{(-\lambda(t_1 - t_2))^k}{k!} \right) = a^2 e^{-2\lambda|t_1 - t_2|}.$$  

The same reasoning can be made for $t_2 \geq t_1$, resulting into

$$\rho_{XX}(\tau) = E(X(t + \tau)X(t)) = a^2 e^{-2\lambda|\tau|}$$
An example of the PSD of a RTS is shown in Fig. 8.7.

The auto-correlation of the RTS signals allows the computation of the power spectrum.

**Theorem 39.** The power spectrum of a RTS with parameter $\lambda$ equals

$$
\Sigma_{XX}^{RTS}(f) = F(\rho_{XX}(\tau)) \\
= a^2 F(e^{-2\lambda|\tau|}) \\
= a^2 \frac{4\lambda}{4\lambda^2 + (2\pi f)^2} \\
= a^2 \frac{\lambda}{\lambda^2 + \pi^2 f^2}.
$$

**Properties of Random Telegraph Signal**

The Random Telegraph Signal originates from a WSS process as the underlying process is a Poisson process (memoryless and a constant rate). Hence, both the mean value and the variance are constant over time (Theorem 17): $E(X(t)) = 0$ and $\sigma_X^2(t) = a^2$.

It was already proven that $E(X(t)) = 0$, but it is also confirmed by (Theo-
rem 20)  

\[ E(X(t)) = \lim_{\tau \to \infty} \rho_{XX}(\tau) = \lim_{\tau \to \infty} a^2 e^{-2\lambda |\tau|} = 0 \]

The same theorem can be used for the variance \( \sigma_X^2 = a^2 \) as

\[ \sigma_X^2 = \rho_{XX}(0) - \mu_X^2 = a^2 \]

The power spectral density of a random telegraph signal has a low-pass characteristic with a 3 dB point at \( f_{3\, \text{dB}} = \frac{\lambda}{\pi} \). An example of the PSD of a RTS is shown in Fig. 8.8.

![Figure 8.8: Power spectral density of a random telegraph signal with both \( a \) and \( \lambda \) equal to one.](image-url)
The Random Telegraph Signal is mean-ergodic (Theorem 28) since
\[
\lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} \rho_{XX}(\tau) d\tau = \lim_{T \to \infty} \frac{a^2}{T} \int_{-T/2}^{T/2} e^{-2\lambda|\tau|} d\tau
\]
\[
= \lim_{T \to \infty} \frac{2a^2}{T} \int_{0}^{T/2} e^{-2\lambda \tau} d\tau
\]
\[
= - \lim_{T \to \infty} \frac{2a^2}{T} \frac{e^{-2\lambda T/2}}{2\lambda} \bigg|_{0}^{T/2}
\]
\[
= - \lim_{T \to \infty} \frac{2a}{\lambda T} \left(e^{-\lambda T} - 1\right)
\]
\[
= 0 = \mu^2_x.
\]

The Random Telegraph Signal is ergodic in the power spectral density (Theorem 35) since
\[
\lim_{T \to \infty} \int_{-T/2}^{T/2} |\tau \rho_{XX}(\tau)| d\tau = \lim_{T \to \infty} \frac{a^2}{T} \int_{-T/2}^{T/2} |\tau| e^{-2\lambda|\tau|} d\tau
\]
\[
= \lim_{T \to \infty} 2a^2 \int_{0}^{T/2} \tau e^{-2\lambda \tau} d\tau
\]
\[
= - \lim_{T \to \infty} \frac{2a^2}{2\lambda^2} \left( (\lambda T + 1)e^{-\lambda T} - 1 \right)
\]
\[
= \frac{a^2}{2\lambda^2} < \infty.
\]

8.6.6 Flicker or 1/f noise

Suppose that the process \( Z(t) \) is composed out of the sum of a large number of orthogonal WSS processes \( X(t, \lambda) \) which depend on both the time \( t \) and a time-invariant stochastic parameter \( \lambda \). As the parameter \( \lambda \) is a stochastic variable, it has a pdf \( p_{\Lambda}(\lambda) \).

Considering the sum of orthogonal processes (Theorem 31) enables the computation of the auto-correlation and the power spectrum of \( Z(t) \) using
\[
\rho_{ZZ}(\tau) = \int \rho_{XX}(\tau, \lambda) f_{\Lambda}(\lambda) d\lambda
\]
and
\[
\Sigma_{ZZ}(f) = \int \Sigma_{XX}(f, \lambda) f_{\Lambda}(\lambda) d\lambda.
\]

Assume that the process \( Z(t) \) is composed out of a large number of uncorrelated random telegraph signals \( X(t, \lambda) \) with parameter \( \lambda \) and amplitude \( a \). As
all the components are uncorrelated (by assumption) and zero-mean (property of the RTS), they are also orthogonal (Lemma 19). This makes it possible to compute

\[ \Sigma_{ZZ}(f) = \int \Sigma_{XX}^{RTS}(f, \lambda) f_{\Lambda}(\lambda) d\lambda. \]

To obtain flicker or \(1/f\) noise, it is necessary to assume that \(f_{\Lambda}(\lambda)\) is inversely proportional to \(\lambda\)

\[ f_{\Lambda}(\lambda) = \frac{1}{\lambda \ln(\lambda_2/\lambda_1)} \]

in the range \([\lambda_1, \lambda_2]\). The scaling factor \(\ln(\lambda_2/\lambda_1)\) guarantees that

\[ \int_{\lambda_1}^{\lambda_2} f_{\Lambda}(\lambda) d\lambda = 1. \]

The PSD of \(Z(t)\) then becomes

\[ \Sigma_{ZZ}(f) = \frac{a^2}{\ln(\lambda_2/\lambda_1)} \int_{\lambda_1}^{\lambda_2} \frac{1}{\lambda^2 + \pi^2 f^2} d\lambda \]

\[ = \frac{a^2}{\ln(\lambda_2/\lambda_1) \pi f} \left( \arctan \left( \frac{\lambda_2}{\pi f} \right) - \arctan \left( \frac{\lambda_1}{\pi f} \right) \right). \]

This expression can be simplified under the condition that \(\lambda_1 \ll \pi f \ll \lambda_2\)

\[ \Sigma_{ZZ}(f) = \frac{a^2}{\ln(\lambda_2/\lambda_1) \pi} \frac{1}{2f} \]

as

\[ \lim_{x \to \infty} \arctan(x) = \pi/2 \]

and

\[ \arctan(0) = 0. \]

Observe that the power spectrum density \(\Sigma_{ZZ}(f)\) decreases proportionally to \(1/f\). The question remains: what kind of physical processes have a distribution that is inversely proportional to \(\lambda\) (8.4)?

**Example 37.** Suppose the carrier trap in semiconductor (i.e. imperfections that can trap charged carriers such as electrons) occurs by the tunneling of charged carriers from a conducting layer to traps inside the oxide layer at depth \(w\). The average number of such events decreases exponentially with the depth \(w\) due to the tunneling effect. Hence,

\[ \lambda(w) = g(w) = \lambda_0 \exp(-\gamma w) \]

where \(\lambda_0\) and \(\gamma\) are constants. If the traps are uniformly distributed between the depth \(w_1\) and \(w_2\), we obtain \(f_{\Lambda}(\lambda)\) through the nonlinear transformation of
the uniform pdf of $w$ ($f_W(w)$), namely that

$$f_{\Lambda}(\lambda) = \frac{1}{|\frac{\partial g(w)}{\partial w}|} f_W(w)$$

$$= \frac{1}{|\gamma \lambda_0 \exp(-\gamma w)|} f_W(w)$$

$$= \frac{1}{|\gamma \lambda(w)|} f_W(w)$$

In conclusion: if $w$ is uniformly distributed between $w_1$ and $w_2$ with $f_W(w) = 1/|w_2 - w_1|$, and if $\lambda(w) = \lambda_0 \exp(-\gamma w)$, then

$$f_{\Lambda}(\lambda) = \frac{1}{|\gamma \lambda(w)|} \frac{1}{|w_2 - w_1|}$$

and hence the condition to generate $1/f$ noise is fulfilled.

8.6.7 Power spectra of modulated processes

Section 7.3.5 showed that modulated WSS processes result into WSS processes. This implies that Wiener-Kinchin

$$\Sigma_{XX}(f) \triangleq \mathcal{F}(\rho_{XX}(\tau))$$

can be used to determine the power spectra of those modulated signals.

Two cases were studied: the amplitude / DSB modulation (Theorem 26) and the more general IQ modulation (Theorem 27).

**Double-Side Band**

Theorem 26 showed that the amplitude modulation of a real-valued WSS process $X(t)$

$$Y(t) = X(t) \cos(2\pi f_c t + \phi)$$

with a fixed frequency $f_c$ and a uniformly distributed initial phase $\phi$ is WSS with

$$\rho_{YY}(\tau) = \frac{1}{2} \rho_{XX} \cos(2\pi f_c \tau).$$

Compute the Fourier transform of this cross-correlation can easily be done using Table 8.1, resulting in

$$\Sigma_{YY}(f) = \mathcal{F}(\rho_{YY}(\tau))$$

$$= \mathcal{F} \left( \frac{1}{2} \rho_{XX}(\tau) \cos(2\pi f_c \tau) \right)$$

$$= \frac{1}{4} \left[ \Sigma_{XX}(f - f_c) + \Sigma_{XX}(f + f_c) \right].$$
This illustrates that the Double Side Band modulation basically shifts the power spectrum up and down with a frequency shift of $f_c$. The power spectrum of the real-valued process $X(t)$ is symmetrical

$$\Sigma_{XX}(f) = \Sigma_{XX}(-f).$$

Therefore, the side bands around the modulation carrier are symmetrical (hence the name DSB).

**Amplitude Modulation**

If $X(t)$ contains a DC value $\mu_X$, then the original power spectrum contains a Dirac at DC $\mu_X^2 \delta(f)$. This will introduce two Dirac functions in the power spectrum of the modulated signal.

**Example 38.** Consider the example of the binary Non-Return-to-Zero (Section 8.6.4) with power spectrum

$$\Sigma_{XX}^{NRZ}(f) = Ts\sigma_X^2 \text{sinc}^2(fTs) + \mu_X^2 \delta(f).$$

Assume that the NRZ signal is modulated with a carrier frequency of $f_c$

$$Y(t) = X(t) \cos(2\pi f_c t + \phi).$$

The power spectrum of $Y(t)$ then equals

$$\Sigma_{YY}(f) = \frac{T_s\sigma_X^2}{4} \left( \text{sinc}^2((f-f_c)T_s) + \text{sinc}^2((f+f_c)T_s) \right)$$

$$+ \frac{\mu_X^2}{4} \left( \delta(f-f_c) + \delta(f+f_c) \right).$$

and is shown in Fig. 8.9 for $f_c = 2.5$. 
IQ modulation

Theorem 27 showed that the IQ modulation of a two real-valued mutual WSS processes $I(t)$ and $Q(t)$

$$Y(t) = I(t) \cos (2\pi f_c t) - Q(t) \sin (2\pi f_c t)$$

with a fixed frequency $f_c$ is WSS if

$$E(I(t)) = E(Q(t)) = 0$$

and

$$\rho_{II}(\tau) = \rho_{QQ}(\tau)$$
$$\rho_{IQ}(\tau) = -\rho_{QI}(\tau).$$

Its auto-correlation equals

$$\rho_{YY}(\tau) = \rho_{II}(\tau) \cos (2\pi f_c \tau) + \rho_{IQ}(\tau) \sin (2\pi f_c \tau).$$

In a classical IQ modulation it is assumed that $I(t)$ and $Q(t)$ independent, i.e. $\rho_{II}(\tau) = \rho_{QQ}(\tau)$ and $\rho_{IQ}(\tau) = -\rho_{QI}(\tau) = 0$ implies that the auto-correlation equals

$$\rho_{YY}(\tau) = \rho_{II}(\tau) \cos (2\pi f_c \tau).$$
This results in a similar expression (besides a factor $1/2$) for the power spectrum compared to a DSB/AM modulation

$$\Sigma_{YY}(f) = \frac{1}{2} [\Sigma_{II}(f - f_c) + \Sigma_{II}(f + f_c)].$$

### Single-Side Band Modulation

In order to suppress one of the two side bands, the Single-Side Band (SSB) Modulation constructs the quadrature signal such that it equals the Hilbert transform of the in-phase signal $I(t)$

$$Q(t) = h_{Hilbert}(t) * I(t)$$

where the Hilbert filter $h_{Hilbert}(t)$ satisfy (by construction)

$$h_{Hilbert}(t) = \frac{1}{\pi t}$$

and

$$\mathcal{F}(h_{Hilbert}(t)) = H_{Hilbert}(f) = \begin{cases} -j & f > 0, \\ 0 & f = 0, \\ j & f < 0. \end{cases}$$

Note that the condition $\rho_{IQ}(\tau) = -\rho_{QI}(\tau)$ is satisfied since

$$\rho_{IQ}(\tau) \overset{(a)}{=} \rho_{II}(\tau) * h_{Hilbert}^*(-\tau) \overset{(b)}{=} -\rho_{II}(\tau) * h_{Hilbert}(\tau) \overset{(a)}{=} -\rho_{QI}(\tau)$$

using (a) Theorem 24 and (b) the fact that $h_{Hilbert}^*(-t) = -h_{Hilbert}(t)$.

It can then be shown that the Fourier transform of $\rho_{IQ}(\tau) \sin(2\pi f_c \tau)$ satisfies

$$\mathcal{F}(\rho_{IQ}(\tau) \sin(2\pi f_c \tau)) = \frac{1}{2j} [\Sigma_{IQ}(f - f_c) - \Sigma_{IQ}(f + f_c)]$$

$$= \frac{1}{2j} [H_{Hilbert}^*(f - f_c) \Sigma_{II}(f - f_c) - H_{Hilbert}^*(f + f_c) \Sigma_{II}(f + f_c)]$$

using (a) Table 8.1, and (b) the results on the cross power spectra of filtered signals (Theorem 32). Different frequency regions can now be considered

- $f > f_c$: the transfer functions are equal to each other
  
  $$H_{Hilbert}^*(f - f_c) = H_{Hilbert}^*(f + f_c) = j.$$

- $-f_c < f < f_c$: the transfer functions have opposite signs
  
  $$H_{Hilbert}^*(f - f_c) = -H_{Hilbert}^*(f + f_c) = -j.$$
CHAPTER 8. SPECTRAL ANALYSIS OF STOCHASTIC PROCESSES

$f < -f_c$: the transfer functions are equal to each other

\[ H_{\text{Hilbert}}(f - f_c) = H_{\text{Hilbert}}(f + f_c) = -j. \]

Compute the Fourier transform of the cross-correlation \( \rho_{YY}(\tau) \) can easily be done considering

\[
|f| > f_c: \quad \Sigma_{YY}(f) = \mathcal{F}(\rho_{YY}(\tau)) = \mathcal{F}(\rho_{II}(\tau) \cos (2\pi f_c \tau)) = \frac{1}{2} [\Sigma_{II}(f - f_c) + \Sigma_{II}(f + f_c)],
\]

\[-f_c < f < f_c: \quad \Sigma_{YY}(f) = \mathcal{F}(\rho_{YY}(\tau)) = \mathcal{F}(\rho_{II}(\tau) \cos (2\pi f_c \tau)) + \mathcal{F}(\rho_{IQ}(\tau) \sin (2\pi f_c \tau)) = \frac{1}{2} [\Sigma_{II}(f - f_c) + \Sigma_{II}(f + f_c)] + \frac{1}{2}j [-j \Sigma_{II}(f - f_c) - j \Sigma_{II}(f + f_c)] = 0.
\]

This clearly illustrates the concept of the Single Side Band: in this case, only the upper side band \((f > f_c)\) is carrying the information, while the lower side band \((0 < f < f_c)\) remains untouched.
Chapter 9

Karhunen–Loève expansion

An arbitrary deterministic function $f(t)$ can be developed in series

$$f(t) = \sum_{n=1}^{\infty} b_n \varphi_n(t) \quad \text{for } |t| < \frac{T}{2}$$

when a sequence (of possibly an infinite number) of orthonormal (possibly complex) base function $\varphi_i(t)$ are chosen in the interval $[-T/2, T/2]$. The orthonormality of the base functions is expressed using the inner product

$$<\varphi_i(t), \varphi_j(t)> = \int_{-T/2}^{T/2} \varphi_i(t)\varphi_j^*(t)dt$$

demanding that

$$<\varphi_i(t), \varphi_j(t)> = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

The series will converge when the sequence of base functions form a complete bases.

This chapter will extend the series expansion techniques for deterministic functions towards stochastic processes / signals. The aim is to represent the stochastic signal $X(t)$ as series expansion

$$X(t) \approx \sum_{n=1}^{\infty} b_n \varphi_n(t) \quad \text{for } |t| < \frac{T}{2}$$

where $\varphi_n(t)$ are deterministic orthonormal basis functions and where $b_n$ are stochastic variables. If the set of basis functions is complete, then the approximation becomes an equality

$$X(t) = \sum_{n=1}^{\infty} b_n \varphi_n(t) \quad \text{for } |t| < \frac{T}{2}.$$
The orthonormality of the basis function enables the simple and efficient computation of the coefficients $b_n$ for a realization $X^{(k)}(t)$ of the process $X(t)$, namely

$$b_n = \int_{-T/2}^{T/2} X^{(k)}(t) \varphi_n^*(t) \, dt$$

since

$$\int_{-T/2}^{T/2} X^{(k)}(t) \varphi_n^*(t) \, dt = \int_{-T/2}^{T/2} \left( \sum_{k=1}^{\infty} b_k \varphi_k(t) \right) \varphi_n^*(t) \, dt = \sum_{k=1}^{\infty} b_k \int_{-T/2}^{T/2} \varphi_k(t) \varphi_n^*(t) \, dt = \sum_{k=1}^{\infty} b_k \delta_{kn} = b_n.$$

The question that remains is: how to find the basis functions $\varphi_i(t)$. Therefore, the same line of thought is used as for the power spectral density: instead of developing a series for one particular realization, the expansion (known as the Karhunen-Loève theorem) will be done on the statistical moments of the ensemble of all possible realizations.

### 9.1 Fourier expansion

Consider as example a Fourier series expansion

$$X(t) = \sum_{n=-\infty}^{\infty} a_n e^{j2\pi n ft} \quad \text{for } |t| < \frac{T}{2}$$

and hence $\varphi_n(t) = e^{j2\pi n ft}$.

The basis function of a Fourier series are not orthonormal for an arbitrary (finite) value of $T$. Consider therefore the inner product

$$< \varphi_n(t), \varphi_m(t) > = \int_{-T/2}^{T/2} e^{j2\pi n ft} e^{-j2\pi m ft} \, dt = \int_{-T/2}^{T/2} e^{j2\pi (n-m) ft} \, dt$$

which must be equal to $\delta_{nm}$ for an orthonormal base. This puts a constraint between the considered time window $T$ and the fundamental frequency $f$, namely $T = k/f$ for integer $k$. This implies that the time window must contain an integer number of periods of the basis functions.

Note that the Fourier series expansion has deterministic coefficients ($a_n$) and the expansion basis consists of sinusoidal functions (sine and cosine). The coefficients in the Karhunen-Loève theorem ($b_n$) are random variables and the
expansion basis depends on the process. In fact, their orthogonal basis functions are determined using the auto-correlation of the process.

**Example 39** (OFDM). Orthogonal frequency division multiplexing (OFDM) is an example of a modulation scheme that uses a large number of frequency carriers in parallel to transport the information. Each carrier is modulated separately and acts as an independent information channel. To remove any interference between the different carriers/channels, the OFDM modulation makes that all the carriers/channels are orthogonal to each other. This is done by demanding that the frequency spacing in-between the different carriers (which corresponds with $f$ in the Fourier series) equals $1/T$.

### 9.2 Karhunen-Loève expansion

The Karhunen-Loève expansion is a representation of a stochastic process as an infinite linear combination of orthogonal functions, analog to a Fourier series representation of a function on a bounded interval. There exist many such expansions for stochastic processes but the Karhunen-Loève theorem yields the best basis in the sense that it minimizes the total mean squared error.

In contrast to a Fourier series, where the coefficients are fixed numbers and the expansion basis consists of sinusoidal functions, the coefficients in the Karhunen-Loève expansion are random variables and the expansion basis depends on the auto-correlation of the process.

Consider the series development of a zero-mean stochastic signal

$$X(t) = \sum_{n=1}^{\infty} b_n \varphi_n(t) \quad \text{for } |t| < \frac{T}{2}$$

with orthonormal basis function $\varphi_n(t)$ which satisfy $< \varphi_i(t), \varphi_j(t) >= \delta_{ij}$.

**Lemma 21.** A zero-mean stochastic signal $X(t)$ directly implies that the coefficients $b_n$ are also zero mean $E(b_n) = 0$ for all $n$.

**Proof.** Considering that

$$E(X(t)) = \sum_{n=1}^{\infty} E(b_n)\varphi_n(t) = 0$$

must be valid for all $|t| < T/2$ and the fact that $\varphi_n(t)$ is a non-zero function of $t$ implies that $E(b_n) = \mu_{b_n} = 0$. $\square$

As the coefficients $b_n$ are stochastic, it would be favorable that the coefficients $b_n$ would be orthogonal, demanding that $E(b_n b_m^*) = \sigma_{b_n}^2 \delta_{nm} + \mu_n \mu_m = \sigma_{b_n}^2 \delta_{nm} + \mu_n \mu_m = \sigma_{b_n}^2 \delta_{nm} + \mu_n \mu_m = \sigma_{b_n}^2 \delta_{nm}$.

**Theorem 40** (Karhunen-Loève theorem). The coefficients $b_n$ of the series development of a zero-mean stochastic signal

$$X(t) = \sum_{n=1}^{\infty} b_n \varphi_n(t) \quad \text{for } |t| < \frac{T}{2}$$
are orthogonal, i.e. \( E(b_n b_m^*) = 0 \) for \( n \neq m \) if and only if the functions \( \varphi_n(t) \) satisfy the integral equation

\[
\int_{-T/2}^{T/2} \rho_{XX}(t_1, t_2) \varphi_n(t_2) dt_2 = \lambda_n \varphi_n(t_1) \quad \text{for} \ |t_1| < \frac{T}{2}
\]  

with eigenvalues \( \lambda_n = \sigma_{b_n}^2 \).

**Proof.** The assumption that \( X(t) \) is zero mean implies that the mean values of all the \( b_n \) coefficients are zero (see previous lemma): \( E(b_n) = \mu_{b_n} = 0 \). Hence, the variance of \( b_n \) equals

\[
\sigma_{b_n}^2 = E(|b_n - \mu_{b_n}|^2) = E(|b_n|^2).
\]

The integral in (9.1) can be rewritten as

\[
\int_{-T/2}^{T/2} \rho_{XX}(t_1, t_2) \varphi_n(t_2) dt_2 = \int_{-T/2}^{T/2} E(X(t_1)X^*(t_2)) \varphi_n(t_2) dt_2
\]

\[
\overset{(a)}{=} E\left( \int_{-T/2}^{T/2} X(t_1)X^*(t_2)\varphi_n(t_2) dt_2 \right)
\]

\[
\overset{(b)}{=} E\left( X(t_1) \int_{-T/2}^{T/2} X^*(t_2)\varphi_n(t_2) dt_2 \right)
\]

\[
\overset{(c)}{=} E\left( X(t_1) \int_{-T/2}^{T/2} \sum_{k=1}^{\infty} b_k^* \varphi_k(t_2) \varphi_n(t_2) dt_2 \right)
\]

\[
\overset{(d)}{=} E\left( X(t_1) \sum_{k=1}^{\infty} b_k^* \left( \int_{-T/2}^{T/2} \varphi_n(t_2)\varphi_k^*(t_2) dt_2 \right) \right)
\]

\[
\overset{(e)}{=} E\left( X(t_1) \sum_{k=1}^{\infty} b_k^* \delta_{nk} \right)
\]

where (a) interchanges the expected value with the integration, (b) uses the linearity of the integral, (c) substitutes \( X(t_2) \) with its series expansion, (d) uses the linearity of the integration, and (e) uses the orthogonality of the basis function \( \varphi_n(t) \).

Substituting \( X(t_1) \) with its series expansion finally results into

\[
\int_{-T/2}^{T/2} \rho_{XX}(t_1, t_2) \varphi_n(t_2) dt_2 = E\left( b_n^* \sum_{k=1}^{\infty} b_k \varphi_k(t_1) \right)
\]

\[
\overset{(a)}{=} \sum_{k=1}^{\infty} E(b_k b_n^*) \varphi_k(t_1)
\]

\[
\overset{(b)}{=} E(b_n b_n^*) \varphi_n(t_1)
\]

\[
= \sigma_{b_n}^2 \varphi_n(t_1)
\]
where (a) uses the fact that the basis functions are deterministic and don’t depend on the realization of a stochastic variable, and (b) uses the assumption that \( E(b_kb_k^*) = 0 \) for \( k \neq n \).

If, additionally to the assumptions made in Theorem 40, the process is wide-sense stationary, then (9.1) becomes

\[
\int_{-T/2}^{T/2} \rho_{XX}(t-\tau)\varphi_n(\tau)d\tau = \lambda_n\varphi_n(t) \quad \text{for } |t| < T/2
\]

Note that the eigenvalues \( \lambda_n = \sigma^2_{b_n} \) equals the variance on the coefficient \( b_n \). Hence, it is possible to determine the variability of the coefficients (= the amount of information) irrespective of the actual realization of the process.

An important observation is that the random coefficients \( b_n \) of the Karhunen-Loève expansion are uncorrelated \( (E(b_nb_m^*) = 0 \text{ for } n \neq m) \). This leads to the following theorem.

**Theorem 41.** The total power of the (zero-mean) process \( X(t) \), which equals its integrated variance over the time interval \([-T/2, T/2]\), equals the sum of the eigenvalues

\[
\int_{-T/2}^{T/2} \sigma^2_X(t)dt = \sum_{n=1}^{\infty} \lambda_n. \quad (9.2)
\]

**Proof.** As the process is zero-mean, it is known that

\[
\sigma^2_X(t) = E(||X(t)||^2)
\]

\[
= E\left( \sum_{n=1}^{\infty} b_n\varphi_n(t) \right) \left( \sum_{m=1}^{\infty} b_m^*\varphi_m^*(t) \right)
\]

\[(a)\]

\[
= E\left( \sum_{n=1}^{\infty} b_n b_n^* \varphi_n(t) \varphi_n^*(t) \right)
\]

\[(b)\]

\[
= \sum_{n=1}^{\infty} \lambda_n \varphi_n(t) \varphi_n^*(t)
\]

where (a) uses \( E(b_nb_m^*) = 0 \text{ for } n \neq m \), and (b) that \( \lambda_n = E(b_nb_n^*) \).

Integration over a time interval \([-T/2, T/2]\) yields

\[
\int_{-T/2}^{T/2} \sigma^2_X(t)dt = \int_{-T/2}^{T/2} \sum_{n=1}^{\infty} \lambda_n \varphi_n(t) \varphi_n^*(t)dt
\]

\[= \sum_{n=1}^{\infty} \lambda_n \int_{-T/2}^{T/2} \varphi_n(t) \varphi_n^*(t)dt
\]

\[= \sum_{n=1}^{\infty} \lambda_n \]
where (a) uses the fact that the functions are normalized such that
\[ \int_{-T/2}^{T/2} \varphi_n(t) \varphi^*_n(t) dt = 1. \]

An alternative proof uses
\[ \int_{-T/2}^{T/2} \sigma^2_X(t) dt = \int_{-T/2}^{T/2} E(|X(t)|^2) dt \]
\[ = E \left( \int_{-T/2}^{T/2} \left( \sum_{n=1}^{\infty} b_n \varphi_n(t) \right) \left( \sum_{m=1}^{\infty} b_m^* \varphi^*_m(t) \right) \right) dt \]
\[ = E \left( \sum_{n=1}^{\infty} b_n b_n^* \left( \int_{-T/2}^{T/2} \varphi_n(t) \varphi^*_n(t) dt \right) \right) \]
\[ = \sum_{n=1}^{\infty} E (b_n b_n^*) = \sum_{n=1}^{\infty} \lambda_n \]

where (a) uses that
\[ \int_{-T/2}^{T/2} \varphi_n(t) \varphi^*_m(t) dt = 0 \quad \text{for} \quad n \neq m \]
and (b) uses
\[ \int_{-T/2}^{T/2} \varphi_n(t) \varphi^*_n(t) dt = 1. \]

This theorem leads to an important interpretation of the eigenvalues. The eigenvalues represent the contributions in variance (= power) of the individual basis function, and hence it become possible to determine the amount of power that remains unexplained when truncating the series expansion.

The total variance of the \( N \)-truncated approximation equals
\[ \sum_{n=1}^{N} \lambda_n. \]

and hence this approximation explains
\[ \frac{\sum_{n=1}^{N} \lambda_n}{\sum_{n=1}^{\infty} \lambda_n} \]

of the variance. If we aim an approximation that explains, e.g. 95\% of the variance, then we can determine an \( N \) such that
\[ \sum_{n=1}^{N} \lambda_n \geq 0.95 \sum_{n=1}^{\infty} \lambda_n. \]
The advantage of the expansion is its optimal convergence. This is why the Karhunen-Loève expansion is important in data compression: the basis functions are computed using the (known) auto-correlation of the process, independently of the realization. They can therefore be made available to the receiver beforehand. Instead of sending the signal through the channel as such, one only has to transmit the coefficients. The receiving side can then reconstruct the signal using the series expansion.

9.3 Karhunen-Loève expansion applied to band-limited signals

Consider an ideal low-pass filtered white noise WSS process \( X(t) \) as defined in Definition 47 with

\[
\Sigma_{XX}(f) = \begin{cases} \frac{N_0}{2} = Cte & |f| < W \\ 0 & |f| \geq W \end{cases}
\]

and auto-correlation

\[
\rho_{XX}(\tau) = N_0 \frac{\sin(2\pi W \tau)}{2\pi \tau}.
\]

This is an adequate model for the noise that is filtered using a (sharp) filter. The Karhunen-Loève expansion now provides the means to determine an optimal series expansion for this process.

The Karhunen-Loève expansion for an ideal low-pass filtered white noise needs to determine the basis function \( \varphi_n(t) \) that satisfy the integral equation (9.1) of the Karhunen-Loève theorem (Theorem 40)

\[
\int_{-T/2}^{T/2} N_0 \frac{\sin(2\pi W (t - \tau))}{2\pi (t - \tau)} \varphi_n(\tau)d\tau = \lambda_n \varphi_n(t) \quad \text{for } |t| < \frac{T}{2}.
\] (9.4)

This equation has been studied by [23] and its solutions

\[
\psi_n(t; c) \quad \text{with } 2c = 2\pi f_c T
\]

are known as prolate spheroidal wave functions. Note that these functions are defined starting from \( n = 0 \), instead of from \( n = 1 \) as used in Section 9.2. It is easy to see that, except for a scale factor, that the eigenvalues \( \lambda_n(c) \) and the eigenfunctions \( \psi_n(t; c) \) of the equation

\[
\int_{-T/2}^{T/2} N_0 \frac{\sin(2\pi W (t - \tau))}{2\pi (t - \tau)} \psi_n(\tau; c)d\tau = \lambda_n(c)\psi_n(t; c) \quad \text{for } \forall t
\] (9.5)

only depend on \( c = \pi WT \).
CHAPTER 9. KARHUNEN–LOÈVE EXPANSION

Note that $\psi_n(t; c)$ satisfies the integral equation (9.5) for all possible $t$, while the Karhunen-Loève theorem only requires that the integral equation (9.4) is fulfilled for $|t| < T/2$. Hence, the prolate spheroidal wave functions can be used to extrapolate a signal over a large time (i.e. $|t| \geq T/2$). It should be noted that such extrapolation is always a risky operation!

**Numerical example for the prolate spheroidal functions for $c = 4$**

Fig. 9.1 shows the first four eigenfunctions $\psi_n(t; c)$ for positive $t$ ($0 \leq t \leq 4T$) and for $c = 4$ [23]. The Karhunen-Loève theorem uses the support of this function in the region $0 \leq 2t/T < 1$. The rest of the presented functions ($1 \leq 2t/T \leq 4$) can be used to extrapolate the signal. The corresponding eigenvalues are tabulated in Table 9.1.

It is also possible to determine the percentage of the variance that can be explained using (9.3). The approximation ratio for various $N$ are listed in Table 9.2 and clearly shows that (in this case) only a small number of basis functions are necessary to accurately describe the signal.

![Figure 9.1: Eigenfunctions corresponding to the prolate spheroidal wave functions for $c = 4$ [23].](image)

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\lambda_n(4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.996</td>
</tr>
<tr>
<td>1</td>
<td>0.912</td>
</tr>
<tr>
<td>2</td>
<td>0.519</td>
</tr>
<tr>
<td>3</td>
<td>0.110</td>
</tr>
<tr>
<td>4</td>
<td>0.009</td>
</tr>
<tr>
<td>5</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Table 9.1: Eigenvalues corresponding to the prolate spheroidal wave functions for $c = 4$ [23].
CHAPTER 9. KARHUNEN–LOÈVE EXPANSION

Table 9.2: Approximation ratio (9.3) of an $N$-truncated approximation for $c = 4$.

<table>
<thead>
<tr>
<th>$N$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sum_{n=0}^{N} \lambda_n / \sum_{n=0}^{\infty} \lambda_n$</td>
<td>39.10%</td>
<td>74.90%</td>
<td>95.31%</td>
<td>99.64%</td>
<td>99.98%</td>
<td>$\approx$ 100%</td>
</tr>
</tbody>
</table>

9.3.1 Discrete Karhunen-Loève expansion applied to band-limited signals

The practice it is very difficult to solve the integral equation (9.1). Even the simple band-limited example defined earlier leads to complex analytical expressions using prolate spheroidal functions. Approximate methods are therefore proposed to make approximate the Karhunen-Loève expansion as closely as possible [24].

We will consider the discrete approximation of the Karhunen-Loève expansion applied to band-limited signals by

- evaluating the continuous time functions $\varphi_n(t)$ in discrete time point $t_k = k\delta T$;
- stacking the values $\varphi_n(t_k)$ in a column vector $\Phi$;
- approximating the integral equation (9.1) of the Karhunen-Loève theorem by a summation

$$\int_{-T/2}^{T/2} \rho_{XX}(t_m - \tau) \varphi_n(\tau) d\tau \approx \sum_{k=-K}^{K} \rho_{XX}(t_m - t_k) \varphi_n(t_k) \delta T$$

with $2K\delta T = T$ and $t_0 = 0$;
- rewriting the summation as a matrix product $R\Phi$ where the element $(m,k)$ of the matrix $R$ equals $\rho_{XX}(t_m - t_k) \delta T$.

All these steps result in the approximation of the integral equation (9.1) by an eigenvalue problem $R\Phi = \lambda\Phi$. This eigenvalue problem can be solved using classical linear algebra numerical tools (e.g. using the “eig” function in Matlab). The eigenvalues $\lambda$ of this eigenvalue problem correspond to the eigenvalues $\lambda_n$ of the Karhunen-Loève expansion; the eigenvectors correspond with the basis functions $\varphi_n(t)$.

Numerical example for the discrete Karhunen-Loève expansion for band-limited signals

Consider the same example as in Section (9.3) with $c = 4$. The aim is to illustrate the similarities and differences between the continuous and discrete version.

In order to transform the continuous time problem into discrete time, we discretized $T$ equidistantly into 1000 segments. This results in a matrix $R$ which
is visualized in Fig. 9.2. After determining the eigenvalues and eigenvectors using Matlab, we considered only the four largest eigenvalues. The eigenvectors and the correspond eigenvalues are shown in Fig. 9.3 and Table 9.3. Note that these eigenvalues corresponds almost exactly with the results with the prolate spheroidal wave functions (Table 9.1).

Figure 9.2: Visual representation of the matrix $R$ for the discrete approximation of (9.4) for $c = 4$. 
These eigenvectors are the discrete versions of their continuous time equivalents. In order to compare the discrete and the continuous time results, we need to rescale the eigenvector as the normalization in an eigenvalue problem differs from the normalization in the prolate spheroidal wave functions. After re-normalizing, we obtain Fig. 9.4. This figure can be compared with Fig. 9.1, considering the x-axis $2t/T \in [0, 1]$. 

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\lambda$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.995</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.912</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.520</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.111</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 9.4: Re-normalized eigenvector to enable the comparison with Fig. 9.1.
Chapter 10

Wiener filter

Consider a telecommunication signal \( S(t) \) which is generated by the transmitter. It is this information bearing signal that is transmitted over the communication channel. This channel will introduce additive noise \( N(t) \). The received signal at the end of the channel is represented by \( X(t) \). The communication channel is therefore regarded as a linear system with an additive noise source. The aim of the Wiener filter is to retrieve the original signal \( S(t) \).

The Wiener filter is a linear filter whose characteristics are determined using a statistical approach based on the knowledge of the spectral properties of the original signal \( S(t) \) and the noise \( N(t) \). This is done by determining the filter (= the Wiener filter) which minimizes the mean squared error between the original signal and its reconstructed version.

**Definition 50 (Wiener filter).** A Wiener filter is a linear time-invariant filter that

1. takes the signal \( X(t) = S(t) + N(t) \) as input with \( S(t) \) the signal and \( N(t) \) the additive noise. Both are wide-sense stationary processes with known first and second order moments (e.g. auto- and cross-correlation);
2. has a linear impulse response \( h(t) \) which results into a process \( Y(t) = h(t) \ast X(t) \) where \( \ast \) represents the convolution product;
3. determines \( h(t) \) such that \( Y(t) \) approaches the original signal \( S(t) \) as closely as possible in minimum mean-square sense for all time instances \( t \).

The Wiener filter basically looking for the continuous function \( h(t) \) which minimizes a mean squared error

\[
MSE(h(t)) = E \left( |S(t) - h(t) \ast X(t)|^2 \right).
\]

This is once more the determination of the extrema of functionals, i.e. "functions of functions", demanding the calculus of variations as introduced in Section 3.2.1. The results of that section will be used to show a general orthogonality principle for minimum least-square problems, followed by the general result to compute a Wiener filter and by its application for the (important) case when the signal \( S(t) \) and the additive noise \( N(t) \) are uncorrelated.
10.1 Orthogonality principle

**Theorem 42** (Orthogonality principle). Consider two zero-mean WSS processes \( X(t) \) and \( Y(t) \) \((\mu_X = \mu_Y = 0)\). Consider furthermore the mean squared error \((MSE)\) between \(Y(t)\) and a filtered version of \(X(t)\) using a linear filter with impulse response \(h(t)\)

\[
MSE(h(t)) = E \left( |Y(t) - h(t) * X(t)|^2 \right)
\]

The orthogonality principle then states that the minimizer of the mean squared error represented by \(\hat{h}(t) = \min_{h(t)} MSE(h(t))\) satisfies

\[
E \left( (Y(t) - \hat{h}(t) * X(t)) X^*(t - \tau) \right) = 0
\]

for all \(\tau\). This is equivalent by demanding that the residuals \(Y(t) - \hat{h}(t) * X(t)\) and the signal \(X(t)\) are uncorrelated and hence orthogonal since both are zero mean.

**Proof.** Consider that \(\hat{h}(t)\) minimizes the mean squared error. Hence, any slight perturbation must increase mean squared error. Let

\[
h_\varepsilon(t) = \hat{h}(t) + \varepsilon \eta(t)
\]

be the result of such a perturbation \(\varepsilon \eta(t)\) of \(\hat{h}(t)\), where \(\varepsilon\) is real valued and \(\eta(t)\) an arbitrary, deterministic function of time.

The mean squared error then equals

\[
MSE(h_\varepsilon(t)) = E \left( |Y(t) - (\hat{h}(t) + \varepsilon \eta(t)) * X(t)|^2 \right)
\]

\[
= E \left( |Y(t) - \hat{h}(t) * X(t)|^2 \right)
\]

\[
- \varepsilon E \left( (\eta(t) * X(t)) (Y(t) - \hat{h}(t) * X(t))^* \right)
\]

\[
- \varepsilon E \left( (Y(t) - \hat{h}(t) * X(t)) (\eta(t) * X(t))^* \right)
\]

\[
+ \varepsilon^2 E \left( |\eta(t) * X(t)|^2 \right)
\]

It was assumed that \(\hat{h}(t)\) minimizes the mean squared error, implying that \(MSE(h_\varepsilon(t))\) is minimal in \(\varepsilon = 0\)

\[
\frac{\partial MSE(h_\varepsilon(t))}{\partial \varepsilon} \bigg|_{\varepsilon=0} = 0.
\]

This is equivalent to demanding that

\[
E \left( (Y(t) - \hat{h}(t) * X(t)) (\eta(t) * X(t))^* \right) = 0.
\]
The function $\eta(t)$ can be put outside the expected value as it is a deterministic function

$$E \left( (Y(t) - \hat{h}(t) * X(t)) (\eta(t) * X(t))^* \right)$$

$$= E \left( (Y(t) - \hat{h}(t) * X(t)) \left( \int_{-\infty}^{\infty} \eta^*(\tau) X^*(t - \tau) d\tau \right) \right)$$

$$= \int_{-\infty}^{\infty} \eta^*(\tau) E \left( (Y(t) - \hat{h}(t) * X(t)) X^*(t - \tau) \right) d\tau$$

As this relationship must be equal to zeros for all possible functions $\eta(t)$, the orthogonality principle demands that $\hat{h}(t)$ must fulfill

$$E \left( (Y(t) - \hat{h}(t) * X(t)) X^*(t - \tau) \right) = 0 \quad \forall \tau.$$ 

\[\square\]

### 10.2 Wiener filter

The Wiener filter (Definition 50) is looking for the filter with impulse response $h(t)$ which minimizes the mean squared error between the original signal $S(t)$ and the output of the filter $h(t) * (S(t) + N(t))$ for all time instances $t$. Hence, the orthogonality principle of the mean squared error can be used (Theorem 42).

**Theorem 43** (Wiener-Hopf integral equation). The Wiener filter (Definition 50) satisfies

$$\rho_{SX}(\tau) = h(\tau) * \rho_{XX}(\tau)$$

$$= \int_{-\infty}^{\infty} h(t) \rho_{XX}(\tau - t) dt \quad \forall \tau \quad (10.1)$$

which is also known as the Wiener-Hopf integral equation. This equation can easily be solved in the frequency domain using

$$\Sigma_{SX}(f) = H(f) \Sigma_{XX}(f) \quad (10.2)$$

with $\Sigma_{SX}(f) = \mathcal{F}(\rho_{SX}(t))$, $\Sigma_{XX}(f) = \mathcal{F}(\rho_{XX}(t))$, and $H(f) = \mathcal{F}(h(t))$.

*Proof.* To use the orthogonality principle of Theorem 42, consider

$$X(t) = S(t) + N(t)$$

and that the Wiener filter $h(t)$ minimizes the mean squared error between $X(t)$ and $S(t)$

$$MSE(h(t)) = E \left( |S(t) - h(t) * X(t)|^2 \right).$$
The orthogonality principle (Theorem 42) then implies,

\[
\begin{align*}
E ( (S(t) - h(t) \ast X(t)) X^\ast(t - \tau)) \\
= E (S(t)X^\ast(t - \tau)) - h(t) \ast E (X(t)X^\ast(t - \tau)) \\
= \rho_{SX}(\tau) - h(\tau) \ast \rho_{XX}(\tau) = 0 \quad \forall \tau
\end{align*}
\]

which finally results into

\[
\rho_{SX}(\tau) = h(\tau) \ast \rho_{XX}(\tau).
\]

Using Wiener-Khintchine (Definition 42) enables the transformation to the frequency domain

\[
\mathcal{F} (\rho_{SX}(\tau)) = \mathcal{F} (h(\tau) \ast \rho_{XX}(\tau))
\]

and the conversion of the convolution product into a simple product

\[
\Sigma_{SX}(f) = H(f) \Sigma_{XX}(f).
\]

The above theorem shows that the Wiener filter can easily be determined in the frequency domain using the auto- and the cross-power spectra

\[
H(f) = \frac{\Sigma_{SX}(f)}{\Sigma_{XX}(f)}.
\]

The impulse response can be determined using the inverse Fourier transform

\[
h(t) = \mathcal{F}^{-1}(H(f)).
\]

### 10.3 Wiener filter for uncorrelated signal and noise

A important case is when the signal \(S(t)\) and the additive noise \(N(t)\) are uncorrelated. Using \(X(t) = S(t) + N(t)\) and \(\rho_{SN}(\tau) = 0\) for all \(\tau\) (or equivalently \(\Sigma_{SN}(f) = 0\) for all frequencies \(f\)) implies that

\[
\begin{align*}
\rho_{XX}(\tau) & = \rho_{SS}(\tau) + \rho_{NN}(\tau) \\
\rho_{SX}(\tau) & = \rho_{SS}(\tau)
\end{align*}
\]

or in the frequency domain

\[
\begin{align*}
\Sigma_{XX}(f) & = \Sigma_{SS}(f) + \Sigma_{NN}(f) \\
\Sigma_{SX}(f) & = \Sigma_{SS}(f).
\end{align*}
\]

This leads directly to the result that the Wiener filter for uncorrelated signal and noise equals

\[
H(f) = \frac{\Sigma_{SX}(f)}{\Sigma_{XX}(f)} = \frac{\Sigma_{SS}(f)}{\Sigma_{SS}(f) + \Sigma_{NN}(f)}.
\]
Fig. 10.1 shows the power spectra and the corresponding Wiener filters for three cases:

- **Top**: The power spectra of the signal and the noise are situated in different frequency bands. Hence, the Wiener filter equals one in the signal band and zero in the noise band.

- **Middle**: The power spectra share a common frequency band. The signal in the noisy band is suppressed only partially.

- **Bottom**: The noise power spectrum is flat (white noise), resulting in a Wiener filter whose frequency response resembles (but is not equal to) the signal power spectrum.
Chapter 11

Matched filter

The matched filter problem is part of the detection theory that aims to detect a (known) signal $s(t)$ that is disturbed by additive noise. It is assumed that $s(t)$ is known for all $t$ and hence deterministic.

Definition 51 (Matched filter). The matched filter is the optimal linear time-invariant filter with impulse response $h(t)$ that maximizes the Signal-to-Noise Ratio (SNR) of

$$Y(t) = h(t) \ast X(t)$$

at a predefined time instance $t_0$ under the assumption that the observed signal $X(t)$ equals a known signal $s(t)$ which is corrupted by additive noise $N(t)$

$$X(t) = s(t) + N(t).$$

11.1 Derivation of matched filter

To determine the expression for the matched filter, we will assume the signal $s(t)$ and that the noise process $N(t)$ is

- independent of $s(t)$,
- wide-sense stationary and ergodic,
- has a power spectrum of $\Sigma_{NN}(f)$.

To determine the matched filter, we will need the Cauchy-Schwarz inequality.

Lemma 22 (Cauchy-Schwarz inequality). The Cauchy-Schwarz inequality states that for all $u$ and $v$ of a space with inner product $\langle u, v \rangle$ that

$$|\langle u, v \rangle|^2 \leq |\langle u, u \rangle|^2 |\langle v, v \rangle|^2$$

where the equality holds if and only if $u$ and $v$ are linearly dependent.
For the inner product defined for complex-valued spectra \( U(f) \) and \( V(f) \) with \[
\langle U, V \rangle = \int_{-\infty}^{\infty} U(f)V(f) df
\]
the Cauchy-Schwarz inequality becomes
\[
\left| \int_{-\infty}^{\infty} U(f)V(f) df \right|^2 \leq \left( \int_{-\infty}^{\infty} |U(f)|^2 df \right) \left( \int_{-\infty}^{\infty} |V(f)|^2 df \right)
\]
with the equality if and only if \( U(f) = kV^*(f) \) with \( k \) a constant. Note that \( U(f) \) and \( V(f) \) must be square-integrable functions to guarantee the existence of the integrals.

**Theorem 44** (Matched filter theorem). The Matched filter with the frequency response
\[
H(f) = k \frac{S^*(f)e^{-j2\pi ft_0}}{\Sigma_{NN}(f)}
\]
maximizes the Signal-to-Noise Ratio at \( t_0 \) with
\[
SNR(t_0) = \int_{-\infty}^{\infty} \frac{|S(f)|^2}{\Sigma_{NN}(f)} df.
\]

**Proof.** Since the matched filter is a linear time-invariant system with impulse response \( h(t) \), it is possible to determine the filter’s output for its two input components, \( s(t) \) and \( N(t) \), namely
\[
Y(t_0) = y_s(t_0) + Y_n(t_0).
\]
The (deterministic) signal output component equals
\[
y_s(t_0) = \int_{-\infty}^{\infty} h(t)s(t_0 - t) dt
\]
and the (stochastic) noise output component becomes
\[
Y_n(t_0) = \int_{-\infty}^{\infty} h(t)N(t_0 - t) dt.
\]
The goal is now to maximize the signal-to-noise (SNR) at time instance \( t_0 \)
\[
SNR(t_0) = \frac{E\left(y_s^2(t_0)\right)}{E\left(y_n^2(t_0)\right)}.
\]
The component \( y_s(t_0) \) is deterministic and hence only depend on the choice of the time instance \( t_0 \). This implies that
\[
SNR(t_0) = \frac{y_s^2(t_0)}{E\left(y_n^2(t_0)\right)}.
\]
CHAPTER 11. MATCHED FILTER

This is the ratio of the power of the signal at time instance \( t = t_0 \) to the average power of the noise at the same time instance. This SNR ratio must be maximized when considering all possible impulse responses functions \( h(t) \).

This SNR at time instance \( t_0 \) can also be written in the frequency domain using the power of the signal component at \( t_0 \)

\[
y_s^2(t_0) = |h(t) \ast s(t)|^2_{t=t_0} = \left| \mathcal{F}^{-1}(\mathcal{F}(h(t))\mathcal{F}(s(t))) \right|^2_{t=t_0} = \left| \int_{-\infty}^{\infty} H(f)S(f)e^{j2\pi ft_0} df \right|^2
\]

with \( H(f) = \mathcal{F}(h(t)) \) and \( S(f) = \mathcal{F}(s(t)) \).

The power of the noise component \( E\left(Y_n^2(t_0)\right) \) has a constant value, independent of \( t_0 \), due to the wide-sense stationarity of the noise and the fact that the filter is linear time-invariant (Theorem 17). Furthermore, it equals the auto-correlation evaluated in \( \tau = 0 \)

\[
E\left(Y_n^2\right) = \rho_{Y_nY_n}(0).
\]

This can be expressed in the frequency domain as

\[
E\left(Y_n^2\right) = \mathcal{F}^{-1}(\Sigma_{Y_nY_n}(f))|_{\tau=0} = \left( \int_{-\infty}^{\infty} \Sigma_{Y_nY_n}(f)e^{j2\pi f\tau} df \right)|_{\tau=0}^{(a)} = \int_{-\infty}^{\infty} |H(f)|^2 \Sigma_{NN}(f) df
\]

where (a) uses Theorem 32 stating that the noise spectrum after filtering equals

\[
\Sigma_{Y_nY_n}(f) = |H(f)|^2 \Sigma_{NN}(f).
\]

The goal is now to determine the LTI filter \( H(f) \) such that \( SNR(t_0) \) is maximized with

\[
SNR(t_0) = \frac{\left| \int_{-\infty}^{\infty} H(f)S(f)e^{j2\pi ft_0} df \right|^2}{\int_{-\infty}^{\infty} |H(f)|^2 \Sigma_{NN}(f) df} = \frac{1}{\int_{-\infty}^{\infty} \left| \frac{\mathcal{F}(U(f))}{\mathcal{F}(V(f))} \right|^2 df} \leq \int_{-\infty}^{\infty} \left| \frac{\mathcal{F}(U(f))}{\mathcal{F}(V(f))} \right|^2 df
\]

where the left hand side is maximized if \( U(f) = kV^*(f) \). Applying this interpretation to (11.2) makes it possible to determine the \( H(f) \) that maximizes the
\( SNR(t_0) \). Therefore consider
\[
U(f) = H(f) \sqrt{\Sigma_{NN}(f)} \\
V(f) = \frac{S(f)e^{j2\pi f t_0}}{\sqrt{\Sigma_{NN}(f)}}.
\]

Applying the Cauchy-Schwarz inequality (Lemma (22)) implies that
\[
SNR(t_0) = \left| \int_{-\infty}^{\infty} H(f) S(f) e^{j2\pi f t_0} df \right|^2 \leq \int_{-\infty}^{\infty} |S(f)|^2 df \leq \int_{-\infty}^{\infty} |H(f)|^2 \Sigma_{NN}(f) df \tag{11.3}
\]

since \( |e^{j2\pi f t_0}| = 1 \). The equality holds (and hence the \( SNR(t_0) \) is maximal) if and only if
\[
H(f) \sqrt{\Sigma_{NN}(f)} = k S^*(f) e^{-j2\pi f t_0} \sqrt{\Sigma_{NN}(f)}.
\]

If the additive noise source is white, then the power spectrum is constant \( \Sigma_{NN}(f) = N_0/2 \) for all \( f \).

**Theorem 45** (Matched filter theorem for white noise source). If the additive noise is white, then the matched filter has a frequency response
\[
H(f) = k' S^*(f) e^{-j2\pi f t_0}
\]
with impulse response
\[
h(t) = k' s^*(t_0 - t).
\]

and with \( k' = 2k/N_0 \) which maximizes the Signal-to-Noise Ratio at \( t_0 \) with
\[
SNR(t_0) = \frac{2}{N_0} \int_{-\infty}^{\infty} |S(f)|^2 df.
\]

**Proof.** The impulse response of the matched filter equals
\[
h(t) = k' \mathcal{F}^{-1} \left( S^*(f) e^{-j2\pi f t_0} \right) \\
= k' \int_{-\infty}^{\infty} S^*(f) e^{-j2\pi f t_0} e^{j2\pi f t} df \\
= k' \left( \int_{-\infty}^{\infty} S(f) e^{j2\pi f(t_0 - t)} df \right)^* \\
= k' s^*(t_0 - t).
\]

If \( k' \) is chosen equal to one, then \( h(t) = s^*(t_0 - t) \) maximizes the SNR at \( t_0 \). This expression can be simplified for real signals to \( h(t) = k's(t_0 - t) \). \( \square \)
CHAPTER 11. MATCHED FILTER

The matched filter can also be interpreted using a cross-correlation. When considering the convolution of the filter

\[ y(t_0) = \int_{-\infty}^{\infty} h(t)x(t_0 - t)dt \]

and the impulse response of the matched filter \( h(t) = s^*(t_0 - t) \), one obtains that

\[ y(t_0) = k' \int_{-\infty}^{\infty} s^*(t_0 - t)x(t_0 - t)dt \]

\[ = k' \int_{-\infty}^{\infty} x(t_0 + t')s^*(t_0 + t')dt'. \]

These equations are similar to the cross-correlation \( \rho_{XS}(t_0, t_0) \). Hence, the output of the matched filter equals the cross-correlation of the signal \( x(t_0 + t) \) with the signal that is to be detected \( s(t_0 + t) \).

11.2 Examples of matched filters

This section provides a wide range of examples of the matched filter problem including

- pulse detection in radar and sonar applications,
- positioning detection in GPS (Global Positioning System),
- time synchronization when receiving a package of digital signals,
- integrate-and-dump circuits within digital demodulation,
- root-raised cosine filters in digital communication.

11.2.1 Radar / Sonar

Radar (RAdio Detection And Ranging) / Sonar (SOund Navigation And Rang-ing) transmit electromagnetic / acoustic pulses towards a target (e.g. an airplane) and these pulses are (partially) reflected by the target. The aim of the radar is to detect and range the target.

Maximizing the signal-to-noise ratio for a radar application is a matched filter problem as it can be assumed that the shape of the received pulse is known (e.g. an attenuated and delayed version of the transmitted pulse).

A graphical representation is shown in Fig. 11.1 where \( s(t) \) represents the transmitted pulse and \( h(t) \) represents the matched filter impulse response.
Figure 11.1: Simulated radar example where \( s(t) \) represents the transmitted pulse and \( h(t) = s(-t) \) the matched filter impulse response.

Fig. 11.2 shows a realization of the noisy received signal \( X(t) \) with an echo at \( t_0 = 0.5 \), which is a scaled version of \( s(t - t_0) \), perturbed with additive white noise. The bottom plot shows \( y(t) \) which represents the output of the matched filter. Note the peak of \( Y(t_0) \) in \( t = t_0 \), indicating the position of the target.
Figure 11.2: Simulated radar example where \( s(t) \) represents the transmitted pulse, \( X(t) \) the noisy received signal which is a scaled and delayed version of \( s(t) \) (with delay equal to \( t_0 = 0.5 \)), perturbed with additive noise. The output of the matched filter is represented by \( Y(t) \) and has a maximum at the position of the target \( t = t_0 \).

11.2.2 GPS

The radar example shows how it is possible to detect the position of a pulse from noisy measurements and how it is possible to maximize its SNR. However, in addition to the SNR, it is also important to separate various individual pulses. It is therefore necessary to use reference signals \( s(t) \) with an auto-correlation that resembles a Dirac function.

GPS (Global Positioning System) uses pseudo-random binary sequences (PRBS) which are known to the receiver. The time of arrival can be determined by time-aligning a known sequence with the measured code sequence. A graphical
representation is shown in Fig. 11.3 where \( s(t) \) represents a maximum length PRBS, and \( X(t) \) the received signal which is a scaled and time shifted version of \( s(t) \) with \( t_0 = 0.5 \), perturbed with additive noise. \( Y(t) \) represents the output of the matched filter and has a peak at \( Y(t_0) \). This delay can then be used to triangulate the position within the GPS application.

![Graphs showing the transmitted, received, and matched filter output signals.]

Figure 11.3: Simulated GPS example where \( s(t) \) represents the transmitted PRBS, \( X(t) \) the noisy received signal for a time shift of \( t_0 = 0.5 \), and \( Y(t) \) the output of the matched filter.

### 11.2.3 Timing synchronization in digital communication

Digital receivers using e.g. QAM-modulation need to determine the start of a data frame and recover the sampling clock for receiving the digital data. This can be done using a preamble which is defined by a fixed preamble \( s(t) \) which is transmitted prior to sending the actual data \( D(t) \). Hence, the transmitted
data frame is a concatenation of the (predefined) preamble $s(t)$ and the data $D(t)$. Fig. 11.4 shows the received signal $X(t)$ after scaling and delaying the transmitted data frame, and after adding additive noise. The output of the matched filter $Y(t)$ with impulse response $h(t) = s(-t)$, with $s(t)$ the preamble, clearly shows a peak at the beginning of the data frame. The time instance of this peak can be used to synchronize both the beginning of the frame, and the data clock used to sample the received data (see the eye-diagrams in the next sections).

Figure 11.4: Timing synchronization in digital communication where $s(t)$ represents the preamble PRBS. The signal $X(t)$ is the (noisy) received signal which is the concatenation of the PAM-2 modulated preamble followed by a PAM-4 (random) information data. The lower figure shows the output of the matched filter.
Consider a PAM-M modulated signal which is transmitted over a channel whose bandwidth is larger than the bandwidth of the signal (e.g., PAM modulation over an optical fiber). In this case, an integrate-and-dump receiver will lead to the maximum SNR.

Under the above assumptions, it is possible to represent transmitted signal as a Dirac-comb of the data

\[ D(t) = \sum_{n=-\infty}^{\infty} d_n \delta(t - nT_s) \]

which is filtered afterwards using a zero-order-hold impulse response

\[ h_{\text{ZOH}}(t) = \begin{cases} 
1 & 0 \leq t < T_s \\
0 & \text{elsewhere,} 
\end{cases} \]

generating

\[ s(t) = h_{\text{ZOH}}(t) \ast D(t). \]

(see also Section 8.6.4).

The matched filter theorem states the signal-to-noise ratio is maximized when using a matched filter which has an impulse response

\[ h(t) = h_{\text{ZOH}}(-t) \]

\[ = \begin{cases} 
1 & -T_s < t \leq 0 \\
0 & \text{elsewhere.} 
\end{cases} \]

The output of the matched filter then equals

\[ Y(t_0) = \int_{-\infty}^{\infty} h(t)[y_s(t_0 - t) + \frac{Y_n(t_0 - t)}{\sqrt{T_s}}]dt \]

\[ = \int_{-T_s}^{0} [y_s(t_0 - t) + \frac{Y_n(t_0 - t)}{\sqrt{T_s}}]dt \]

\[ = \int_{0}^{T_s} [y_s(t_0 + t') + \frac{Y_n(t_0 + t')}{\sqrt{T_s}}]dt'. \]

This corresponds to integrating over a time window \( T_s \) and then dumping the past before computing the response of the next sample.

Fig. 11.5 shows an example of the various signals along the chain. Fig. 11.6 and 11.7 show the eye diagram of the same signals. In the noiseless case, the signal goes from one state to the other in a linear way. The noisy case shows that the eye-diagram is open (highest SNR to sample) at \( t = nT_s \).
Figure 11.5: Example of the signals for an integrate-and-dump receiver. The top figure shows the data representation as a Dirac-comb and the zero-order-hold reconstructed data $s(t)$ by filtering of the Dirac-comb using $h_{ZOH}(t)$. The second plot shows the input of the matched filter $X(t)$ which equals $s(t) + N(t)$. This signal is processed by the matched filter, resulting in $y_s(t)$ for the noiseless input $s(t)$ and $Y(t)$ for the noisy input $X(t)$. 
Figure 11.6: The eye diagram of an integrate-and-dump. Note that the signals move from on state to another in a linear way.

Figure 11.7: The eye diagram of a noisy integrate-and-dump receiver. Notice that the eye gets closed and that the SNR is maximal in $t = nT_s$. 
11.2.5 Root-raised-cosine filter as matched filter

Digital communication systems often split the wanted overall transfer function into two parts, one realized by the transmitter and one by the receiver. Additionally, it is preferred that the overall transfer function satisfies the Nyquist ISI criterion, and hence the choice to realize a raised-cosine characteristic for the overall transfer function. The question to be answered is: how to divide the raised-cosine characteristic over the transmitter and the receiver to obtain the best signal-to-noise ratio?

Consider that the transmitter is sending the data by passing the data stream, namely impulses with value $d_n$ through the transmit filter $h_{TX}(t)$ generating

$$s(t) = h_{TX}(t) * D(t)$$

with

$$D(t) = \sum_{n=-\infty}^{\infty} d_n \delta(t - nT_s).$$

The transmitted data is send over the channel and corrupted with additive white noise $n(t)$. Hence, the received signal equals $X(t) = s(t) + n(t)$. The matched filter learns us that the signal-to-noise ratio at the moment of detection, i.e. $t_0 = nT_s$, can be maximized using the matched filter at the receiver side, namely $h_{RX}(t) = h_{TX}^*(-t)$. Hence, the overall transfer function in the frequency domain becomes

$$H_{RX}(f)H_{TX}(f) = |H_{RX}(f)|^2$$

which must be equal to $H_{RC}(f)$ to avoid ISI. As $H_{RC}(f)$ is real-valued, it can be shown that using the root-raised-cosine filter for both the transmitter and the receiver

$$H_{TX}(f) = H_{RX}(f) = H_{RRC}(f) = \sqrt{H_{RC}(f)}$$

maximizes the signal-to-noise ratio. Hence,

$$h_{TX}(t) = h_{RX}(t) = h_{RRC}(t).$$

Fig. 11.8 shows a binary data sequence that is filtered by the root-raised-cosine filter of the transmitter ($s(t) = h_{TX}(t) * D(t)$) and then perturbed with additive noise. After filtering with the second root-raised-cosine filter at the receiver side, one can see that sampling the received signal at the correct time instances leads to the reconstruction of the data without any inter-symbol interference (see $y_s(t)$). When considering the noisy received signal, one observes that the error-less samples of $y_s(t)$ are perturbed by additive noise. Fig. 11.9 shows the probability density function of the sampled $Y(t)$, clearly indicating the two Gaussian distributions that correspond with the transmission of the binary '0' or '1'. Fig. 11.10 and 11.11 show the eye diagrams of the received signals, clearly showing that there is no ISI in the noiseless case (Fig. 11.10) and that the signal-to-noise ratio is maximized at the sampling instances (Fig. 11.11).
Figure 11.8: Example of the signals when using root-raised-cosine filters at both the transmitter and the receiver. The top figure shows the data representation as a Dirac-comb and the root-raised cosine reconstructed data $s(t)$ by filtering of the Dirac-comb using $h_{TX}(t) = h_{RRC}(t)$. The second plot shows input of the matched filter $X(t)$ which equals $s(t) + N(t)$. This signal is then processed by the matched filter, resulting in $y_s(t)$ for the noiseless input $s(t)$ and $Y(t)$ for the noisy input $X(t)$. 

```plaintext
\begin{align*}
S(t) & = \text{Dirac-comb representation of the data} \\
X(t) & = s(t) + N(t) \\
Y_s(t) & = \text{Matched filter output for noiseless input} \\
Y(t) & = \text{Matched filter output for noisy input}
\end{align*}
```
Figure 11.9: The pdf of the sampled data of $Y(t)$ of Fig. 11.8.

Figure 11.10: The eye diagram of a noiseless raised-cosine filtered signal. Note the absence of the inter-symbol-interference.
Figure 11.11: The eye diagram of a noisy root-raised-cosine based transmitter and receiver. Note that the SNR is maximal at $t = nT_s$. 
Bibliography


