The book is a result of author’s 33 years of experience in teaching and research in signal processing.

This book will guide you from a review of continuous-time signals and systems, through the world of digital signal processing, up to some of the most advanced theory and techniques in adaptive systems, time-frequency analysis and sparse signal processing.

It provides simple examples and explanations for each, including the most complex transform, method, algorithm or approach presented in the book. The most sophisticated results in signal processing theory are illustrated on simple numerical examples.

The book is written for students learning digital signal processing and for engineers and researchers refreshing their knowledge in this area. The selected topics are intended for advanced courses and for preparing the reader to solve problems in some of the state of art areas in signal processing.

Ljubisa Stankovic is a professor at the University of Montenegro, IEEE Fellow for contributions to the Time-Frequency Signal Analysis, a member of the Montenegrin and European Academy of Sciences and Arts. He has been an Associate Editor of several world-leading journals in Signal Processing.
DIGITAL SIGNAL PROCESSING

with selected topics

ADAPTIVE SYSTEMS
TIME-FREQUENCY ANALYSIS
SPARSE SIGNAL PROCESSING

Ljubiša Stanković

2015
To
my parents
Božo and Cana,

my wife Snežana,
and our
Irena, Isidora, and Nikola.
## Contents

### I  Review

**Chapter 1  Continuous-Time Signals and Systems**

1.1 Continuous-Time Signals
1.2 Periodic Signals and Fourier Series
    - 1.2.1 Fourier Series of Real-Valued Signals
    - 1.2.2 Linear Systems
1.3 Fourier Transform
    - 1.3.1 Fourier Transform and Linear Time-Invariant Systems
    - 1.3.2 Properties of the Fourier Transform
    - 1.3.3 Relationship Between the Fourier Series and the Fourier Transform
1.4 Fourier Transform and Stationary Phase Method
1.5 Laplace Transform
    - 1.5.1 Linear Systems Described by Differential Equations
    - 1.5.2 Table of the Laplace Transform
1.6 Butterworth Filter

**Chapter 2  Discrete-Time Signals and Transforms**

2.1 Discrete-Time Signals
2.2 Fourier Transform of Discrete-Time Signals
    - 2.2.1 Properties
    - 2.2.2 Spectral Energy and Power Density

### II  Discrete Signals and Systems
## Contents

2.3 Sampling Theorem in the Time Domain 69  
2.4 Problems 74  
2.5 Solutions 79  
2.6 Exercise 99  

Chapter 3 Discrete Fourier Transform 101  
3.1 DFT Definition 101  
3.2 DFT Properties 107  
3.3 Zero-Padding and Interpolation 112  
3.4 Relation among the Fourier Representations 116  
3.5 Fast Fourier Transform 118  
3.6 Sampling of Periodic Signals 124  
3.7 Analysis of a Sinusoid by Using the DFT 127  
3.7.1 Leakage Effect 128  
3.7.2 Displacement 131  
3.8 Discrete Cosine and Sine Transforms 134  
3.9 Discrete Walsh-Hadamard and Haar Transforms 139  
3.9.1 Discrete Walsh-Hadamard Transform 142  
3.9.2 Discrete Haar Wavelet Transform 146  
3.10 Problems 150  
3.11 Solutions 152  
3.12 Exercise 161  

Chapter 4 $z$-Transform 163  
4.1 Definition of the $z$-transform 163  
4.2 Properties of the $z$-transform 165  
4.2.1 Linearity 165  
4.2.2 Time-Shift 165  
4.2.3 Multiplication by exponential signal: Modulation 166  
4.2.4 Differentiation 166  
4.2.5 Convolution in time 167  
4.2.6 Table of the $z$-transform 167  
4.2.7 Initial and Stationary State Signal Value 168  
4.3 Inverse $z$-transform 168  
4.3.1 Direct Power Series Expansion 168  
4.3.2 Theorem of Residues Based Inversion 172  
4.4 Discrete systems and the $z$-transform 174  
4.5 Difference equations 177  
4.5.1 Solution Based on the $z$-transform 177  
4.5.2 Solution of Difference Equations in the Time Domain 180  
4.6 Relation of the $z$-transform to other Transforms 185
4.7 Problems 187
4.8 Solutions 191
4.9 Exercise 207

Chapter 5 From Continuous to Discrete Systems 211
5.1 Impulse Invariance Method 212
5.2 Matched z-transform method 217
5.3 Differentiation and Integration 220
5.4 Bilinear Transform 224
5.5 Discrete Filters Design 230
  5.5.1 Lowpass filters 230
  5.5.2 Highpass Filters 236
  5.5.3 Bandpass Filters 238
  5.5.4 Allpass Systems - System Stabilization 239
  5.5.5 Inverse and Minimum Phase Systems 241
5.6 Problems 246
5.7 Solutions 249
5.8 Exercise 258

Chapter 6 Realization of Discrete Systems 261
6.1 Realization of IIR systems 261
  6.1.1 Direct realization I 262
  6.1.2 Direct realization II 262
  6.1.3 Sensitivity of the System Poles/Zeros to Errors in Coefficients 265
  6.1.4 Cascade Realization 270
  6.1.5 Parallel realization 274
  6.1.6 Inverse realization 277
6.2 FIR Systems and their Realizations 278
  6.2.1 Linear Phase Systems and Group Delay 279
  6.2.2 Windows 281
  6.2.3 Design of FIR System in the Frequency Domain 285
  6.2.4 Realizations of FIR system 287
6.3 Problems 292
6.4 Solutions 296
6.5 Exercise 308

Chapter 7 Discrete-Time Random Signals 313
7.1 Basic Statistical Definitions 313
  7.1.1 Expected Value 313
  7.1.2 Probability and Probability Density Function 320
  7.1.3 Median 323
Contents

7.1.4 Variance 325
7.2 Second-Order Statistics 330
7.2.1 Correlation and Covariance 330
7.2.2 Stationarity and Ergodicity 331
7.2.3 Power Spectral Density 332
7.3 Noise 334
7.3.1 Uniform Noise 334
7.3.2 Binary Noise 335
7.3.3 Gaussian Noise 338
7.3.4 Complex Gaussian Noise and Rayleigh Distribution 343
7.3.5 Impulsive Noises 344
7.3.6 Noisy Signals 346
7.4 Discrete Fourier Transform of Noisy Signals 346
7.4.1 Detection of a Sinusoidal Signal Frequency 350
7.5 Linear Systems and Random Signals 354
7.5.1 Spectral Estimation of Narrowband Signals 360
7.6 Detection and Matched Filter 362
7.6.1 Matched Filter 363
7.7 Optimal Wiener Filter 366
7.8 Quantization effects 370
7.8.1 Input signal quantization 371
7.8.2 Quantization of the results 376
7.9 Problems 388
7.10 Solutions 394
7.11 Exercise 412

III Selected Topics 415

Chapter 8 Adaptive Systems 417
8.1 Introduction 417
8.2 Linear Adaptive Adder 421
8.2.1 Error Signal 423
8.2.2 Autocorrelation Matrix Eigenvalues and Eigenvectors 432
8.2.3 Error Signal Analysis 437
8.2.4 Orthogonality Principle 439
8.3 Steepest Descend Method 440
8.4 LMS Algorithm 451
8.4.1 Convergence of the LMS algorithm 452
8.5 LMS Application Examples 454
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.5.1</td>
<td>Identification of Unknown System</td>
<td>454</td>
<td></td>
</tr>
<tr>
<td>8.5.2</td>
<td>Noise Cancellation</td>
<td>458</td>
<td></td>
</tr>
<tr>
<td>8.5.3</td>
<td>Sinusoidal Disturbance Cancellation</td>
<td>461</td>
<td></td>
</tr>
<tr>
<td>8.5.4</td>
<td>Signal Prediction</td>
<td>462</td>
<td></td>
</tr>
<tr>
<td>8.5.5</td>
<td>Adaptive Antenna Arrays</td>
<td>467</td>
<td></td>
</tr>
<tr>
<td>8.5.6</td>
<td>Acoustic Echo Cancellation</td>
<td>473</td>
<td></td>
</tr>
<tr>
<td>8.6</td>
<td>Variations on the LMS Algorithm</td>
<td>475</td>
<td></td>
</tr>
<tr>
<td>8.6.1</td>
<td>Sign LMS</td>
<td>475</td>
<td></td>
</tr>
<tr>
<td>8.6.2</td>
<td>Block LMS</td>
<td>476</td>
<td></td>
</tr>
<tr>
<td>8.6.3</td>
<td>Normalized LMS Algorithm</td>
<td>477</td>
<td></td>
</tr>
<tr>
<td>8.6.4</td>
<td>LMS with Variable Step Size</td>
<td>479</td>
<td></td>
</tr>
<tr>
<td>8.6.5</td>
<td>Complex LMS</td>
<td>483</td>
<td></td>
</tr>
<tr>
<td>8.7</td>
<td>RLS Algorithm</td>
<td>483</td>
<td></td>
</tr>
<tr>
<td>8.8</td>
<td>Adaptive Recursive Systems</td>
<td>487</td>
<td></td>
</tr>
<tr>
<td>8.9</td>
<td>From the LMS algorithm to the Kalman filters</td>
<td>489</td>
<td></td>
</tr>
<tr>
<td>8.10</td>
<td>Neural Networks</td>
<td>495</td>
<td></td>
</tr>
<tr>
<td>8.10.1</td>
<td>Neuron</td>
<td>496</td>
<td></td>
</tr>
<tr>
<td>8.10.2</td>
<td>Network Function</td>
<td>496</td>
<td></td>
</tr>
<tr>
<td>8.10.3</td>
<td>Activation Function</td>
<td>497</td>
<td></td>
</tr>
<tr>
<td>8.10.4</td>
<td>Neural Network Topology</td>
<td>499</td>
<td></td>
</tr>
<tr>
<td>8.10.5</td>
<td>Network with Supervised Learning</td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>8.10.6</td>
<td>One-Layer Network with Binary Output - Perceptron</td>
<td>502</td>
<td></td>
</tr>
<tr>
<td>8.10.7</td>
<td>One-Layer Neural Network with Continuous Output</td>
<td>506</td>
<td></td>
</tr>
<tr>
<td>8.10.8</td>
<td>Multilayer Neural Networks</td>
<td>510</td>
<td></td>
</tr>
<tr>
<td>8.10.9</td>
<td>Neural Networks with Unsupervised Learning</td>
<td>513</td>
<td></td>
</tr>
<tr>
<td>8.10.10</td>
<td>Voting Machines</td>
<td>514</td>
<td></td>
</tr>
</tbody>
</table>

Chapter 9  Time-Frequency Analysis  515
| 9.1     | Short-Time Fourier Transform | 516 |
| 9.2     | Windows | 523 |
| 9.2.1   | Rectangular Window | 523 |
| 9.2.2   | Triangular (Bartlett) Window | 524 |
| 9.2.3   | Hann(ing) Window | 525 |
| 9.2.4   | Hamming Window | 527 |
| 9.2.5   | Blackman and Kaiser Windows | 528 |
| 9.2.6   | Discrete Form and Realizations of the STFT | 529 |
| 9.2.7   | Recursive STFT Realization | 535 |
| 9.2.8   | Filter Bank STFT Implementation | 536 |
| 9.2.9   | Signal Reconstruction from the Discrete STFT | 540 |
9.2.10 Time-Varying Windows 550
9.2.11 Frequency-Varying Window 561
9.2.12 Hybrid Time-Frequency-Varying Windows 563
9.3 Wavelet Transform 563
  9.3.1 Filter Bank and Discrete Wavelet 568
  9.3.2 S-Transform 601
9.4 Local Polynomial Fourier Transform 604
  9.4.1 Fractional Fourier Transform with Relation to
      the LPFT 607
9.5 High-Resolution STFT 608
  9.5.1 Capon’s STFT 608
  9.5.2 MUSIC STFT 612
  9.5.3 Capon’s LPFT 614
9.6 Wigner Distribution 616
  9.6.1 Auto-Terms and Cross-Terms in the Wigner
      Distribution 621
  9.6.2 Wigner Distribution Properties 625
  9.6.3 Pseudo and Smoothed Wigner Distribution 630
  9.6.4 Discrete Pseudo Wigner Distribution 633
  9.6.5 From the STFT to the Wigner Distribution via
      S-Method 641
9.7 General Quadratic Time-Frequency Distributions 647
  9.7.1 Reduced Interference Distributions 651
  9.7.2 Kernel Decomposition Method 656

Chapter 10 Sparse Signal Processing 659
10.1 Illustrative Examples 660
10.2 Sparsity and Reduced Set of Samples/Observations 670
10.3 Transformation Matrix Parameters 674
  10.3.1 Unitary Matrix 676
  10.3.2 Isometry and Restricted Isometry Property 677
  10.3.3 Coherence 678
  10.3.4 Restricted Isometry and Coherence 681
  10.3.5 Restricted Isometry and Eigenvalues 685
  10.3.6 Unique Reconstruction Condition and RIP 695
  10.3.7 Rank and Spark of a Matrix 698
  10.3.8 Spark and the Solution Uniqueness 700
10.4 Norm-Zero Based Reconstruction 704
  10.4.1 Direct Combinatorial Search 704
  10.4.2 Pseudoinverse matrix 707
  10.4.3 Estimation of Unknown Positions 708
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.4.4</td>
<td>Unavailable/Missing Samples Noise in Initial Estimation</td>
<td>711</td>
</tr>
<tr>
<td>10.4.5</td>
<td>Iterative Procedure</td>
<td>721</td>
</tr>
<tr>
<td>10.4.6</td>
<td>Influence of Additive Input Noise</td>
<td>723</td>
</tr>
<tr>
<td>10.4.7</td>
<td>Nonsparse Signal Reconstruction</td>
<td>727</td>
</tr>
<tr>
<td>10.5</td>
<td>Norm-One Based Reconstruction</td>
<td>729</td>
</tr>
<tr>
<td>10.5.1</td>
<td>Illustrations in the Signal Domain</td>
<td>731</td>
</tr>
<tr>
<td>10.5.2</td>
<td>Illustration in the Sparsity Domain</td>
<td>736</td>
</tr>
<tr>
<td>10.5.3</td>
<td>Equivalence of the Norm-Zero and Norm-One Based Minimization Solutions</td>
<td>750</td>
</tr>
<tr>
<td>10.6</td>
<td>Median Based Formulation</td>
<td>754</td>
</tr>
<tr>
<td>10.7</td>
<td>Norm-one Based Reconstruction Algorithms</td>
<td>760</td>
</tr>
<tr>
<td>10.7.1</td>
<td>LASSO- Minimization</td>
<td>761</td>
</tr>
<tr>
<td>10.7.2</td>
<td>Signal Domain Reconstruction with a Gradient Algorithm</td>
<td>764</td>
</tr>
<tr>
<td>10.8</td>
<td>On the Uniqueness of the DFT of Sparse Signals</td>
<td>776</td>
</tr>
<tr>
<td>10.9</td>
<td>Indirect Measurements/Sampling</td>
<td>785</td>
</tr>
<tr>
<td>10.10</td>
<td>Processing of Sparse Signals with Impulsive Noise</td>
<td>795</td>
</tr>
<tr>
<td>10.10.1</td>
<td>Direct Search Procedure</td>
<td>796</td>
</tr>
<tr>
<td>10.10.2</td>
<td>Criteria for Selecting Samples</td>
<td>797</td>
</tr>
<tr>
<td>10.10.3</td>
<td>Uniqueness of the Obtained Solution</td>
<td>800</td>
</tr>
<tr>
<td>10.11</td>
<td>Image Reconstruction</td>
<td>802</td>
</tr>
</tbody>
</table>

Index                                                                 805

About the Author                                                      810
Preface

This book is a result of author’s thirty-three years of experience in teaching and research in signal processing. It is written for students and engineers as a first book in digital signal processing, assuming that a reader is familiar with the basic mathematics, including integrals, differential calculus, and linear algebra. Although a review of continuous-time analysis is presented in the first chapter, a prerequisite for the presented content is a basic knowledge about continuous-time signal processing.

The book consists of three parts. After an introductory review part, the basic principles of digital signal processing are presented within Part two of the book. This part starts with Chapter two which deals with basic definitions, transforms, and properties of discrete-time signals. The sampling theorem, providing essential relation between continuous-time and discrete-time signals, is presented in this chapter as well. Discrete Fourier transform and its applications to signal processing are the topic of the third chapter. Other common discrete transforms, like Cosine, Sine, Walsh-Hadamard, and Haar are also presented in this chapter. The z-transform, as a powerful tool for analysis of discrete-time systems, is the topic of Chapter four. Various methods for transforming a continuous-time system into a corresponding discrete-time system are derived and illustrated in Chapter five. Chapter six is dedicated to the forms of discrete-time system realizations. Basic definitions and properties of random discrete-time signals are given in Chapter six. Systems to process random discrete-time signals are considered in this chapter as well. Chapter six concludes with a short study of quantization effects.

The presentation is supported by numerous illustrations and examples. Chapters within Part two are followed by a number of solved and unsolved problems for practice. Theory is explained in a simple way with a necessary mathematical rigor. The book provides simple examples and
explanations for each presented transform, method, algorithm or approach. Sophisticated results in signal processing theory are illustrated by simple numerical examples.

Part three of the book contains few selected topics in digital signal processing: adaptive discrete-time systems, time-frequency signal analysis, and processing of discrete-time sparse signals. This part could be studied within an advanced course in digital signal processing, following the basic course. Some parts from the selected topics may be included in tailoring a more extensive first course in digital signal processing as well.


The author thanks the colleagues that helped in preparing the special topics part of the book. Many thanks to Miloš Daković who coauthored all three chapters of Part three of this book and to other coauthors of chapters in this part: Thayaparan Thayananthan, Srdjan Stanković, and Irena Orović. Special thanks to M.Sc. Miloš Brajović and M.Sc. Stefan Vujović for their careful double-check of the presented theory and examples, numerous comments, and for the help in proofreading the final version of the book.

London,

Author
Introduction

Signal is a physical process, mathematical function, or any other physical or symbolic representation of an information. Signal theory and processing are the areas dealing with the efficient generation, description, transformation, transmission, reception, and interpretation of information. In the beginning, the most common physical processes used for these purposes were the electric signals, for example, varying current or electromagnetic waves. Signal theory is most commonly studied within electrical engineering. Signal theory theory are strongly related to the applied mathematics and information theory. Examples of signals include speech, music, image, video, medical, biological, geophysical, sonar, radar, biomedical, car engine, financial, and molecular data. In terms of signal generation, the main topics are in sensing, acquisition, synthesis, and reproduction of information. Various mathematical transforms, representations, and algorithms are used for describing signals. Signal transformations are a set of methods for decomposition, filtering, estimation and detection. Modulation, demodulation, detection, coding, and compression are the most important aspects of the signal transmission. In the process of interpretation, various approaches may be used, including adaptive and learning-based tools and analysis.

Mathematically, signals are presented by functions of one or more variables. Examples of one-dimensional signals are speech and music signals. A typical example of a two-dimensional signal is an image while video sequence is a sample of a three-dimensional signal. Some signals, for example, geophysical, medical, biological, radar, or sonar, may be represented and interpreted as one-dimensional, two-dimensional, or multidimensional.

Signals may be continuous functions of independent variables, for example, functions of time and/or space. Independent variables may also be discrete, with the signal values being defined only over an ordered set
of discrete independent variable values. This is a discrete-time signal. The discrete-time signals, after being stored in a general computer or special-purpose hardware, are discretized (quantized) in amplitude as well, so that they can be memorized within the registers of a finite length. These kinds of signals are referred to as digital signals, Fig. 1. A continuous-time and continuous amplitude (analog) signal is transformed into a discrete-time and discrete-amplitude (digital) signal by using analog-to-digital (A/D) converters, Fig. 2. Their processing is known as digital signal processing. In modern systems, the amplitude quantization errors are very small. Common A/D converters are with sampling frequency of up to megasample (some even up to few gigasample) per second with 8 to 24 bits of resolution in amplitude. The digital signals are usually mathematically treated as continuous (nondiscretized) in amplitude, while the quantization error is studied, if needed, as a small disturbance in processing, reduced to a noise in the input signal. Digital signals are transformed back into analog form by digital-to-analog (D/A) converters.

According to the nature of their behavior, all signals could be deterministic or stochastic. For deterministic signals, the values are known in the past and future, while the stochastic signals are described by probabilistic methods. The deterministic signals are commonly used for theoretical description, analysis, and syntheses of systems for signal processing.

Figure 1  Illustration of a continuous signal and its discrete-time and digital version.
Advantages of processing signals in digital form are in their flexibility and adaptability with possibilities ranging up to our imagination to implement a transformation with an algorithm on a computer. The time required for processing in real time (all calculations have to be completed between two signal samples) is a limitation as compared to the analog systems that are limited with a physical delay of electrical components and circuits only.
Part I

Review
Chapter 1

Continuous-Time Signals and Systems

Most of discrete-time signals are obtained by sampling continuous-time signals. In many applications, the result of signal processing is presented and interpreted in the continuous-time domain. Throughout the course of digital signal processing, the results will be discussed and related to the continuous-time forms of signals and their parameters. This is the reason why the first chapter is dedicated to a review of signals and transforms in the continuous-time domain. This review will be of help in establishing proper correspondence and notation for the presentation that follows in the next chapters.

1.1 CONTINUOUS-TIME SIGNALS

One-dimensional signals, represented by a function of time as a continuous independent variable, are referred to as continuous-time signals (continuous signals). Some simple forms of deterministic continuous-time signals are presented next.

The unit-step signal (Heaviside function) is defined by

\[
    u(t) = \begin{cases} 
    1, & \text{for } t \geq 0 \\
    0, & \text{for } t < 0 
    \end{cases} \quad (1.1)
\]

In the Heaviside function definition, the value of \( u(0) = 1/2 \) is also used. Note that the independent variable \( t \) is continuous, while the signal itself is not a continuous function. It has a discontinuity at \( t = 0 \).

The boxcar signal (rectangular window) is formed as \( b(t) = u(t + 1/2) - u(t - 1/2) \), that is, \( b(t) = 1 \) for \(-1/2 \leq t < 1/2\) and \( b(t) = 0 \) elsewhere. A signal obtained by multiplying the unit-step signal by \( t \) is called the ramp signal, with notation \( R(t) = tu(t) \).
The impulse signal (or delta function) is defined as
\[
\delta(t) = 0, \text{ for } t \neq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} \delta(t) dt = 1. \quad (1.2)
\]

The impulse signal is equal to 0 everywhere, except at \( t = 0 \), where it assumes an infinite value, so that its area is 1. From the definition of the impulse signal, it follows \( \delta(at) = \delta(t)/|a| \). This function cannot be implemented in real-world systems due to its infinitely short duration and infinitely large amplitude at \( t = 0 \).

In theory, any signal can be expressed by using the impulse signal, as
\[
x(t) = \int_{-\infty}^{\infty} x(t-\tau)\delta(\tau) d\tau = \int_{-\infty}^{\infty} x(\tau)\delta(t-\tau) d\tau. \quad (1.3)
\]

Using the previous relation, it is possible to relate the unit-step signal and the impulse signal,
\[
u(t) = \int_{-\infty}^{\infty} \delta(\tau)u(t-\tau) d\tau = \int_{-\infty}^{t} \delta(\tau) d\tau
\]
or
\[
\frac{du(t)}{dt} = \delta(t). \quad (1.4)
\]

A sinusoidal signal, with amplitude \( A \), frequency \( \Omega_0 \), and initial phase \( \varphi \), is a signal of the form
\[
x(t) = A \sin(\Omega_0 t + \varphi). \quad (1.5)
\]
This signal is periodic in time, since it satisfies the periodicity condition
\[
x(t + T) = x(t). \quad (1.6)
\]
In this case, the period is \( T = 2\pi/\Omega_0 \).

A signal periodic with a basic period \( T \) could also be considered as periodic with periods \( kT \), where \( k \) is an integer.

A complex sinusoidal signal
\[
x(t) = A e^{j(\Omega_0 t + \varphi)} = A \cos(\Omega_0 t + \varphi) + jA \sin(\Omega_0 t + \varphi) \quad (1.7)
\]
is also periodic with period \( T = 2\pi/\Omega_0 \). Fig. 1.1 depicts basic continuous-time signals.
Example 1.1. Find the period of a signal

\[ x(t) = \sum_{n=0}^{N} A_n e^{j\Omega_0 t}. \]

★This signal consists of \( N + 1 \) components. The constant component \( A_0 \) can be considered as periodic with any period. The remaining components \( A_1 e^{j\Omega_0 t}, A_2 e^{j2\Omega_0 t}, A_3 e^{j3\Omega_0 t}, \ldots, A_N e^{jN\Omega_0 t} \) are periodic with periods, \( T_1 = 2\pi/\Omega_0, T_2 = 2\pi/(2\Omega_0), T_3 = 2\pi/(3\Omega_0), \ldots, T_N = 2\pi/(N\Omega_0) \), respectively. A sum of periodic signals is periodic with the period being equal to the smallest time interval \( T \) containing all of the periods \( T_1, T_2, T_3, \ldots, T_N \) an integer number of times. In this case, it is \( T = 2\pi/\Omega_0 \).

Example 1.2. Find the periods of signals: \( x_1(t) = \sin(2\pi t/36), x_2(t) = \cos(4\pi t/15 + 2), x_3(t) = \exp(j0.1t), x_4(t) = x_1(t) + x_2(t) \), and \( x_5(t) = x_1(t) + x_3(t) \).

★Periods are calculated according to (1.6). For \( x_1(t) \) the period follows from \( 2\pi T_1/36 = 2\pi \), as \( T_1 = 36 \). Similarly, \( T_2 = 15/2 \) and \( T_3 = 20\pi \). The period of \( x_4(t) \) is the smallest interval containing \( T_1 \) and \( T_2 \). It is \( T_4 = 180 \) (5 periods of \( x_1(t) \) and 24 periods of \( x_2(t) \)). For signal \( x_5(t) \), when the periods of components are \( T_1 = 36 \) and \( T_3 = 20\pi \), there is no common interval \( T_5 \) such that the periods \( T_1 \) and \( T_3 \) are contained an integer number of times. Thus, the signal \( x_5(t) \) is not periodic.

Some parameters that can be used to describe a signal are:
• Maximum absolute value (magnitude) of a signal

\[ M_x = \max_{-\infty < t < \infty} |x(t)|, \]  
(1.8)

• Signal energy

\[ E_x = \int_{-\infty}^{\infty} |x(t)|^2 \, dt, \]  
(1.9)

• Signal instantaneous power

\[ P_x(t) = |x(t)|^2. \]  
(1.10)

The average signal power is defined by

\[ P_{AV} = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} |x(t)|^2 \, dt. \]

The average power is a time average of energy. Energy signals are signals with a finite energy, while power signals have finite and nonzero power. The average signal power of energy signals is zero.

1.2 PERIODIC SIGNALS AND FOURIER SERIES

Consider a periodic signal \( x(t) \) with a period \( T \). It can be expressed as a sum of weighted periodic complex sinusoidal functions \( e^{j2\pi nt/T}, -\infty < n < \infty \),

\[
x(t) = \cdots + X_{-1}e^{-j2\pi nt/T} + X_0e^{-j0} + X_1e^{j2\pi nt/T} + \cdots
\]  
(1.11)

\[
= \sum_{n=-\infty}^{\infty} X_ne^{j2\pi nt/T}
\]

if the Dirichlet conditions are met: (1) the signal \( x(t) \) has a finite number of discontinuities within the period \( T \); (2) it has a finite average value in the period \( T \); and (3) the signal has a finite number of maxima and minima. Since the signal analysis deals with real-world physical signals, rather than with mathematical generalizations, these conditions are almost always met.
The set of basis functions \( \{ e^{j2\pi nt/T} : -\infty < n < \infty \} \), is an orthonormal set of functions since their inner product is

\[
\langle e^{j2\pi mt/T}, e^{j2\pi nt/T} \rangle = \frac{1}{T} \int_{-T/2}^{T/2} e^{j2\pi mt/T} e^{-j2\pi nt/T} dt = \begin{cases} 
1 & \text{for } m = n \\
\frac{\sin(\pi(m-n))}{\pi(m-n)} & \text{for } m \neq n
\end{cases}
\]

It means that the inner product of any two different basis functions is zero (orthogonal set), while the inner product of a function with itself is 1 (normal set). In the case of orthonormal set of basis functions, it is easy to show that the weighting coefficients \( X_n \) can be calculated as projections of \( x(t) \) onto the basis functions \( e^{j2\pi nt/T} \),

\[
X_n = \langle x(t), e^{j2\pi nt/T} \rangle = \frac{1}{T} \int_{-T/2}^{T/2} x(t) e^{-j2\pi nt/T} dt.
\] (1.12)

This relation follows after a simple multiplication of the right and left sides of (1.11) by \( e^{-j2\pi mt/T} \) and an integration within the period \( \frac{1}{T} \int_{-T/2}^{T/2} (\cdot) dt \).

**Example 1.3.** Show that the Fourier series coefficients \( X_n \) of a periodic signal \( x(t) \) can be obtained by minimizing the mean square error between the signal and \( \sum_{n=-N}^{N} X_n e^{j2\pi nt/T} \) within the period \( T \).

\[ I = \frac{1}{T} \int_{-T/2}^{T/2} \left| x(t) - \sum_{n=-N}^{N} X_n e^{j2\pi nt/T} \right|^2 dt. \]

From \( \frac{\partial I}{\partial X_m^*} = 0 \) follows

\[
\frac{1}{T} \int_{-T/2}^{T/2} e^{-j2\pi mt/T} \left( x(t) - \sum_{n=-N}^{N} X_n e^{j2\pi nt/T} \right) dt = 0
\]

\[
X_m = \frac{1}{T} \int_{-T/2}^{T/2} x(t) e^{-j2\pi mt/T} dt.
\] (1.13)
Note: A derivative of complex function \( F(z) = u + jv = u(x, y) + jv(x, y) \) with \( z = x + jy \) is defined by

\[
\frac{\partial F(z)}{\partial z} = \left( \frac{\partial}{\partial x} - j \frac{\partial}{\partial y} \right) F(x, y),
\]

\[
\frac{\partial F(z)}{\partial z^*} = \left( \frac{\partial}{\partial x} + j \frac{\partial}{\partial y} \right) F(x, y).
\]

Often a half of these values is used in the definition, what does not change our results.

In order to prove the form \( \frac{\partial I}{\partial X^*} = 0 \) let us denote \( X_m \) by \( z \), all terms in \( x(t) - \sum_{n=-N}^{N} X_m e^{j2\pi nt/T} = f(z) \) that not depend on \( z = x + jy = X_m \) by \( a + jb \) and \( e^{j2\pi nt/T} = e^{j\alpha} \), then we have to show that

\[
\frac{\partial F(z)}{\partial z^*} = \frac{\partial |f(z)|^2}{\partial z^*} = 2e^{-j\alpha} f(z).
\]

In our case

\[
|f(z)|^2 = |a + jb + e^{j\alpha}(x + jy)| = (a + x \cos \alpha - y \sin \alpha)^2 + (b + x \sin \alpha + y \cos \alpha)^2.
\]

For the minimization of a function of two variables \( x \) and \( y \) we need partial derivatives

\[
\frac{\partial |f(z)|^2}{\partial x} = 2 \cos \alpha (a + x \cos \alpha - y \sin \alpha) + 2 \sin \alpha (b + x \sin \alpha + y \cos \alpha)
\]

\[
= 2 \Re \{e^{-j\alpha} f(z)\}
\]

and

\[
\frac{\partial |f(z)|^2}{\partial y} = 2 \Im \{e^{-j\alpha} f(z)\}.
\]

Therefore, all calculations with two real-valued equations (1.14) and (1.15) are the same as using one complex valued relation

\[
\frac{\partial |f(z)|^2}{\partial x} + j \frac{\partial |f(z)|^2}{\partial y} = \left( \frac{\partial}{\partial x} + j \frac{\partial}{\partial y} \right) F(x, y) = \frac{\partial F(z)}{\partial z^*}.
\]

Since the signal and the basis functions are periodic with period \( T \), in all previous integrals, we can use

\[
\frac{1}{T} \int_{-T/2}^{T/2} x(t) e^{-j2\pi nt/T} dt = \frac{1}{T} \int_{-T/2+\Lambda}^{T/2+\Lambda} x(t) e^{-j2\pi nt/T} dt
\] (1.16)
where \( \Lambda \) is an arbitrary constant.

The signal expansion \((1.11)\) is known as the Fourier series, and the coefficients \(X_n\) are the Fourier series coefficients.

**Example 1.4.** Calculate the Fourier series coefficients of a periodic signal \(x(t) = \cos^2(\pi t/4)\). What will be the coefficient values if period \(T = 8\) is assumed?

- The signal \(x(t)\) can be written as \(x(t) = (1 + \cos(\pi t/2))/2\). The period is \(T = 4\). Assuming that the Fourier series coefficients are calculated with \(T = 4\), after transforming the signal into \((1.11)\) form, we get

\[
x(t) = \frac{1}{4} e^{-j\pi t/2} + \frac{1}{2} + \frac{1}{4} e^{j\pi t/2}.
\]

The Fourier series coefficients are recognized as \(X_{-1} = 1/4\), \(X_0 = 1/2\) and \(X_1 = 1/4\) (without the calculation defined by \((1.12)\)). Other coefficients are equal to zero. In the above transformation, the relation \(\cos(\pi t/2) = (e^{j\pi t/2} + e^{-j\pi t/2})/2\) is used. If the period \(T = 8\) is used, then the signal is decomposed into complex sinusoids \(e^{2\pi nt/8} = e^{j\pi nt/4}\), \((1.11)\). The signal can be written as

\[
x(t) = \frac{1}{4} e^{-2\pi t/4} + \frac{1}{2} + \frac{1}{4} e^{2\pi t/4}.
\]

Thus, comparing the signal definition with the basis functions \(e^{j\pi nt/4}\), we may write \(X_{-2} = 1/4\), \(X_0 = 1/2\), and \(X_2 = 1/4\). Other coefficients are equal to zero.

**Example 1.5.** Calculate the Fourier series coefficients of a periodic signal \(x(t)\) defined as

\[
x(t) = \sum_{n=-\infty}^{\infty} x_0(t + 2n)
\]

with

\[
x_0(t) = u(t + 1/4) - u(t - 1/4).
\]

- The signal \(x(t)\) is a periodic extension of \(x_0(t)\), with period \(T = 2\). This signal is equal to 1 for \(-1/4 \leq t < 1/4\), within its basic period. Thus,

\[
X_n = \frac{1}{2} \int_{-1/4}^{1/4} e^{-j\pi nt/2} dt = \frac{\sin(\pi n/4)}{\pi n},
\]

with \(X_0 = 1/4\). Values of \(X_n\) are presented in Fig. 1.2.

The signal \(x(t)\) can be reconstructed by using the Fourier series \((1.11)\). In calculations, a finite number of terms denoted by \(M\) should be used,

\[
x_M(t) = \sum_{n=-M}^{M} X_n e^{j\pi nt}.
\]

The reconstructed signal, with \(M = 1, 2, 6,\) and 30, is shown in Fig. 1.3. \(\square\)
Figure 1.2  Periodic signal (left) and its Fourier series coefficients (right).

Figure 1.3  Illustration of signal reconstruction by using a finite Fourier series with: (a) coefficients $X_n$ within $-1 \leq n \leq 1$, (b) coefficients $X_n$ within $-2 \leq n \leq 2$, (c) coefficients $X_n$ within $-6 \leq n \leq 6$, and (d) coefficients $X_n$ within $-30 \leq n \leq 30$.

1.2.1 Fourier Series of Real-Valued Signals

For a real-valued signal $x(t)$ the Fourier series coefficients can be written as

$$X_n = \frac{1}{T} \int_{-T/2}^{T/2} x(t) \cos \left( \frac{2\pi nt}{T} \right) dt - j \frac{1}{T} \int_{-T/2}^{T/2} x(t) \sin \left( \frac{2\pi nt}{T} \right) dt$$

$$= \frac{A_n - jB_n}{2}.$$  \hspace{1cm} (1.20)
where $A_n/2$ and $-B_n/2$ are real and imaginary part of $X_n$. Since $X_n = X^*_{-n}$ holds for real-valued signals, the values of $A_n$ and $B_n$ are

\[
A_n = X_n + X_{-n} = \frac{2}{T} \int_{-T/2}^{T/2} x(t) \cos\left(\frac{2\pi n t}{T}\right) dt, \\
B_n = \frac{X_n - X_{-n}}{-j} = \frac{2}{T} \int_{-T/2}^{T/2} x(t) \sin\left(\frac{2\pi n t}{T}\right) dt.
\]  

(1.21)

The Fourier series form for real-valued signals is

\[
x(n) = -\sum_{n=-\infty}^{-1} X_n e^{j2\pi nT/T} + X_0 + \sum_{n=1}^{\infty} X_n e^{j2\pi nT/T}
\]

\[
= X_0 + \sum_{n=1}^{\infty} \left( X_n e^{j2\pi nT/T} + X_{-n} e^{-j2\pi nT/T} \right)
\]

\[
= A_0 + \sum_{n=1}^{\infty} A_n \cos\left(\frac{2\pi n T}{T}\right) + \sum_{n=1}^{\infty} B_n \sin\left(\frac{2\pi n T}{T}\right)
\]

(1.22)

(1.23)

with $|X_n| = \sqrt{A_n^2 + B_n^2}/2$. For real-valued signals the integrals in (1.21), corresponding to $A_n$ and $B_n$, are even and odd functions with respect to $n$. Therefore, it is possible to calculate

\[
H_n = \frac{1}{T} \int_{-T/2}^{T/2} x(t) \left[ \cos\left(\frac{2\pi n t}{T}\right) + \sin\left(\frac{2\pi n t}{T}\right) \right] dt
\]

(1.24)

and to get

\[
A_n = H_n + H_{-n}, \\
B_n = H_n - H_{-n}.
\]

The coefficients calculated by (1.24) are the Hartley series coefficients. For a real-valued and even signal $x(t) = x(-t)$ this transform reduces to

\[
C_n = \frac{1}{T} \int_{-T/2}^{T/2} x(t) \cos\left(\frac{2\pi n t}{T}\right) dt
\]

corresponding to the Fourier cosine series coefficients. Similar expression is obtained for an odd real-valued signal $x(n)$ when the Fourier series reduces to the Fourier sine series coefficients.
Example 1.6. For a signal
\[ x(t) = t[u(t) - u(t - 1/2)] \]
we are interested in its reconstruction based on the Fourier series coefficients. The rate of coefficients convergence may depend on the way how the periodic extension of this signal is formed.

(a) Calculate the Fourier series of the original signal periodically extended with period \( T = 1/2 \),
\[ x_p(t) = \sum_{n=-\infty}^{\infty} x(t + \frac{1}{2}n). \]
Write the reconstruction formula with \( M \) Fourier series coefficients.

(b) What are the Fourier transform coefficients and the reconstruction formula for
\[ x_p(t) = \sum_{n=-\infty}^{\infty} x(t + n), \]
when the period is \( T = 1 \).

(c) The signal is first extended with its reversed version
\[ x_c(t) = x(t) + x(1 - t) \]
and then periodically extended with period \( T = 1 \). Find the Fourier series coefficients and the reconstruction formula.

(d) Comment the coefficients convergence in all cases.

★ (a) The Fourier series coefficients of this signal are
\[ X_n = \frac{1}{1/2} \int_{0}^{1/2} t e^{-j2\pi n t/(1/2)} dt = \frac{1}{-j4\pi n} \]
with \( X_0 = 1/4 \). The reconstructed signal with \( M \) coefficients is
\[ x_M(t) = \frac{1}{4} + \sum_{n=1}^{M} \left[ -\frac{1}{4j\pi n} e^{j4\pi nt} + \frac{1}{4j\pi n} e^{-j4\pi nt} \right] \]
\[ = \frac{1}{4} - \sum_{n=1}^{M} \frac{\sin(4\pi nt)}{2\pi n}. \]
The reconstructed signal for some values of \( M \) is presented in Fig.1.4.

(b) The Fourier series coefficients of the signal extended with period 1 are
\[ X_n = \frac{1}{1} \int_{0}^{1/2} t e^{-j2\pi n t} dt = \frac{1}{-j2\pi n} \left. t e^{-j2\pi nt} \right|_{0}^{1/2} + \frac{1}{(2\pi n)^2} \left. e^{-j2\pi nt} \right|_{0}^{1/2} \]
\[ = (-1)^n \frac{-1}{-j4\pi n} + (-1)^n - 1 \frac{1}{(2\pi n)^2}. \]
Figure 1.4  Reconstruction of a signal using the Fourier series. Reconstructed signal is denoted by \( x_M(t) \), where \( M \) indicates the number of coefficients used in reconstruction.

with \( X_0 = 1/8 \). Note that the relation between the Fourier coefficients in (a) and (b) is \( 2X_n^{(b)} = X_n^{(a)} \). The reconstruction is presented in Fig. 1.5.

(c) For the signal \( x_c(t) \) extended with its reversed version follows

\[
C_n = \int_{0}^{1/2} te^{-j2\pi nt} \, dt + \int_{1/2}^{1} (1-t)e^{-j2\pi nt} \, dt
\]

\[
= 2 \int_{0}^{1/2} t \cos(2\pi nt) \, dt = \frac{(-1)^n - 1}{2\pi^2 n^2}
\]

with \( C_0 = 1/4 \). The reconstruction formula is

\[
x_M(t) = \frac{1}{4} - 2 \sum_{n=1}^{M} \frac{1 - (-1)^n}{2\pi^2 n^2} \cos(2\pi nt)
\]

\[
= \frac{1}{4} - 2 \sum_{n=1}^{M} \frac{1}{\pi^2 (2n-1)^2} \cos(2\pi (2n-1)t).
\]

The reconstructed signal in this case is presented in Fig. 1.6.

(d) The coefficients convergence in cases (a) and (b) is of order \( 1/n \) while the convergence in the last case (c) is of order \( 1/n^2 \). The best signal reconstruction with the given number of coefficients will be achieved in the case (c). Also for a given reconstruction error we will need the smallest
Figure 1.5  Reconstruction of a periodic signal, with a zero interval extension before using the Fourier series.

Figure 1.6  Reconstruction of a periodic signal after an even extension before using the Fourier series (cosine Fourier series).
number of reconstruction terms $M$ in case (c). This kind of signal extension will be later used as a basis for a definition of the so-called cosine signal transforms.

1.2.2 Linear Systems

A system transforms one signal (input signal) into another signal (output signal). Assume that $x(t)$ is the input signal. The system transformation will be denoted by an operator $T\{\cdot\}$. The output signal can be written as

$$y(t) = T\{x(t)\}. \quad (1.25)$$

A system is linear if, for any two signals $x_1(t)$ and $x_2(t)$ and arbitrary constants $a_1$ and $a_2$, it holds

$$y(t) = T\{a_1x_1(t) + a_2x_2(t)\} = a_1T\{x_1(t)\} + a_2T\{x_2(t)\}. \quad (1.26)$$

We say that a system is time-invariant if its properties and parameters do not change in time. For a time-invariant system, it holds:

if $y(t) = T\{x(t)\}$, then $T\{x(t - t_0)\} = y(t - t_0), \quad (1.27)$

for any $t_0$.

Linear time-invariant (LTI) systems are fully described by their response to the impulse signal. If we know the impulse response of these systems,

$$h(t) = T\{\delta(t)\},$$

then for arbitrary signal $x(t)$ at the input, the output can be calculated, by using (1.3), as

$$y(t) = T\{x(t)\} = T\left\{\int_{-\infty}^{\infty} x(\tau)\delta(t-\tau)d\tau\right\}$$

$$\overset{\text{Linearity}}{=} \int_{-\infty}^{\infty} x(\tau)T\{\delta(t-\tau)\}d\tau \overset{\text{Time-invariance}}{=} \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau.$$ 

The last integral is of particular significance. It is called a convolution in time of $x(t)$ and $h(t)$. Its notation is

$$y(t) = x(t) *_1 h(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau. \quad (1.28)$$
The convolution is a commutative operation

\[ x(t) * h(t) = h(t) * x(t). \]  

(1.29)

**Example 1.7.** Find a convolution of signals \( x(t) = u(t + 1) - u(t - 1) \) and \( h(t) = e^{-t}u(t). \)

 relied to as causal signals.

A system is causal if there is no response before the input signal appears. For causal systems \( h(t) = 0 \) for \( t < 0 \). In general signals that satisfy the property that they may be an impulse response of a causal system may be referred to as causal signals.

A system is stable if any input signal with a finite magnitude \( M_x = \max_{-\infty < t < \infty} |x(t)| \) produces an output \( y(t) \) whose values are finite, \( |y(t)| < \infty \). Sufficient condition that a linear time-invariant system is stable is

\[ \int_{-\infty}^{\infty} |h(\tau)| d\tau < \infty \]  

(1.30)

since

\[ |y(t)| = | \int_{-\infty}^{\infty} x(t - \tau)h(\tau) d\tau | \leq \int_{-\infty}^{\infty} |x(t - \tau)h(\tau)| d\tau \]

\[ = \int_{-\infty}^{\infty} |x(t - \tau)||h(\tau)| d\tau \leq M_x \int_{-\infty}^{\infty} |h(\tau)| d\tau < \infty, \]

if (1.30) holds.

It can be shown that the absolute value integrability of the impulse response is the necessary condition for a linear time-invariant system to be stable as well.
1.3 FOURIER TRANSFORM

The Fourier series has been introduced and presented for periodic signals, with a period $T$. Assume now that we extend the period to infinity, while not changing the signal. This case corresponds to the analysis of an aperiodic signal $x(t)$. Its transform, the Fourier series coefficients normalized by the period, is given by

$$\lim_{T \to \infty} X_n T = \lim_{T \to \infty} \int_{-T/2}^{T/2} x(t) e^{-j2\pi nt/T}dt = \int_{-\infty}^{\infty} x(t) e^{-j\Omega t}dt \quad (1.31)$$

with $2\pi/T = \Delta \Omega \to d\Omega$ (being infinitesimal) and $2\pi n/T \to \Omega$ becoming a continuous variable, as $T \to \infty$ and $-\infty < n < \infty$.

The function $X(\Omega)$, defined by

$$X(\Omega) = \int_{-\infty}^{\infty} x(t) e^{-j\Omega t}dt, \quad (1.32)$$

is called the Fourier transform (FT) of a signal $x(t)$. For the Fourier transform existence it is sufficient that a signal is absolutely integrable. There are some signals that do not satisfy this condition, whose Fourier transform exists in a form of generalized functions, such as delta function.

The inverse Fourier transform (IFT) can be obtained by multiplying both sides of (1.32) by $e^{j\Omega \tau}$ and integrating over $\Omega$,

$$\int_{-\infty}^{\infty} X(\Omega) e^{j\Omega \tau}d\Omega = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(t) e^{j\Omega (\tau-t)}dtd\Omega.$$

Using the fact that

$$\int_{-\infty}^{\infty} e^{j\Omega (\tau-t)}d\Omega = 2\pi \delta(\tau-t)$$

we get the inverse Fourier transform

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\Omega) e^{j\Omega \tau}d\Omega. \quad (1.33)$$
Example 1.8. Calculate the Fourier transform of \( x(t) = Ae^{-at}u(t), a > 0 \).

\[ X(\Omega) = \int_{0}^{\infty} Ae^{-at}e^{-j\Omega t}dt = \frac{A}{(a + j\Omega)}. \]

\[ \square \]

Example 1.9. Find the Fourier transform of

\[ x(t) = \text{sign}(t) = \begin{cases} 1 & \text{for } t > 0 \\ 0 & \text{for } t = 0 \\ -1 & \text{for } t < 0 \end{cases}. \quad (1.34) \]

\[ \star \]Since a direct calculation of the Fourier transform for this signal is not possible, let us consider the signal

\[ x_a(t) = \begin{cases} e^{-at} & \text{for } t > 0 \\ 0 & \text{for } t = 0 \\ -e^{at} & \text{for } t < 0 \end{cases} \]

where \( a > 0 \) is a real-valued constant. It is obvious that

\[ \lim_{a \to 0} x_a(t) = x(t). \]

The Fourier transform of \( x(t) \) can be obtained as

\[ X(\Omega) = \lim_{a \to 0} X_a(\Omega), \]

where

\[ X_a(\Omega) = \int_{-\infty}^{0} -e^{at}e^{-j\Omega t}dt + \int_{0}^{\infty} e^{-at}e^{-j\Omega t}dt = \frac{2\Omega}{ja^2 + j\Omega^2}. \quad (1.35) \]

It results in

\[ X(\Omega) = \frac{2}{j\Omega}. \quad (1.36) \]

\[ \square \]

Based on the definitions of the Fourier transform and the inverse Fourier transform, it is easy to conclude that the duality property holds:

If \( X(\Omega) \) is the Fourier transform of \( x(t) \), then the Fourier transform of \( X(t) \) is \( 2\pi x(-\Omega) \)

\[ X(\Omega) = \text{FT}\{x(t)\} \]

\[ 2\pi x(-\Omega) = \text{FT}\{X(t)\}, \quad (1.37) \]

where \( \text{FT}\{\diamond\} \) denotes the Fourier transform operator.
Example 1.10. Find the Fourier transform of $\delta(t)$, $x(t) = 1$ and $u(t)$.

The Fourier transform of $\delta(t)$ is

$$\text{FT}\{\delta(t)\} = \int_{-\infty}^{\infty} \delta(t)e^{-j\Omega t}dt = 1. \quad (1.38)$$

According to the duality property,

$$\text{FT}\{1\} = 2\pi\delta(\Omega).$$

Finally,

$$\text{FT}\{u(t)\} = \text{FT}\left\{\frac{\text{sign}(t) + 1}{2}\right\} = \frac{1}{j\Omega} + \pi\delta(\Omega). \quad (1.39)$$

### 1.3.1 Fourier Transform and Linear Time-Invariant Systems

Consider a linear, time-invariant system with an impulse response $h(t)$ and the input signal $x(t) = Ae^{j(\Omega_0 t + \phi)}$. The output signal is

$$y(t) = x(t) \ast h(t) = \int_{-\infty}^{\infty} Ae^{j(\Omega_0(t-\tau) + \phi)}h(\tau)d\tau$$

$$= Ae^{j(\Omega_0 t + \phi)} \int_{-\infty}^{\infty} h(\tau)e^{-j\Omega_0 \tau}d\tau = H(\Omega_0)x(t), \quad (1.40)$$

where

$$H(\Omega) = \int_{-\infty}^{\infty} h(t)e^{-j\Omega t}dt \quad (1.41)$$

is the Fourier transform of $h(t)$. The linear time-invariant system does not change the form of an input complex harmonic signal $Ae^{j(\Omega_0 t + \phi)}$. It remains complex harmonic signal after passing through the system, with the same frequency $\Omega_0$. The amplitude of the input signal $x(t)$ is changed for $|H(\Omega_0)|$ and the phase is changed for $\text{arg}\{H(\Omega_0)\}$.

### 1.3.2 Properties of the Fourier Transform

The Fourier transform satisfies the following properties:
1. Linearity

\[ \text{FT}\{a_1 x_1(t) + a_2 x_2(t)\} = a_1 X_1(\Omega) + a_2 X_2(\Omega), \]  

where \( X_1(\Omega) \) and \( X_2(\Omega) \) are the Fourier transforms of signals \( x_1(t) \) and \( x_2(t) \), separately.

2. Realness

The Fourier transform of a signal is real (i.e., \( X^*(\Omega) = X(\Omega) \)), if

\[ x^*(-t) = x(t), \]

since

\[ X^*(\Omega) = \int_{-\infty}^{\infty} x^*(t)e^{j\Omega t}dt = \int_{-\infty}^{\infty} x^*(-t)e^{-j\Omega t}dt = X(\Omega), \]  

if \( x^*(-t) = x(t) \).

3. Modulation

\[ \text{FT}\{x(t)e^{j\Omega_0 t}\} = \int_{-\infty}^{\infty} x(t)e^{j\Omega_0 t}e^{-j\Omega t}dt = X(\Omega - \Omega_0) \]  

\[ \text{FT}\{2x(t)\cos(\Omega_0 t)\} = X(\Omega - \Omega_0) + X(\Omega + \Omega_0). \]

4. Shift in time

\[ \text{FT}\{x(t - t_0)\} = \int_{-\infty}^{\infty} x(t - t_0)e^{-j\Omega t}dt = X(\Omega)e^{-j\Omega t_0}. \]  

5. Time-scaling

\[ \text{FT}\{x(at)\} = \int_{-\infty}^{\infty} x(at)e^{-j\Omega t}dt = \frac{1}{|a|} X\left(\frac{\Omega}{a}\right). \]  

6. Convolution

\[ \text{FT}\{x(t) *_t h(t)\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(\tau)h(t - \tau)e^{-j\Omega t}d\tau dt \]

\[ = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(\tau)h(u)e^{-j\Omega(\tau - u)}d\tau du = X(\Omega)H(\Omega). \]
7. Multiplication

\[ \text{FT}\{x(t)h(t)\} = \frac{1}{2\pi} \int_{-\infty}^{\infty} x(t) \int_{-\infty}^{\infty} H(\theta)e^{j\theta t}d\theta e^{-j\Omega t}dt \]  

\[ = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(\theta)X(\Omega - \theta)d\theta = X(\Omega) \ast_{\Omega} H(\Omega) = H(\Omega) \ast_{\Omega} X(\Omega). \]

Convolution in frequency domain is denoted by \( \ast_{\Omega} \) with a factor of \( 1/2\pi \) being included.

8. Parseval’s theorem

\[ \int_{-\infty}^{\infty} x(t)y^*(t)dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\Omega)Y^*(\Omega)d\Omega \]  

\[ \int_{-\infty}^{\infty} |x(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |X(\Omega)|^2 d\Omega. \]

9. Differentiation

\[ \text{FT}\left\{ \frac{dx(t)}{dt}\right\} = \text{FT}\left\{ \frac{d}{dt}\left( \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\Omega)e^{j\Omega t}d\Omega \right)\right\} = j\Omega X(\Omega). \]  

10. Integration

The Fourier transform of

\[ \int_{-\infty}^{t} x(\tau)d\tau \]

can be calculated as the Fourier transform of

\[ x(t) \ast_t u(t) = \int_{-\infty}^{\infty} x(\tau)u(t-\tau)d\tau = \int_{-\infty}^{t} x(\tau)d\tau. \]

Then,

\[ \text{FT}\left\{ \int_{-\infty}^{t} x(\tau)d\tau \right\} = \text{FT}\{x(t)\} \text{FT}\{u(t)\} = \]

\[ X(\Omega) \left( \frac{1}{j\Omega} + \pi\delta(\Omega) \right) = \frac{1}{j\Omega} X(\Omega) + \pi X(0)\delta(\Omega). \]
11. An analytic part of a signal \( x(t) \), whose Fourier transform is \( X(\Omega) \), is a signal with the Fourier transform defined by

\[
X_a(\Omega) = \begin{cases} 
2X(\Omega) & \text{for } \Omega > 0 \\
X(0) & \text{for } \Omega = 0 \\
0 & \text{for } \Omega < 0 
\end{cases}
\] (1.52)

It can be written as

\[
X_a(\Omega) = X(\Omega) + X(\Omega)\text{sign}(\Omega) = X(\Omega) + jX_h(\Omega)
\] (1.53)

where \( X_h(\Omega) \) is the Fourier transform of the Hilbert transform of the signal \( x(t) \). From Example 1.9 with the signal \( x(t) = \text{sign}(t) \) and the duality property of the Fourier transform pair, obviously the inverse Fourier transform of \( \text{sign}(\Omega) \) is \( j/\pi t \). Therefore, the analytic part of a signal, in the time domain, reads as

\[
x_a(t) = x(t) + jx_h(t) = x(t) + x(t) * \frac{j}{\pi t} = x(t) + \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x(\tau)}{t - \tau} d\tau.
\] (1.54)

where \( p.v. \) stands for Cauchy principal value of the considered integral.

### 1.3.3 Relationship Between the Fourier Series and the Fourier Transform

Consider an aperiodic signal \( x(t) \), with the Fourier transform \( X(\Omega) \). Assume that the signal is of a limited duration (i.e., \( x(t) = 0 \) for \( |t| > T_0/2 \)). Then,

\[
X(\Omega) = \int_{-T_0/2}^{T_0/2} x(t) e^{-j\Omega t} dt.
\] (1.55)

If we make a periodic extension of \( x(t) \), with a period \( T \), we get a signal

\[
x_p(t) = \sum_{n=-\infty}^{\infty} x(t + nT).
\]

The periodic signal \( x_p(t) \) can be expanded into Fourier series with the coefficients

\[
X_n = \frac{1}{T} \int_{-T/2}^{T/2} x_p(t) e^{-j2\pi nt/T} dt.
\] (1.56)
If $T > T_0$ it is easy to conclude that

$$
\frac{T}{2} \int_{-T/2}^{T/2} x_p(t) e^{-j2\pi n t/T} dt = \frac{T_0}{2} \int_{-T_0/2}^{T_0/2} x(t) e^{-j\Omega t} dt|_{\Omega = 2\pi n / T}
$$

or

$$X_n = \frac{1}{T} X(\Omega)|_{\Omega = 2\pi n / T}. \tag{1.57}
$$

It means that the Fourier series coefficients are the samples of the Fourier transform, divided by $T$. The only condition in the derivation of this relation is that the signal duration is shorter than the period of periodic extension (i.e., $T > T_0$). The sampling interval in frequency is

$$\Delta \Omega = \frac{2\pi}{T}, \quad \Delta \Omega < \frac{2\pi}{T_0}.
$$

It should be smaller than $2\pi / T_0$, where $T_0$ is the signal $x(t)$ duration. This is a form of the sampling theorem in the frequency domain. The sampling theorem in the time domain will be discussed later.

In order to write the Fourier series coefficients in the Fourier transform form, note that a periodic signal $x_p(t)$, formed by a periodic extension of $x(t)$ with period $T$, can be written as

$$x_p(t) = \sum_{n=-\infty}^{\infty} x(t + nT) = x(t) * \sum_{n=-\infty}^{\infty} \delta(t + nT). \tag{1.58}
$$

The Fourier transform of this periodic signal is

$$X_p(\Omega) = \text{FT} \left\{ x(t) * \sum_{n=-\infty}^{\infty} \delta(t + nT) \right\}
= X(\Omega) \cdot \frac{2\pi}{T} \sum_{n=-\infty}^{\infty} \delta \left( \Omega - \frac{2\pi}{T} n \right)
= \frac{2\pi}{T} \sum_{n=-\infty}^{\infty} X \left( \frac{2\pi}{T} n \right) \delta \left( \Omega - \frac{2\pi}{T} n \right)
$$

since

$$\text{FT} \left\{ \sum_{n=-\infty}^{\infty} \delta(t + nT) \right\} = \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(t + nT) e^{-j\Omega t} dt
= \sum_{n=-\infty}^{\infty} e^{j\Omega n T} = \frac{2\pi}{T} \sum_{n=-\infty}^{\infty} \delta \left( \Omega - \frac{2\pi}{T} n \right). \tag{1.60}$$
The Fourier transform of a periodic signal is a series of generalized impulse signals at $\Omega = 2\pi n / T$ with weighting factors $X(2\pi n)/T$ being equal to the Fourier series coefficients $X_n$. The relation between periodic generalized impulse signals in the time and frequency domain will be explained (derived) later, (see Example 2.8).

### 1.4 FOURIER TRANSFORM AND STATIONARY PHASE METHOD

When a signal

$$x(t) = A(t)e^{j\phi(t)} \tag{1.61}$$

is not of a simple analytic form, it may be possible, in some cases, to obtain an approximative expression for its Fourier transform by using the method of stationary phase.

The method of stationary phase states that if the phase function $\phi(t)$ is monotonous and the amplitude $A(t)$ is sufficiently smooth function, then

$$\int_{-\infty}^{\infty} A(t)e^{j\phi(t)}e^{-j\Omega t}dt \simeq A(t_0)e^{j\phi(t_0)}e^{-j\Omega t_0} \sqrt{\frac{2\pi j}{|\phi''(t_0)|}}, \tag{1.62}$$

where $t_0$ is the solution of

$$\phi'(t_0) = \Omega.$$ 

The most significant contribution to the integral on the left side of (1.62) comes from the region where the phase $\phi(t) - \Omega t$ of the exponential function $\exp(j(\phi(t) - \Omega t))$ is stationary in time, since the contribution of the intervals with fast varying $\phi(t) - \Omega t$ tends to zero. It means that locally around an instant $t$ the signal behaves as $\exp(j(\phi'(t)t))$. Value

$$\Omega_i(t) = \phi'(t)$$

is called instantaneous frequency of a signal. Around the stationary phase instant $t_0$ holds

$$\left. \frac{d(\phi(t) - \Omega t)}{dt} \right|_{t=t_0} = 0,$$

$$\phi'(t_0) - \Omega = 0.$$ 

Around this point the phase can be expanded into a Taylor series as

$$\phi(t) - \Omega t = [\phi(t_0) - \Omega t_0] + [\phi'(t_0) - \Omega] + \frac{1}{2} \phi''(t_0)t^2 + ...$$
Since $\phi'(t_0) - \Omega = 0$ the integral (1.62) is

$$\int_{-\infty}^{\infty} A(t) e^{i(\phi(t) - \Omega t)} dt \cong A(t_0) e^{i(\phi(t_0) - \Omega t_0)} \int_{-\infty}^{\infty} e^{i\frac{1}{2} \phi''(t_0) t^2} dt$$

where $A(t) \cong A(t_0)$ is also used. With

$$\int_{-\infty}^{\infty} e^{\frac{1}{2}at^2} dt = \sqrt{\frac{2\pi}{a}}$$

the stationary phase approximation follows.

If the equation $\phi'(t_0) = \Omega$ has two (or more) solutions $t_0^+$ and $t_0^-$, then the integral on the left side of (1.62) is equal to the sum of functions at both (or more) stationary phase points. Finally, this relation holds for $\phi''(t_0) \neq 0$. If $\phi''(t_0) = 0$, then similar analysis may be performed, using the lowest nonzero phase derivative at the stationary phase point.

**Example 1.11.** Consider signal

$$x(t) = \exp\left(-\left(t^2 - 1\right)^2\right) \exp(j4\pi t^2 + j10\pi t).$$

Find its Fourier transform approximation by using the stationary phase method.

According to the stationary phase method,

\begin{align*}
8\pi t_0 + 10\pi &= \Omega \\
t_0 &= \frac{\Omega - 10\pi}{8\pi}
\end{align*}

and

$$\phi''(t_0) = 8\pi.$$  \hspace{1cm} (1.63)

The amplitude of $X(\Omega)$ is

$$|X(\Omega)| \simeq A(t_0) \sqrt{\frac{2\pi}{\phi''(t_0)}} = \exp\left(-\left(t_0^2 - 1\right)\frac{t_0^2}{2}\right) \sqrt{\frac{2\pi}{8\pi}}$$

$$= \frac{1}{2} \exp\left(-\left[\left(\frac{\Omega - 10\pi}{8\pi}\right)^2 - 1\right] \left(\frac{\Omega - 10\pi}{8\pi}\right)^2\right)$$  \hspace{1cm} (1.64)

The signal, stationary phase approximation of the Fourier transform and the numerical value of the Fourier transform amplitudes are shown in Fig.1.7
Example 1.12. Consider a frequency-modulated signal

\[ x(t) = A(t) \exp(jat^{2N}). \]

where \( A(t) \) is a slow-varying non-negative function. Find its Fourier transform approximation by using the stationary phase method.

\[ \star \text{According to the stationary phase method, we get that the stationary phase point is } 2Nat_0^{2N-1} = \Omega \text{ with} \]

\[ t_0 = \left( \frac{\Omega}{2Na} \right)^{1/(2N-1)} \]
and
\[
\phi''(t_0) = 2N(2N - 1)a \left( \frac{\Omega}{2Na} \right)^{(2N-2)/(2N-1)}. \tag{1.65}
\]

The amplitude and phase of \(X(\Omega)\), according to (1.62), are
\[
|X(\Omega)|^2 \simeq A^2(t_0) \left| \frac{2\pi}{\phi''(t_0)} \right| \tag{1.66}
\]
\[
= A^2 \left( \left( \frac{\Omega}{2Na} \right)^{1/(2N-1)} \right) \left| \frac{2\pi}{(2N-1)\Omega} \left( \frac{\Omega}{2nN} \right)^{1/(2N-1)} \right| \]
\[
\arg \{X(\Omega)\} \simeq \phi(t_0) - \Omega t_0 + \pi/4 = \left( \frac{1 - 2N}{2N} \Omega \right) \left( \frac{\Omega}{2nN} \right)^{1/(2N-1)} + \pi/4.
\]
for a large value of \(a\).

For \(N = 1\) and \(A(t) = 1\), we get \(|X(\Omega)|^2 = |\pi/a|\) and \(\arg \{X(\Omega)\} = -\Omega^2/(4a) + \pi/4\).

The method of stationary phase may be defined in the frequency domain as well. For a Fourier transform
\[
X(\Omega) = B(\Omega)e^{j\theta(\Omega)} \tag{1.67}
\]
the method of stationary phase states that if the Fourier transform phase function \(\theta(t)\) is monotonous and the amplitude \(B(t)\) is sufficiently smooth function, then
\[
x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} B(\Omega)e^{j\theta(\Omega)} e^{j\Omega t} d\Omega \simeq B(\Omega_0)e^{j\theta(\Omega_0)} e^{j\Omega_0 t} \sqrt{\frac{j}{2\pi |\theta''(\Omega_0)|}}, \tag{1.68}
\]
where \(\Omega_0\) is the solution of
\[
-\theta'(\Omega_0) = t,
\]
and
\[
t_g = -\theta'(\Omega)
\]
is the group delay.

**Example 1.13.** Consider a system with transfer function
\[
H(\Omega) = \exp(-\Omega^2) \exp(-ja\Omega^2/2 - jb\Omega).
\]
Find its impulse response by using the stationary phase method.
According to the stationary phase method,

\[ a\Omega_0 + b = t \]

\[ \Omega_0 = \frac{t - b}{a} \]

and

\[ \theta''(\Omega_0) = -a. \]

The impulse response is

\[ h(t) \approx \exp\left(-\Omega_0^2\right)e^{-ja\Omega_0^2/2-jb\Omega_0+jbt} \sqrt{\frac{j}{2\pi|\theta''(\Omega_0)|}} \]

\[ = \exp\left(-\left(\frac{t-b}{a}\right)^2\right)e^{j((t-b)^2/(2a)+\pi/4)} \sqrt{\frac{1}{2\pi a}}. \]

The signal amplitude is delayed for \( b \). The second order parameter \( a \) in the phase function scales time axis of the impulse response. This is an undesirable effect in common systems.

\[ \square \]

**Example 1.14.** For a system with frequency response \( H(\Omega) = |H(\Omega)|e^{i0} \) the impulse response is \( h(t) \). Find the impulse response of the systems with transfer functions shown in Fig.1.8 with:

(a) \( H_a(\Omega) = |H(\Omega)|e^{-j4\Omega} \),

(b) \( H_b(\Omega) = |H(\Omega)|e^{-j2\pi\Omega^2} \), and

(c) \( H_c(\Omega) = |H(\Omega)|\left[\frac{3}{4} + \frac{1}{4}\cos(2\pi\Omega^2)\right]e^{i0}. \)

\[ \star \](a) The impulse response is

\[ h_a(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(\Omega)e^{-j4\Omega}e^{i\Omega t} = h(t - 4). \]

It is delayed with respect to \( h(t) \) for \( t_0 = 4 \).

(b) In this case

\[ h_b(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(\Omega)e^{-j2\pi\Omega^2}e^{i\Omega t}d\Omega. \]

The group delay is \( t_g = -\theta'(\Omega) = 4\pi\Omega \). According to the stationary phase method, by replacing \( \Omega \) with \( t/(4\pi) \), we get

\[ h_b(t) = H\left(\frac{t}{4\pi}\right)e^{i(t^2/8\pi+\pi/4)} \sqrt{\frac{1}{8\pi^2}}. \]
Figure 1.8  Frequency response of systems (amplitude, top row, and phase, middle row) with corresponding impulse responses (amplitude - bottom row).
Nonlinear group delay causes time scaling and form change of the impulse response. Note: Check that \( \int_{-\infty}^{\infty} h^2(t)dt = \int_{-\infty}^{\infty} h_b^2(t)dt \). The impulse response is calculated numerically as well. The agreement with the approximative result is high, Fig. 1.8 (third row, middle column).

(c) Since \( 2\cos(2\pi\Omega^2) = \exp(j2\pi\Omega^2) + \exp(-j2\pi\Omega^2) \)

\[
h_c(t) = \frac{3}{4}h(t) + \frac{1}{8}h_b(t) + \frac{1}{8}h_b(-t) = \frac{3}{4}h(t) + \frac{1}{4}h_b(t).
\]

Here, fast variations of the amplitude result in a two-component impulse response, one being proportional to the impulse response from case (a) and the other proportional to the form from (b).

\[ \hfill \square \hfill \]

### 1.5 LAPLACE TRANSFORM

The Fourier transform could be considered as a special case of the Laplace transform. At the beginning, Fourier’s work was even not published as an original contribution due to this fact. The Laplace transform is defined by

\[
X(s) = \mathcal{L}\{x(t)\} = \int_{-\infty}^{\infty} x(t)e^{-st}dt, \tag{1.69}
\]

where \( s = \sigma + j\Omega \) is a complex number. It is obvious that the Fourier transform is the value of a Laplace transform along the imaginary axis, \( \sigma = 0 \) or \( s = j\Omega \). This form of the Laplace transform is also known as the bilateral Laplace transform (in contrast to unilateral one, where the integration limits are from 0 to \( \infty \)).

**Example 1.15.** Calculate the Laplace transform of \( x(t) = e^{-at}u(t) \).

According to the definition

\[
X(s) = \int_{0}^{\infty} e^{-at}e^{-st}dt = -\frac{e^{-\{s+a\}t}}{s+a} \bigg|_{0}^{\infty} = \frac{1}{s+a}
\]

if

\[
\lim_{t \to \infty} e^{-(s+a)t} = 0
\]

or \( \sigma + a > 0 \), that is, \( \sigma > -a \). Therefore, the region of convergence of this Laplace transform is the region where \( \sigma > -a \). The point \( s = -a \) is the pole of
the Laplace transform. The region of convergence is limited by a vertical line in the complex s-plane, passing through a pole.

The Laplace transform may be considered as a Fourier transform of a signal \( x(t) \) multiplied by \( \exp(-\sigma t) \), with varying parameter \( \sigma \),

\[
\text{FT}\{ x(t)e^{-\sigma t} \} = \int_{-\infty}^{\infty} x(t)e^{-\sigma t}e^{-j\Omega t}dt = \int_{-\infty}^{\infty} x(t)e^{-st}dt = X(s).
\] (1.70)

In this way, we may calculate the Laplace transform of functions that are not absolutely integrable (i.e., do not satisfy condition for the Fourier transform existence, \( \int_{-\infty}^{\infty} |x(t)|dt < \infty \)) In these cases, for some values of \( \sigma \), the new signal \( x(t)e^{-\sigma t} \) may be absolutely integrable and the Laplace transform could exist. In the previous example, the Fourier transform does not exist for \( a < 0 \), while for \( a = 0 \) it exists in the generalized functions sense only. Laplace transform of the considered signal always exists, with the region of convergence \( \sigma > -a \). If \( a > 0 \), then the region of convergence \( \sigma > -a \) includes the line \( \sigma = 0 \), meaning that the Fourier transform exists.

The inverse Laplace transform is

\[
x(t) = \frac{1}{2\pi j} \lim_{T \to \infty} \int_{\gamma-jT}^{\gamma+jT} X(s)e^{st}ds
\]

where the integration is performed along a path in the region of convergence of \( X(s) \).

**Example 1.16.** Consider a signal \( x(t) \) such that \( x(t) = 0 \) for \( |t| > T \) (time-limited signal). Its Fourier transform is \( X(\Omega) \). Derive the relation to calculate the Laplace transform \( X(s) \) for any \( \sigma \) within the region of convergence, based on the value of \( X(\Omega) \).

★Based on \( X(\Omega) \) the signal values are \( x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\Omega)e^{j\Omega t}d\Omega \). The Laplace transform is

\[
X(s) = \int_{-T}^{T} \left( \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\Omega)e^{j\Omega t}d\Omega \right) e^{-st}dt
\]

\[
= \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\Omega) \int_{-T}^{T} e^{-st+j\Omega t}dt d\Omega = \frac{1}{\pi} \int_{-\infty}^{\infty} X(\Omega) \frac{\sinh((j\Omega - s)T)}{j\Omega - s}d\Omega.
\] (1.71)

within the region of convergence.  \( \Box \)
Properties of the Laplace transform may easily be generalized from those presented for the Fourier transform, like for example

\[ \mathcal{L}\{ax(t) + by(t)\} = a\mathcal{L}\{x(t)\} + b\mathcal{L}\{y(t)\} = aX(s) + bY(s), \]

\[ \mathcal{L}\{x(t) * h(t)\} = \mathcal{L}\{x(t)\} \mathcal{L}\{h(t)\} = X(s)H(s). \]

Since the Laplace transform will be used to describe linear systems described by linear differential equations we will consider only the relation of the signal derivatives with the corresponding forms in the Laplace domain. In general the Laplace transform of the first derivative \( dx(t)/dt \) of a signal \( x(t) \) is

\[
\int_{-\infty}^{\infty} \frac{dx(t)}{dt} e^{-st} dt = s \int_{-\infty}^{\infty} x(t) e^{-st} dt = sX(s).
\]

This relation follows by integration in part of the first integral, with the assumption that the values of \( x(t)e^{-st} \) are zero as \( t \to \pm \infty \).

In many applications it has been assumed that the systems are causal with corresponding causal signals used in calculations. In these cases \( x(t) = 0 \) for \( t < 0 \), i.e., \( x(t) = x(t)u(t) \). Then the so called one-sided Laplace transform (unilateral Laplace transform) is used. Its definition is

\[ X(s) = \int_{0}^{\infty} x(t)e^{-st} dt. \]

When dealing with the derivatives of causal signals we have to take care about possible discontinuity at \( t = 0 \).

In general the first derivative of the function \( x(t)u(t) \) is

\[ \frac{d(x(t)u(t))}{dt} = \frac{dx(t)}{dt} u(t) + x(0)\delta(t). \]

The Laplace transform of the first derivative of a causal signal is

\[
\int_{0}^{\infty} \frac{dx(t)}{dt} e^{-st} dt = x(t)e^{-st}|_{0}^{\infty} + s \int_{0}^{\infty} x(t)e^{-st} dt = sX(s) - x(0).
\]

Value of signal at \( t = 0 \), denoted by \( x(0) \), is the initial condition.
These relations can easily be generalized to higher order derivatives

\[
\int_0^\infty \frac{d^n x(t)}{dt^n} e^{-st} dt = \frac{s^n}{0} \int_{0}^{\infty} x(t) e^{-st} dt - s^{n-1} x(0) - s^{n-2} x'(0) - ... - x^{(n-1)}(0)
\]

\[
= s^n X(s) - s^{n-1} x(0) - s^{n-2} x'(0) + ... - x^{(n-1)}(0).
\]

The Laplace transform of an integral of \(x(t)\) is

\[
\mathcal{L}\{\int_0^t x(\tau) d\tau\} = \mathcal{L}\{u(t) * x(t)\} = \frac{1}{s}X(s),
\]

since \(\mathcal{L}\{u(t)\} = \int_0^\infty e^{-st} dt = 1/s\).

The initial and final values of the signal are \(x(0) = \lim_{s \to \infty} sX(s)\) and \(x(\infty) = \lim_{s \to 0} sX(s)\), respectively.

### 1.5.1 Linear Systems Described by Differential Equations

After we have established the relation between the Laplace transform and signals derivatives we may use it to analyze the systems described by differential equations. Consider a causal system

\[
a_N \frac{d^N y(t)}{dt^N} + ... + a_1 \frac{dy(t)}{dt} + a_0 y(t) = b_M \frac{d^M x(t)}{dt^M} + ... + b_1 \frac{dx(t)}{dt} + b_0 x(t)
\]

with the initial conditions \(x(0) = x'(0) = x^{(n-1)}(0) = 0\). The Laplace transform of both sides of this differential equation is

\[
a_N s^N Y(s) + ... + a_1 s Y(s) + a_0 Y(s) = b_M s^M X(s) + ... + b_1 s X(s) + b_0 X(s).
\]

Transfer function of this system is of the form

\[
H(s) = \frac{Y(s)}{X(s)} = \frac{b_M s^M + ... + b_1 s + b_0}{a_N s^N + ... + a_1 s + a_0}
\]

**Example 1.17.** A causal system is described by the differential equation

\[
\frac{d^2 y(t)}{dt^2} + 3 \frac{dy(t)}{dt} + 2y(t) = x(t)
\]

with the initial conditions \(y'(0) = 1\) and \(y(0) = 0\). Find the system output \(y(t)\) for \(x(t) = e^{-4t} u(t)\).
The Laplace transform of both sides is
\[ (s^2 Y(s) - sy(0) - y'(0)) + 3(sY(s) - y(0)) + 2Y(s) = X(s) \]
or
\[ Y(s)(s^2 + 3s + 2) = X(s) + sy(0) + y'(0) + 3y(0). \]
With \( X(s) = 1/(s + 4) \) follows
\[ Y(s) = \frac{s + 5}{(s + 4)(s^2 + 3s + 2)} = \frac{A_1}{s + 4} + \frac{A_2}{s + 1} + \frac{A_3}{s + 2}. \]
The coefficients \( A_i \) are obtained from
\[ A_i = (s - s_i)Y(s) |_{s = s_i}. \]
For example,
\[ A_1 = (s + 4) \frac{s + 5}{(s + 4)(s^2 + 3s + 2)} |_{s = -4} = 1/6. \]
The other two coefficients are \( A_2 = -3/2 \) and \( A_3 = 4/3. \)
The output signal is
\[ y(t) = \frac{1}{6} e^{-4t} u(t) - \frac{3}{2} e^{-2t} u(t) + \frac{4}{3} e^{-t} u(t). \]

### 1.5.2 Table of the Laplace Transform

<table>
<thead>
<tr>
<th>Signal ( x(t) )</th>
<th>Laplace transform ( X(s) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \delta(t) )</td>
<td>1</td>
</tr>
<tr>
<td>( u(t) )</td>
<td>( \frac{1}{s} )</td>
</tr>
<tr>
<td>( e^{at} u(t) )</td>
<td>( \frac{1}{s-a} )</td>
</tr>
<tr>
<td>( tu(t) )</td>
<td>( \frac{1}{s^2} )</td>
</tr>
<tr>
<td>( e^{at} \cos(\Omega_0 t) u(t) )</td>
<td>( \frac{1}{(s-a)^2 + \Omega_0^2} )</td>
</tr>
<tr>
<td>( e^{at} \sin(\Omega_0 t) u(t) )</td>
<td>( \frac{\Omega_0}{(s-a)^2 + \Omega_0^2} )</td>
</tr>
<tr>
<td>( te^{at} u(t) )</td>
<td>( \frac{1}{(s-a)^2} )</td>
</tr>
<tr>
<td>( x'(t) u(t) )</td>
<td>( sX(s) - x(0) )</td>
</tr>
<tr>
<td>( tx(t) u(t) )</td>
<td>( -dX(s)/ds )</td>
</tr>
<tr>
<td>( x(t) u(t) / t )</td>
<td>( \int_{s}^{\infty} F(s) , ds )</td>
</tr>
<tr>
<td>( e^{at} x(t) u(t) )</td>
<td>( X(s - a) )</td>
</tr>
<tr>
<td>( x(t) \ast u(t) = \int_{0}^{t} x(t) , dt )</td>
<td>( \frac{X(s)}{s} )</td>
</tr>
</tbody>
</table>
1.6 BUTTERWORTH FILTER

The most common processing systems in communications and signal processing are filters, used to selectively pass a part of the input signal in the frequency domain and to reduce possible interferences. The basic form is a lowpass filter. Here we will present a simple Butterworth lowpass filter.

The squared frequency response of the Butterworth lowpass filter is

\[ |H(j\Omega)|^2 = \frac{1}{1 + \left(\frac{\Omega}{\Omega_c}\right)^{2N}}. \]

It is shown in Fig. 1.9. This filter definition contains two parameters. Order of the filter is \( N \). It is a measure of the transition sharpness from the passband to the stopband region. For \( N \to \infty \) the amplitude form of an ideal lowpass filter is achieved. The second parameter is the critical frequency. At \( \Omega = \Omega_c \) we get \( |H(j\Omega_c)|^2 = |H(0)|^2 / 2 = 1/2 \), corresponding to \(-3\,dB\) gain for any filter order \( N \).

The squared frequency response may be written as

\[ H(j\Omega)H(-j\Omega) = \frac{1}{1 + \left(\frac{j\Omega}{\Omega_c}\right)^{2N}} \]
\[ H(s)H(-s) = \frac{1}{1 + \left(\frac{s}{\Omega_c}\right)^{2N}} \text{ for } s = j\Omega. \]
Poles of the product of transfer functions $H(s)H(-s)$ are

$$
\left( \frac{s_k}{\Omega_c} \right)^{2N} = -1 = e^{i(2\pi k + \pi)}
$$

$$
s_k = \Omega_c e^{i(2\pi k + \pi)/2N + j\pi/2} \text{ for } k = 0, 1, 2, ..., 2N - 1.
$$

Poles of the Butterworth filter are located on a circle whose radius is $\Omega_c$ at the angles

$$
\alpha_k = \frac{2\pi k + \pi}{2N} + \frac{\pi}{2} \text{ for } k = 0, 1, 2, ..., 2N - 1.
$$

For a given filter order $N$ and frequency $\Omega_c$, the only remaining decision is to select a half of the poles $s_k$ that belong to $H(s)$ and to declare that the remaining half of the poles belong to $H(-s)$. Since we want that a filter is stable then we chose the poles

$$s_0, s_1, ..., s_{N-1}$$

within the left side of the $s$ plane, where $\text{Re}\{s\} < 0$, i.e., $\pi/2 < \alpha_k < 3\pi/2$. The symmetric poles with $\text{Re}\{s\} > 0$ are the poles of $H(-s)$. They are not used in the filter design.

**Example 1.18.** Design a lowpass Butterworth filter with:
(a) $N = 3$ with $\Omega_c = 1$,
(b) $N = 4$ with $\Omega_c = 3$.

★(a) Poles for $N = 3$ with $\Omega_c = 1$ have the phases

$$
\alpha_k = \frac{2\pi k + \pi}{6} + \frac{\pi}{2} \text{ for } k = 0, 1, 2.
$$
Their values are

\[ s_0 = \cos\left(\frac{2\pi}{3}\right) + j\sin\left(\frac{2\pi}{3}\right) = -\frac{1}{2} + j\frac{\sqrt{3}}{2} \]
\[ s_1 = \cos\left(\frac{2\pi}{3} + \frac{\pi}{3}\right) + j\sin\left(\frac{2\pi}{3} + \frac{\pi}{3}\right) = -1 \]
\[ s_2 = \cos\left(\frac{2\pi}{3} + \frac{2\pi}{3}\right) + j\sin\left(\frac{2\pi}{3} + \frac{2\pi}{3}\right) = -\frac{1}{2} - j\frac{\sqrt{3}}{2} \]

with

\[ H(s) = \frac{c}{(s + \frac{1}{2} - j\frac{\sqrt{3}}{2})(s + \frac{1}{2} + j\frac{\sqrt{3}}{2})(s + 1)} = \frac{1}{(s^2 + s + 1)(s + 1)} \]

where \( c = 1 \) is used to make \( H(0) = 1 \).

(b) Poles for \( N = 4 \) with \( \Omega_c = 3 \) are at angles

\[ \alpha_k = \frac{2\pi k + \pi}{8} + \frac{\pi}{2}, \text{for } k = 0, 1, 2, 3. \]

Their values are

\[ s_0 = 3\cos\left(\frac{\pi}{2} + \frac{\pi}{8}\right) + j3\sin\left(\frac{\pi}{2} + \frac{\pi}{8}\right) \]
\[ s_1 = 3\cos\left(\frac{\pi}{2} + \frac{3\pi}{8}\right) + j3\sin\left(\frac{\pi}{2} + \frac{3\pi}{8}\right) \]
\[ s_2 = 3\cos\left(\frac{\pi}{2} + \frac{5\pi}{8}\right) + j3\sin\left(\frac{\pi}{2} + \frac{5\pi}{8}\right) \]
\[ s_3 = 3\cos\left(\frac{\pi}{2} + \frac{7\pi}{8}\right) + j3\sin\left(\frac{\pi}{2} + \frac{7\pi}{8}\right) \]

with

\[ H(s) = \frac{c}{(s^2 + 2.296s + 9)(s^2 + 5.543s + 9)} = \frac{9}{(s^2 + 2.296s + 9)(s^2 + 5.543s + 9)} \]

where \( c = 9 \) is used to make \( H(0) = 1 \).

\[ \square \]

In practice we usually do not know the filter order, but its passband frequency \( \Omega_p \) and stopband frequency \( \Omega_s \), with a maximal attenuation in the passband \( a_p[\text{dB}] \) and a minimal attenuation in the stopband \( a_p[\text{dB}] \), as shown in Fig. 1.11. Based on these values we can calculate the order \( N \) and the critical frequency \( \Omega_c \) needed for a filter design.
The relations for $N$ and $\Omega_c$ are

\[
\frac{1}{1 + \left(\frac{\Omega_p}{\Omega_c}\right)^2} \geq \left(\frac{A_p}{A}\right)^2
\]

\[
\frac{1}{1 + \left(\frac{\Omega_s}{\Omega_c}\right)^2} \leq \left(\frac{A_s}{A}\right)^2.
\]

Using equality in both relations, it follows

\[
N = 1 - \frac{\frac{1}{A_p} - 1}{\ln \Omega_p - \ln \Omega_s} - \frac{\frac{1}{A_s} - 1}{\ln \Omega_p - \ln \Omega_s}.
\]

Nearest greater integer is assumed for the filter order $N$. Then we can use any of the relations in (1.72) with equality sign to calculate $\Omega_c$. If we choose the first one then $\Omega_c$ will satisfy \[|H(j\Omega_p)|^2 = A_p^2,\] while if we use the second relation the value of $\Omega_c$ will satisfy \[|H(j\Omega_s)|^2 = A_s^2.\] These two values differ. However both of them are within the defined criteria for the transfer function.

The relation

\[ a = 20 \log A \]

or $A = 10^{a/20}$ should be used for the attenuation given in [dB].

All other filter forms, like passband and highpass, may be obtained from a lowpass filter with appropriate signal modulations. These modulations will be discussed for discrete-time filter forms in Chapter V.
Chapter 4

z-Transform

The Fourier transform of discrete signals and the DFT are used for direct signal processing and calculations. A transform that generalizes these transforms, in the same way as the Laplace transform generalizes the Fourier transform of continuous signals, is the z-transform. This transform provides an efficient tool for qualitative analysis and design of the discrete systems.

4.1 DEFINITION OF THE Z-TRANSFORM

The Fourier transform of a discrete-time signal \( x(n) \) can be considered as a special case of the z-transform defined by

\[
X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n}, \quad (4.1)
\]

where \( z = r\exp(j\omega) \) is a complex number. Value of the z-transform along the unit circle \(|z| = 1\) or \( z = \exp(j\omega) \) is equal to the Fourier transform of discrete-time signals.

The z-transform, in general, converges only for some values of the complex argument \( z \). The region of values of \( z \) where \( X(z) \) is finite is the region of convergence (ROC) of the z-transform.

**Example 4.1.** Consider a discrete-time signal

\[
x(n) = a^n u(n) + b^n u(n),
\]

where \( a \) and \( b \) are complex numbers, \(|a| < |b|\). Find the z-transform of this signal and its region of convergence.
The $z$-transform of $x(n)$ is
\[ X(z) = \sum_{n=0}^{\infty} (a^n z^{-n} + b^n z^{-n}) = \sum_{n=0}^{\infty} (a/z)^n + \sum_{n=0}^{\infty} (b/z)^n \]
\[ = \frac{1}{1 - a/z} + \frac{1}{1 - b/z} = \frac{z}{z - a} + \frac{z}{z - b}. \]

Infinite geometric series with progression coefficient $(a/z)$ converges for $|a/z| < 1$. The other series converges for $|b/z| < 1$. The last term has pole at $z = 0$. Since $|b| > |a|$, the region of convergence is $|a| < |z| < |b|$, Fig. 4.1.

**Example 4.2.** Consider a discrete-time signal
\[ x(n) = a^n u(n - 1) - b^n u(-n - 1) + 2\delta(n - 2), \]
where $a$ and $b$ are complex numbers, $|b| > |a|$. Find the $z$-transform of $x(n)$ and its region of convergence.

The $z$-transform is
\[ X(z) = \sum_{n=1}^{\infty} a^n z^{-n} - \sum_{n=-\infty}^{-1} b^n z^{-n} + 2z^{-2} = \sum_{n=1}^{\infty} a^n z^{-n} - \sum_{n=1}^{\infty} b^{-n} z^n + 2z^{-2} \]
\[ = \frac{a/z}{1 - a/z} - \frac{z/b}{1 - z/b} + 2z^{-2} = \frac{a}{z - a} + \frac{z}{z - b} + 2z^{-2}. \]

Infinite geometric series with progression coefficient $(a/z)$ converges for $|a/z| < 1$. The other series converges for $|z/b| < 1$. The last term has pole at $z = 0$. Since $|b| > |a|$ the region of convergence is $|a| < |z| < |b|$, Fig. 4.2.

Note that in this example and the previous one two different signals $b^n u(n)$ and $-b^n u(-n - 1)$ produced the same $z$-transform $X_b(z) = z/(z - b)$, but with different regions of convergence.
4.2 PROPERTIES OF THE Z-TRANSFORM

4.2.1 Linearity

The $z$-transform is linear since

$$Z\{ax(n) + by(n)\} = \sum_{n=-\infty}^{\infty} [ax(n) + by(n)]z^{-n} = aX(z) + bY(z)$$

with the region of convergence being at least the intersection of the regions of convergence of $X(z)$ and $Y(z)$. In special cases the region can be larger than the intersection of the regions of convergence of $X(z)$ and $Y(z)$ if some poles, defining the region of convergence, cancel out in the linear combination of transforms.

4.2.2 Time-Shift

For a shifted signal $x(n - n_0)$ the $z$-transform is

$$Z\{x(n - n_0)\} = \sum_{n=-\infty}^{\infty} x(n - n_0)z^{-n} = \sum_{n=-\infty}^{\infty} x(n)z^{-(n+n_0)} = X(z)z^{-n_0}.$$ 

Additional pole at $z = 0$ is introduced for $n_0 > 0$. The region of convergence is the same except for $z = 0$ or $z \to \infty$, depending on the value of $n_0$.

Example 4.3. For a causal signal $x(n) = x(n)u(n)$ find the $z$-transform of $x(n + n_0)u(n)$, for $n_0 \geq 0$. 

---

Figure 4.2  Regions of convergence (gray area)
The signal \( x(n + n_0)u(n) \) has a \( z \)-transform

\[
\mathcal{Z}\{x(n + n_0)u(n)\} = \sum_{n=0}^{\infty} x(n + n_0)z^{-n} = \sum_{n=0}^{\infty} x(n + n_0)z^{-(n + n_0)}z^{n_0}
\]

\[
= z^{n_0} \left[ \sum_{n=0}^{\infty} x(n)z^{-n} - (0) - x(1)z^{-1} - \ldots - x(n_0 - 1)z^{-n_0+1} \right]
\]

\[
= z^{n_0} \left[ X(z) - x(0) - x(1)z^{-1} - \ldots - x(n_0 - 1)z^{-n_0+1} \right].
\]

For \( n_0 = 1 \) follows \( \mathcal{Z}\{x(n + 1)u(n)\} = zX(z) - x(0) \). Note that for this signal \( x(n + n_0)u(n) \neq x(n + n_0)u(n + n_0) \).

### 4.2.3 Multiplication by exponential signal: Modulation

For a signal multiplied by an exponential signal the \( z \)-transform is

\[
\mathcal{Z}\{a^n x(n)\} = \sum_{n=-\infty}^{\infty} x(n)a^n z^{-n} = X\left(\frac{z}{a}\right),
\]

with region of convergence being scaled by \( |a| \). In a special case when \( a = e^{j\omega_0} \), the \( z \)-transform plane is just rotated

\[
\mathcal{Z}\{e^{j\omega_0 n} x(n)\} = \sum_{n=-\infty}^{\infty} x(n)e^{j\omega_0 n}z^{-n} = X(ze^{-j\omega_0})
\]

with the same region of convergence as \( X(z) \).

### 4.2.4 Differentiation

Consider the \( z \)-transform of a causal signal \( x(n) \)

\[
X(z) = \sum_{n=0}^{\infty} x(n)z^{-n} \quad \text{and} \quad \frac{dX(z)}{dz} = \sum_{n=0}^{\infty} -nx(n)z^{-n-1}.
\]

We can conclude that

\[
\mathcal{Z}\{nx(n)u(n)\} = -z\frac{dX(z)}{dz}.
\]

This kind of the \( z \)-transform derivations can be generalized to

\[
\mathcal{Z}\{n(n - 1)...(n - N - 1)x(n)u(n)\} = (-1)^Nz^N \frac{d^N X(z)}{dz^N}.
\]
4.2.5 Convolution in time

The z-transform of a convolution of signals \(x(n)\) and \(y(n)\) is

\[
\mathcal{Z}\{x(n) \ast y(n)\} = \mathcal{Z}\left\{ \sum_{m=-\infty}^{\infty} x(m)y(n-m) \right\}
\]

\[
= \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} x(m)y(n-m)z^{-n} = \sum_{l=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} x(m)y(l)z^{-m-l} = X(z)Y(z)
\]

with the region of convergence being at least the intersection of the regions of convergence of \(X(z)\) and \(Y(z)\). In the case of a product of two z-transforms it may happen that some poles are canceled out causing that the resulting region of convergence is larger than the intersection of the individual regions of convergence.

4.2.6 Table of the z-transform

<table>
<thead>
<tr>
<th>Signal (x(n))</th>
<th>z-transform (X(z))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\delta(n))</td>
<td>1</td>
</tr>
<tr>
<td>(u(n))</td>
<td>(\frac{z}{1-z}, \</td>
</tr>
<tr>
<td>(a^{m}u(n))</td>
<td>(\frac{z}{a-z}, \</td>
</tr>
<tr>
<td>(na^{n-1}u(n))</td>
<td>(\frac{z}{(a-z)^{n}}, \</td>
</tr>
<tr>
<td>(-a^{m}u(-n-1))</td>
<td>(\frac{z}{a-z}, \</td>
</tr>
<tr>
<td>(a^{m}x(n))</td>
<td>(X(z/a))</td>
</tr>
<tr>
<td>(a^{n},</td>
<td>a</td>
</tr>
<tr>
<td>(x(n-n_{0}))</td>
<td>(z^{-n_{0}}X(z))</td>
</tr>
<tr>
<td>(nx(n)u(n))</td>
<td>(-zdX(z)/dz)</td>
</tr>
<tr>
<td>(n(n-1)x(n)u(n))</td>
<td>(z^{2}d^{2}X(z)/dz^{2})</td>
</tr>
<tr>
<td>(\cos(\omega_{0}n)u(n))</td>
<td>(\frac{1-z^{-1}\cos(\omega_{0})}{1-2z^{-1}\cos(\omega_{0})+z^{-2}})</td>
</tr>
<tr>
<td>(\sin(\omega_{0}n)u(n))</td>
<td>(\frac{1-z^{-1}\sin(\omega_{0})}{1-2z^{-1}\cos(\omega_{0})+z^{-2}})</td>
</tr>
<tr>
<td>(\frac{1}{n}u(n))</td>
<td>(\exp(z))</td>
</tr>
<tr>
<td>([x(n)u(n)] \ast u(n) = \sum_{m=-\infty}^{\infty} x(m))</td>
<td>(\frac{z}{z^{2}}X(z))</td>
</tr>
</tbody>
</table>
4.2.7 Initial and Stationary State Signal Value

The initial value of a causal signal may be calculated as

\[ x(0) = \lim_{z \to \infty} X(z). \] (4.2)

According to the \( z \)-transform definition all terms with \( z^{-n} \) vanishes as \( z \to \infty \). The term which does not depend on \( z \) follows then. It is the term with \( x(0) \).

The stationary state value of a causal signal \( x(n) \) is

\[ \lim_{n \to \infty} x(n) = \lim_{z \to 1} (z - 1)X(z). \] (4.3)

This relation follows from

\[
\mathcal{Z}\{x(n + 1)u(n)\} - \mathcal{Z}\{x(n)u(n)\} = zX(z) - x(0) - X(z)
\]

\[
\mathcal{Z}\{x(n + 1)u(n)\} - \mathcal{Z}\{x(n)u(n)\} = \lim_{N \to \infty} \left[ \sum_{n=0}^{N} x(n + 1)z^{-n} - \sum_{n=0}^{N} x(n)z^{-n} \right]
\]

\[
= \lim_{N \to \infty} [x(N + 1) - x(0)].
\]

Thus,

\[
\lim_{N \to \infty} [x(N + 1) - x(0)] = zX(z) - x(0) - X(z),
\]

produces the stationary state value (4.3).

4.3 INVERSE \( Z \)-TRANSFORM

4.3.1 Direct Power Series Expansion

Most common approach to the \( z \)-transform inversion is based on a direct expansion of the given transform into power series with respect to \( z^{-1} \) within the region of convergence. After the \( z \)-transform is expanded into series

\[ X(z) = \sum_{n=-\infty}^{\infty} X_n z^{-n} \]

the signal is identified as \( x(n) = X_n \) for \(-\infty < n < \infty \).
In general various techniques may be used to expand a function into power series. Most of the cases in signal processing, after some transformations, reduce to a simple form of an infinite geometric series
\[ \frac{1}{1-q} = 1 + q + q^2 + ... = \sum_{n=0}^{\infty} q^n \]
for \( |q| < 1 \).

**Example 4.4.** For the \( z \)-transform
\[ X(z) = \frac{1}{1 - \frac{1}{2}z^{-1}} + \frac{1}{1 - \frac{1}{3}z} \]
identify possible regions of convergence and find the inverse \( z \)-transform for each of them.

★ Obviously the \( z \)-transform has the poles \( z_1 = \frac{1}{2} \) and \( z_2 = \frac{1}{3} \). Since there are no poles in the region of convergence there are three possibilities to define the region of convergence: 1) \( |z| > \frac{1}{2} \), 2) \( \frac{1}{3} < |z| < \frac{1}{2} \), and 3) \( |z| < \frac{1}{3} \). The signals are obtained by using power series expansion for each case.

1) For the region of convergence \( |z| > \frac{1}{2} \) the \( z \)-transform should be written in the form
\[ X(z) = \frac{1}{1 - \frac{1}{2}z^{-1}} + \frac{1}{1 - \frac{1}{3}z} \]
Now we have two sums of the geometric series
\[ \frac{1}{1 - \frac{1}{2}z^{-1}} = \sum_{n=0}^{\infty} \left( \frac{1}{2} \right)^n z^{-n} \text{ for } \left| \frac{1}{2} \right| < 1 \text{ or } |z| > \frac{1}{2} \]
\[ \frac{1}{1 - \frac{1}{3}z} = \sum_{n=0}^{\infty} \left( \frac{1}{3} \right)^n z^{-n} \text{ for } \left| \frac{1}{3} \right| < 1 \text{ or } |z| > \frac{1}{3} \]
Both of these sums converge for \( |z| > \frac{1}{2} \). The resulting power series expansion of \( X(z) \) is
\[ X(z) = \sum_{n=0}^{\infty} \frac{1}{2^n} z^{-n} - \sum_{n=0}^{\infty} \frac{1}{3^n} z^{-n} \]
\[ = \sum_{n=0}^{\infty} \frac{1}{2^n} z^{-n} - \sum_{n=1}^{\infty} \frac{1}{3^n} z^{-n} \]
The inverse \( z \)-transform, for this region of convergence, is
\[ x(n) = \frac{1}{2^n} u(n) - \frac{1}{3^n} u(n - 1). \]
2) For $1/3 < |z| < 1/2$ the $z$-transform should be written in the form

$$X(z) = \frac{-2z}{1 - 2z} + \frac{1}{-3z(1 - \frac{1}{3})}.$$ 

The corresponding geometric series are

$$\frac{1}{1 - 2z} = \sum_{n=0}^{\infty} (2z)^n = \sum_{n=-\infty}^{0} 2^{-n}z^{-n} \quad \text{for} \quad |2z| < 1 \text{ or } |z| < \frac{1}{2}$$

$$\frac{1}{1 - \frac{1}{3}z} = \sum_{n=0}^{\infty} \left(\frac{1}{3}z\right)^n = \sum_{n=0}^{\infty} \frac{1}{3^n}z^{-n} \quad \text{for} \quad \left|\frac{1}{3}z\right| < 1 \text{ or } |z| > \frac{1}{3}.$$ 

They converge for $1/3 < |z| < 1/2$. The resulting power series expansion is

$$X(z) = -2z \sum_{n=-\infty}^{0} 2^{-n}z^{-n} - \frac{1}{3z} \sum_{n=0}^{\infty} \frac{1}{3^n}z^{-n}$$

$$= - \sum_{n=-\infty}^{0} \frac{1}{2^n}z^{-n} - \sum_{n=1}^{\infty} \frac{1}{3^n}z^{-n}.$$ 

The inverse $z$-transform for this region of convergence is

$$x(n) = -\frac{1}{2^n}u(-n - 1) - \frac{1}{3^n}u(n - 1).$$

3) For $|z| < 1/3$ we can write

$$X(z) = \frac{-2z}{1 - 2z} + \frac{1}{1 - 3z}.$$ 

The corresponding geometric series are

$$\frac{1}{1 - 2z} = \sum_{n=0}^{\infty} (2z)^n = \sum_{n=-\infty}^{0} 2^{-n}z^{-n} \quad \text{for} \quad |2z| < 1 \text{ or } |z| < \frac{1}{2}$$

$$\frac{1}{1 - 3z} = \sum_{n=0}^{\infty} (3z)^n = \sum_{n=-\infty}^{0} 3^{-n}z^{-n} \quad \text{for} \quad |3z| < 1 \text{ or } |z| < \frac{1}{3}.$$ 

Both series converge for $|z| < 1/3$. The expansion is

$$X(z) = -2z \sum_{n=-\infty}^{0} 2^{-n}z^{-n} + \sum_{n=-\infty}^{0} 3^{-n}z^{-n}$$

$$= - \sum_{n=-\infty}^{0} \frac{1}{2^n}z^{-n} + \sum_{n=-\infty}^{0} \frac{1}{3^n}z^{-n}.$$ 

The inverse $z$-transform, in this case, is

$$x(n) = -\frac{1}{2^n}u(-n - 1) + \frac{1}{3^n}u(-n).$$
Example 4.5. For the z-transform
\[ X(z) = e^{a/z} \]
identify the region of convergence and find the inverse z-transform.

★ Expanding \( e^{a/z} \) into a complex Taylor (Laurant) series
\[ X(z) = e^{a/z} = 1 + \frac{a}{2!}(a/z)^2 + \frac{1}{3!}(a/z)^3 + ... \]
follows
\[ x(n) = \delta(n) + a\delta(n-1) + \frac{1}{2!}a^2\delta(n-2) + \frac{1}{3!}a^3\delta(n-3) + \]
\[ = a^n \frac{1}{n!} u(n). \]
The series converges for any \( z \) except \( z = 0 \). \qed

Example 4.6. For the z-transform
\[ X(z) = \frac{z^2 + 1}{(z - 1/2)(z^2 - 3z/4 + 1/8)} \]
find the signal \( x(n) \) if the region of convergence is \( |z| > 1/2 \).

★ The denominator of \( X(z) \) will be rewritten in the form
\[ X(z) = \frac{z^2 + 1}{(z - 1/2)(z - z_1)(z - z_2)} = \frac{z^2 + 1}{(z - 1/2)^2(z - 1/4)} \]
where \( z_1 = 1/2 \) and \( z_2 = 1/4 \). Writing \( X(z) \) in the form of partial fractions
\[ X(z) = \frac{A}{(z - \frac{1}{2})^2} + \frac{B}{z - \frac{1}{2}} + \frac{C}{z - \frac{1}{4}} \]
the coefficients \( A, B, \) and \( C \) follow from
\[ \frac{(z^2 + 1)}{(z - \frac{1}{2})^2(z - \frac{1}{4})} = \frac{A(z - \frac{1}{4}) + B(z - \frac{1}{2})(z - \frac{1}{4}) + C(z - \frac{1}{2})^2}{(z - \frac{1}{2})^2(z - \frac{1}{4})} \]
or from
\[ (z^2 + 1) = A(z - \frac{1}{4}) + B(z - \frac{1}{2})(z - \frac{1}{4}) + C(z - \frac{1}{2})^2. \] (4.4)
For \( z = 1/4 \) we get \( 17/16 = C/16 \) or \( C = 17 \). Value of \( z = 1/2 \) gives
\[ \frac{1}{4} + 1 = A(\frac{1}{2} - \frac{1}{4}), \]
and $A = 5$ is obtained. Finally if the highest order coefficients in the relation (4.4) with $z^2$ are equated
\[ z^2 = Bz^2 + Cz^2 \]
we get $1 = B + C$, producing $B = -16$. The $z$-transform is
\[ X(z) = \frac{5}{(z - \frac{1}{2})^2} + \frac{-16}{z - \frac{1}{2}} + \frac{17}{z - 1/4}. \]
For the region of convergence $|z| > 1/2$ and a parameter $|a| \leq 1/2$ holds
\[ \frac{1}{z - a} = \frac{1}{z(1 - \frac{a}{z})} = z^{-1}(1 + az^{-1} + a^2z^{-2} + \ldots) = \sum_{n=1}^{\infty} a^{n-1}z^{-n}. \]
Differentiating both sides of the previous equation with respect to $a$ we get
\[ \frac{d}{da} \left( \frac{1}{z - a} \right) = \frac{1}{(z - a)^2} = \sum_{n=2}^{\infty} (n-1)a^{n-2}z^{-n} \]
Using this relation with $a = 1/2$ the inverse $z$-transform of $X(z)$ is
\[ x(n) = 5\frac{n-1}{2^{n-2}}u(n-2) - 16\frac{1}{2^{n-1}}u(n-1) + 17\frac{1}{4^{n-1}}u(n-1). \]
Note: In general, the relation
\[ \frac{1}{(z-a)^{m+1}} = \frac{1}{m!} \frac{d^m}{da^m} \left( \frac{1}{z-a} \right) = \frac{m!}{m!} \frac{1}{m!} \frac{d^m}{da^m} \left( \sum_{n=1}^{\infty} a^{n-1}z^{-n} \right) = \frac{(n-1)(n-2)\ldots(n-m)}{m!} \sum_{n=1}^{\infty} a^{n-m-1}z^{-n} \]
produces the inverse $z$-transform
\[ x(n) = \frac{(n-1)(n-2)\ldots(n-m)}{m!} a^{n-m-1}u(n) \]
\[ = \frac{(n-1)(n-2)\ldots(n-m)}{m!} a^{n-m-1}u(n-m-1) \]
\[ = \binom{n}{m} a^{n-m-1}u(n-m-1). \]

4.3.2 Theorem of Residues Based Inversion

In general the inversion is calculated by using the Cauchy relation from the complex analysis
\[ \frac{1}{2\pi j} \oint_C z^{m-1}dz = \delta(m), \]
where $C$ is any closed contour line within the region of convergence. The complex plane origin is within the contour. By multiplying both sides of $X(z)$ by $z^{m-1}$, after integration along the closed contour within the region of convergence we get

$$\frac{1}{2\pi j} \oint_C z^{m-1} X(z) dz = \sum_{n=-\infty}^{\infty} \frac{1}{2\pi j} \oint_C z^{m-1} x(n) z^{-n} dz = x(m).$$

The integral is calculated by using the theorem of residues

$$x(n) = \frac{1}{2\pi j} \oint_C z^{n-1} X(z) dz = \sum_{z_i} \left\{ \frac{1}{(k-1)!} \frac{d^{(k-1)} [z^{n-1} X(z) (z - z_i)^k]}{dz^{k-1}} \right|_{z=z_i} \right\},$$

where $z_i$ are the poles of $z^{n-1} X(z)$ within the integration contour $C$ that is in the region of convergence and $k$ is the pole order. If the signal is causal, $n \geq 0$, and all poles of $z^{n-1} X(z)$ within contour $C$ are simple (first-order poles with $k = 1$) then, for a given instant $n$,

$$x(n) = \sum_{z_i} \left\{ [z^{n-1} X(z) (z - z_i)]_{z=z_i} \right\}. \quad (4.5)$$

**Example 4.7.** For the $z$-transform

$$X(z) = \frac{2z + 3}{(z - 1/2)(z - 1/4)}$$

find a causal signal $x(n)$.

- According to the residuum theorem for $n \geq 1$

$$x(n) = \sum_{z_i} \left\{ [z^{n-1} X(z) (z - z_i)]_{z=z_i} \right\}$$

$$= \frac{z^{n-1}(2z + 3)}{(z - 1/2)(z - 1/4)} (z - \frac{1}{2})_{z=1/2} + \frac{z^{n-1}(2z + 3)}{(z - 1/2)(z - 1/4)} (z - \frac{1}{4})_{z=1/4}$$

$$= \frac{1}{4} + \frac{1}{4} = 16 \frac{1}{2^n - 1} - 14 \frac{1}{4^n - 1}.$$

For $n = 0$ additional pole at $z = 0$ exists

$$x(0) = \frac{z^{-1}(2z + 3)}{(z - 1/2)(z - 1/4)} z|_{z=0} + \frac{z^{-1}(2z + 3)}{(z - 1/2)(z - 1/4)} (z - \frac{1}{2})_{z=1/2}$$

$$+ \frac{z^{-1}(2z + 3)}{(z - 1/2)(z - 1/4)} (z - \frac{1}{4})_{z=1/4} = 0.$$
An easy way to get \( x(0) \) is \( x(0) = \lim_{z \to \infty} X(z) \).

The resulting inverse \( z \)-transform is

\[
x(n) = 16 \frac{1}{2n-1} u(n-1) - 14 \frac{1}{4n-1} u(n-1).
\]

It has been assumed that the signal is causal. Using the theorem of residuum prove that \( x(n) = 0 \) for \( n < 0 \) with \( |z| > 1/2 \).

Hint: Since for each \( n < 0 \) there is a pole at \( z = 0 \) of the order \( n + 1 \), to avoid different derivatives for each \( n \) we can make a substitution of variables \( z = 1/p \), with \( dz = -dp/p^2 \). New region of convergence in the complex plane \( p \) will be \( p < 2 \). All poles are now outside this region and outside the integration contour, producing the zero-valued integral.

\[\square\]

### 4.4 DISCRETE SYSTEMS AND THE \( Z \)-TRANSFORM

For a linear time-invariant discrete system described by

\[
y(n) = x(n) * h(n) = \sum_{m=\infty}^{-\infty} x(m)h(n-m)
\]

the \( z \)-transform is given by

\[
Y(z) = X(z)H(z).
\]

The output signal \( z \)-transform is obtained by multiplying the input signal \( z \)-transform by the transfer function

\[
H(z) = \sum_{n=\infty}^{-\infty} h(n)z^{-n}.
\]

It is possible to relate two important properties of a system with the transfer function properties.

The system is stable if

\[
\sum_{m=\infty}^{-\infty} |h(m)| < \infty.
\]

It means that the \( z \)-transform exists at \( |z| = 1 \), i.e., that the circle

\[
|z| = 1
\]
belongs to the region of convergence for a stable system.

The system is causal if \( h(n) = 0 \) for \( n < 0 \). Since \( H(z) = h(0) + h(1)z^{-1} + h(2)z^{-2} + \ldots \) it is obvious that \( z \to \infty \) belongs to the region of convergence for a causal system.

From the previous two properties we can conclude that a linear time-invariant system is stable and causal if the unit circle \( |z| = 1 \) and \( z \to \infty \) belong to the region of convergence. Since there are no poles within the region of convergence one may conclude that a transfer function \( H(z) \) may correspond to a stable and causal system only if all of its poles are inside the unit circle.

**Example 4.8.** For the systems whose transfer functions are

\[
H_1(z) = \frac{1}{(z - 1/3)(z - 3/2)}, \quad |z| > 3/2
\]

\[
H_2(z) = \frac{1}{z(z - 1/3)(z - 3/2)}, \quad 1/3 < |z| < 3/2
\]

\[
H_3(z) = \frac{1}{(z - 1/3)(z - 3/4)}, \quad |z| > 3/4
\]

plot the regions of convergence and discuss the stability and causality. Find and plot the impulse response for each case.
The regions of convergence are shown in Fig. 4.3. The system described by \( H_1(z) \) is causal but not stable. The system \( H_2(z) \) is stable but not causal, while the system \( H_3(z) \) is both stable and causal. Their impulse responses are presented in Fig. 4.3 as well.

Amplitude of the frequency response (gain) of a discrete system is related to the transfer function as

\[
|H(e^{j\omega})| = |H(z)|_{z=e^{j\omega}}.
\]

Consider a discrete system whose transfer function assumes the form of a ratio of two polynomials

\[
H(z) = \frac{B_0 + B_1 z^{-1} + \ldots + B_M z^{-M}}{A_0 + A_1 z^{-1} + \ldots + A_N z^{-N}} = \frac{B_0 z_{-M}}{A_0} \frac{(z - z_{01})(z - z_{02})\ldots(z - z_{0M})}{(z - z_{p1})(z - z_{p2})\ldots(z - z_{pN})}
\]

where \( z_{0i} \) are zeros and \( z_{pi} \) are poles of the transfer function. For the amplitude of frequency response we may write

\[
|H(e^{j\omega})| = \left| \frac{B_0}{A_0} \frac{TO_1 TO_2 \ldots TO_M}{TP_1 TP_2 \ldots TP_N} \right|
\]

where \( TO_i \) are the distances from point \( T \) at a given frequency \( z = e^{j\omega} \) to zero \( O_i \) at \( z_{0i} \). Distances from point \( T \) to poles \( P_i \) at \( z_{pi} \) are denoted by \( TP_i \).

**Example 4.9.** Plot the frequency response of the causal notch filter with the transfer function

\[
H(z) = \frac{z - e^{j\pi/3}}{z - 0.95e^{j\pi/3}}
\]

The transfer functions calculation is illustrated in Fig. 4.4. Its value is

\[
|H(e^{j\omega})| = \frac{TO_1}{TP_1}
\]

where \( O_1 \) is positioned at \( z_{01} = e^{j\pi/3} \) and pole \( P_1 \) is at \( z_{p1} = 0.95e^{j\pi/3} \). For any point \( T \) at \( z = e^{j\omega} \), \( \omega \neq \pi/3 \), the distances \( TO_1 \) and \( TP_1 \) from \( T \) to \( O_1 \) and from \( T \) to \( P_1 \) are almost the same, \( TO_1 \cong TP_1 \). Then \( |H(z)|_{z=e^{j\omega}} \cong 1 \) except at \( \omega = \pi/3 \), when \( TO_1 = 0 \) and \( TP_1 \neq 0 \) resulting in \( |H(z)|_{z=e^{j\pi/3}} = 0 \). The frequency response \( |H(e^{j\omega})| \) is shown in Fig. 4.4.
4.5 DIFFERENCE EQUATIONS

An important class of discrete systems can be described by difference equations. They are obtained by converting corresponding differential equations or by describing an intrinsically discrete system relating the input and output signal in a recursive way. A general form of a linear difference equation with constant coefficients, that relates the output signal at an instant $n$ with the input signal $x(n)$ and the previous input and output samples, is

$$ y(n) + A_1 y(n-1) + \ldots + A_N y(n-N) = B_0 x(n) + B_1 x(n-1) + \ldots + B_M x(n-M). $$

4.5.1 Solution Based on the $z$-transform

The $z$-transform of the linear difference equation, assuming zero-valued initial conditions, is

$$ [1 + A_1 z^{-1} + \ldots + A_N z^{-N}] Y(z) = [B_0 + B_1 z^{-1} + \ldots + B_M z^{-M}] X(z), $$

since $Z\{x(n-i)\} = X(z)z^{-i}$ and $Z\{y(n-k)\} = Y(z)z^{-k}$. The solution $y(n)$ of the difference equation is obtained as an inverse $z$-transform of

$$ Y(z) = \frac{B_0 + B_1 z^{-1} + \ldots + B_M z^{-M}}{1 + A_1 z^{-1} + \ldots + A_N z^{-N}} X(z). $$
Example 4.10. A causal discrete system is described by the difference equation

\[ y(n) - \frac{5}{6}y(n-1) + \frac{1}{6}y(n-2) = x(n). \quad (4.6) \]

If the input signal is \( x(n) = 1/4^n u(n) \) find the output signal.

The \( z \)-transform domain form of the system is

\[ Y(z) = \frac{1}{1 - \frac{1}{2}z^{-1} + \frac{1}{6}z^{-2}} X(z). \]

The \( z \)-transform of the input signal is \( X(z) = \frac{1}{1 - \frac{1}{4}z^{-1}} \) for \( |z| > 1/4 \). The output signal \( z \)-transform is

\[ Y(z) = \frac{z^3}{(z - \frac{1}{2})(z - \frac{1}{3})(z - \frac{1}{4})}. \]

For a causal system the region of convergence is \( |z| > 1/2 \). The output signal is the inverse \( z \)-transform of \( Y(z) \). For \( n > 0 \) it is

\[
y(n) = \sum_{z_i=1/2,1/3,1/4} \left\{ [z^{n-1}Y(z)(z - z_i)]|_{z=z_i} \right\} \\
= \frac{z^{n+2}}{(z - \frac{1}{2})(z - \frac{1}{3})|_{z=1/2}} + \frac{z^{n+2}}{(z - \frac{1}{2})(z - \frac{1}{4})|_{z=1/3}} + \frac{z^{n+2}}{(z - \frac{1}{3})(z - \frac{1}{4})|_{z=1/4}} \\
= 6 \frac{1}{2^n} - \frac{8}{3^n} + \frac{3}{4^n}. \]

For \( n = 0 \) there is no pole at \( z = 0 \). Thus, the above expressions hold for \( n = 0 \) as well. The output signal is

\[ y(n) = \left[ \frac{6}{2^n} - \frac{8}{3^n} + \frac{3}{4^n} \right] u(n). \]

Note: This kind of solution assumes the initial values from the system causality and \( x(n) \) as \( y(0) = x(0) = 1 \) and \( y(1) - 5y(0)/6 = x(1) \), i.e., \( y(1) = 13/12 \).

Example 4.11. A first-order causal discrete system is described by the following difference equation

\[ y(n) + A_1y(n-1) = B_0 x(n) + B_1 x(n-1). \quad (4.7) \]

Find its impulse response and discuss its behavior in terms of the system coefficients.
For the impulse response calculation the input signal is $x(n) = \delta(n)$ with $X(z) = 1$. Then we have

$$(1 + A_1 z^{-1}) Y(z) = (B_0 + B_1 z^{-1})$$

$$Y(z) = \frac{B_0 + B_1 z^{-1}}{1 + A_1 z^{-1}}.$$ 

The pole of this system is $z = -A_1$. There are two possibilities for the region of convergence $|z| > |A_1|$ and $|z| < |A_1|$. For a causal system the region of convergence is $|z| > |A_1|$. Thus, the $z$-transform $Y(z)$ can be expanded into a geometric series with $q = A_1 z^{-1} = (A_1 / z) < 1$

$$Y(z) = \left( B_0 + B_1 z^{-1} \right) \left( 1 - A_1 z^{-1} + A_1^2 z^{-2} - A_1^3 z^{-3} + ... + (-A_1 z^{-1})^n + ... \right)$$

$$= B_0 + B_0 \sum_{n=1}^{\infty} (-A_1)^n z^{-n} + B_1 \sum_{n=1}^{\infty} (-A_1)^{(n-1)} z^{-n}$$

with

$$y(n) = B_0 \delta(n) + (-A_1)^{n-1}(-A_1 B_0 + B_1)u(n-1).$$

We can conclude that, in general, the impulse response has an infinite duration for any $A_1 \neq 0$. It is a result of the recursive relation between the output $y(n)$ and its previous value(s) $y(n-1)$. This kind of systems are referred to as **infinite impulse response (IIR) systems or recursive systems**. If the value of coefficient $A_1$ is $A_1 = 0$ then there is no recursion and

$$y(n) = B_0 \delta(n) + B_1 \delta(n-1).$$

Then we have a system with a **finite impulse response (FIR)**. This kind of system produces an output to a signal $x(n)$ as

$$y(n) = B_0 x(n) + B_1 x(n-1).$$

They are called **moving average (MA) systems**. Systems without recursion are always stable since a finite sum of finite signal values is always finite.

Systems that would contain only $x(n)$ and the output recursions, in this case,

$$y(n) + A_1 y(n-1) = B_0 x(n)$$

are **auto-regressive (AR) systems or all pole systems**. This kind of systems could be unstable, due to recursion. In our case the system is obviously unstable if $|A_1| > 1$. Systems (4.7) are in general **auto-regressive moving average (ARMA) systems**.
If the region of convergence were \(|z| < |A_1|\) then the function \(Y(z)\) would be expanded into series with \(q = z/A_1 < 1\) as

\[
Y(z) = \frac{B_0 + B_1 z^{-1}}{A_1 z^{-1} (z/A_1 + 1)} = \left( \frac{B_0 z + B_1}{A_1} \right) \sum_{n=0}^{\infty} (-A_1^{-1} z)^n
\]

\[
= B_0 \sum_{n=-\infty}^{0} (-A_1)^{(n-1)} z^{-(n-1)} + \frac{B_1}{A_1} \sum_{n=-\infty}^{0} (-A_1)^n z^{-n}
\]

\[
= B_0 \sum_{n=-\infty}^{-1} (-A_1)^n z^{-n} + \frac{B_1}{A_1} \sum_{n=-\infty}^{0} (-A_1)^n z^{-n}
\]

with

\[
y(n) = B_0 (-A_1)^n u(-n-1) + \frac{B_1}{A_1} (-A_1)^n u(-n).
\]

This system would be stable if \(|1/A_1| < 1\) and unstable if \(|1/A_1| > 1\), having in mind that \(y(n)\) is nonzero for \(n < 0\). This is an anticausal system since it has impulse response satisfying \(h(n) = 0\) for \(n \geq 1\).

Here, we have just introduced the notions. These systems will be considered in Chapter 5 in details.

### 4.5.2 Solution of Difference Equations in the Time Domain

A direct way to solve a linear difference equation with constant coefficients of the form

\[
y(n) + A_1 y(n - 1) + \ldots + A_N y(n - N) = x(n)
\]

in the time domain will be described next.

A homogeneous part of this difference equation is

\[
y(n) + A_1 y(n - 1) + \ldots + A_N y(n - N) = 0.
\]

Solution for the homogeneous equation is of the form

\[
y_i(n) = C_i \lambda_i^n.
\]

Replacing \(y_i(n)\) into (4.9), the characteristic polynomial equation follows

\[
C_i \lambda_i^n + C_i A_1 \lambda_i^{n-1} + \ldots + C_i A_N \lambda_i^{n-N} = 0,
\]

or

\[
\lambda_i^N + A_1 \lambda_i^{N-1} + \ldots + A_N = 0.
\]

This is a polynomial of the \(N\)th order. In general, it has \(N\) solutions \(\lambda_i, i = 1, 2, \ldots, N\). All functions \(y_i(n) = \lambda_i^n, i = 1, 2, \ldots, N\) are the solutions of
equation (4.9). Since the equation is linear, a linear combination of these solutions,
\[ y_h(n) = \sum_{i=1}^{N} C_i \lambda_i^n \]
is also a solution of the homogeneous equation (4.9). This solution is called homogeneous part of the solution of (4.8).

Next a particular solution \( y_p(n) \), corresponding to the form of input signal \( x(n) \), should be found using the form of \( x(n) \). The solution of equation (4.8) is then
\[ y(n) = y_h(n) + y_p(n). \]
The constants \( C_i \), \( i = 1, 2, \ldots, N \) are calculated based on initial conditions \( y(i-1), i = 1, 2, \ldots, N \).

**Example 4.12.** Find the output of a causal discrete system
\[ y(n) - \frac{5}{6} y(n-1) + \frac{1}{6} y(n-2) = x(n) \] (4.10)
to the input signal \( x(n) = (n + 11/6) u(n) \) by solving the difference equation in the discrete-time domain. The initial conditions are \( y(0) = 1 \) and \( y(1) = 5 \).

★Solution of the homogeneous part of (4.10)
\[ y(n) - \frac{5}{6} y(n-1) + \frac{1}{6} y(n-2) = 0 \]
is of the form \( y_i(n) = C_i \lambda_i^n \). Its replacement into the equation results in the characteristic polynomial
\[ \lambda_i^2 - \frac{5}{6} \lambda_i + \frac{1}{6} = 0, \]
producing \( \lambda_1 = 1/2 \) and \( \lambda_2 = 1/3 \). The homogeneous part of the solution is
\[ y_h(n) = C_1 \frac{1}{2^n} + C_2 \frac{1}{3^n}. \]

Since \( x(n) \) is a linear function of \( n \), a particular solution is of the form \( y_p(n) = An + B \). Replacing \( y_p(n) \) into (4.10) we obtain
\[ y_p(n) - \frac{5}{6} y_p(n-1) + \frac{1}{6} y_p(n-2) = n + 11/6 \]
\[ An + B - \frac{5}{6} (An - A + B) + \frac{1}{6} (An - 2A + B) = n + 11/6, \]
and \( A = 3, B = 1 \) follow. The solution of (4.10) is a sum of homogeneous and particular solutions,
\[ y(n) = y_h(n) + y_p(n) = C_1 \frac{1}{2^n} + C_2 \frac{1}{3^n} + 3n + 1. \]
Using the initial conditions
\[ y(0) = C_1 + C_2 + 1 = 1 \]
\[ y(1) = \frac{C_1}{2} + \frac{C_2}{3} + 4 = 5 \]
the constants \( C_1 = 6 \) and \( C_2 = -6 \) follow. The final solution is
\[ y(n) = \left[ \frac{6}{2^n} - \frac{6}{3^n} + 3n + 1 \right] u(n). \]

Note: The \( z \)-transform based solution would assume \( y(0) = x(0) = \frac{11}{6} \) and \( y(1) = 5y(0)/6 + x(1) = 157/36 \). The solution with the initial conditions \( y(0) = 1 \) and \( y(1) = 5 \) could be obtained from this solution with appropriate changes of the first two samples of the input signal in order to take into account the previous system state and to produce the given initial conditions \( y(0) = 1 \) and \( y(1) = 5 \).

If multiple polynomial roots are obtained, for example \( \lambda_i = \lambda_{i+1} \), then \( y_i(n) = \lambda_i^n \) and \( y_{i+1}(n) = n\lambda_i^n \).

**Example 4.13. Goertzel algorithm:** Show that a discrete-time signal
\[ y(n) = e^{j(2\pi k_0 n / N + \phi)} \]
is a solution of the homogeneous difference equation
\[ y(n) - e^{2\pi k_0 / N} y(n - 1) = 0. \]  \hfill (4.11)

Consider a periodic signal \( x(n) \) with a period \( N \) and its DFT values \( X(k) \),
\[ x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j2\pi nk / N}. \]  \hfill (4.12)

If the signal within one of its periods, for \( 0 \leq n \leq N - 1 \), is applied as the input to the system described by difference equation (4.11) show that the output signal at \( n = N - 1 \) is equal to the DFT of signal at frequency \( k = k_0 \), i.e.,
\[ y(N - 1) = X(k_0). \]

\[ \star \] For the signal \( y(n) \) holds
\[ y(n) = e^{j(2\pi k_0 n / N + \phi)} = e^{j(2\pi k_0 (n-1+1) / N + \phi)} = e^{j(2\pi k_0 / N)} y(n - 1). \]

Consider now the case when the input signal \( x(n) \) is applied to the system. Since the system is linear, consider one component of the input signal (4.12)
\[ x_k(n) = \frac{1}{N} X(k) e^{j2\pi nk / N}, \]
for an arbitrary $0 \leq k \leq N - 1$. Then the difference equation for this input signal reads
\[
y_k(n) - e^{2\pi jk/N}y_{k}(n-1) = x_k(n)
\]
\[
Y_k(z) = \frac{Z\{x_k(n)\}}{1 - e^{2\pi jk/N}z^{-1}}.
\] (4.13)

The $z$-transform of $x_k(n)$, for $0 \leq n \leq N - 1$, is
\[
Z\{x_k(n)\} = Z\left\{\frac{1}{N}X(k)e^{2\pi jk/N}\right\}
\]
\[
= \frac{1}{N}X(k)\sum_{n=0}^{N-1}e^{2\pi jk/N}z^{-n} = \frac{1}{N}X(k)\frac{1 - e^{2\pi jk/N}}{1 - e^{2\pi jk/N}z^{-1}}.
\] (4.14)

The transform $Z\{x_k(n)\}$, for a given $k$, has zeros at
\[
z_0^N = e^{2\pi jk+2\pi j}, l = 0,1,2,...,N - 1
\] or
\[
z_0 = e^{2\pi j(k+1)/N}, l = 0,1,2,...,N - 1.
\]

Note that the zero
\[
z_0 = e^{2\pi jk/N}, \text{ obtained for } l = 0
\]
is canceled with the pole $z_p = e^{2\pi jkn/N}$ in (4.14). Therefore the remaining zeros are at
\[
z_0 = e^{2\pi j(k+1)/N}, l = 1,2,...,N - 1
\]
The output $z$-transform $Y_k(z)$, defined by (4.13), has a pole at
\[
z_p = e^{2\pi jk_0/N}
\]
- If $k \neq k_0$ then one of zeros $z_0^N = e^{2\pi j(k+1)/N}, l = 1,2,...,N - 1$ will coincide with the pole $z_p = e^{2\pi jk_0/N}$ and will cancel it. Thus for $k \neq k_0$ the function $Y_k(z)$ will not have any poles. Then
\[
y_k(N-1) = \frac{1}{2\pi j}\oint_C z^{-2}Y_k(z)dz = 0
\] (4.15)
since there are no poles, Fig.4.5.

- If $k = k_0$ then the pole at $k = k_0$ is already canceled in $Z\{x_k(n)\}$ and $z_p = e^{2\pi jk_0/N}$ remains as a pole of $Y(z)$. In this case the signal value at $n = N - 1$ is equal to the residuum of function in (4.15) at the pole $z_p = e^{2\pi jk_0/N}$, relation (4.5),
\[
y_{k_0}(N-1) = z^{N-2}Y_{k_0}(z)(z - e^{2\pi jk_0/N})\bigg|_{z=e^{2\pi jk_0/N}}
\]
\[
= z^{N-1}\frac{1}{N}X(k_0)\left\{1 - \frac{e^{2\pi jk_0z^{-N}}}{1 - e^{2\pi jk_0/N}z^{-1}}\right\}_{z=e^{2\pi jk_0/N}}
\]
\[
= \frac{1}{N}X(k_0)\lim_{z\to e^{2\pi jk_0/N}} \frac{z^N - e^{2\pi jk_0}}{z - e^{2\pi jk_0/N}} = X(k_0).
\]
Therefore the output of the system, at \( n = N - 1 \), is

\[ y_k(N - 1) = X(k)\delta(k - k_0). \]

**Note:** The difference relation

\[ y(n) - e^{j2\pi k_0 n/N}y(n - 1) = x(n) \]  \hspace{1cm} (4.16)

with the \( z \)-transform domain form

\[ Y(z) = \frac{X(z)}{1 - e^{j2\pi k_0 n/N}z^{-1}} \]

is often extended to

\[
Y(z) = \frac{X(z)}{1 - e^{j2\pi k_0 n/N}z^{-1}} - \frac{1}{1 - e^{-j2\pi k_0 n/N}z^{-1}} \frac{1}{1 - e^{-j2\pi k_0 n/N}z^{-1}} - \frac{1}{2\cos(2\pi k_0 n / N)z^{-1} + z^{-2}} X(z)
\]

In the discrete-time domain the system

\[ y(n) - 2\cos(2\pi k_0 / N)y(n - 1) + y(n - 2) = x(n) - e^{-j2\pi k_0 n / N}x(n - 1) \]  \hspace{1cm} (4.17)

is called Goertzel algorithm for the DFT calculation at a given single frequency \( X(k_0) \).

It is interesting to note that the computation of (4.17) is more efficient than the computation of (4.16). For the calculation of (4.16), for one \( k_0 \), we need one complex multiplication (4 real multiplications) and one complex addition (2 real additions). For \( N \) instants and one \( k_0 \) we need \( 4N \) real multiplications and \( 2N \) real additions. For the calculation of (4.17) we can use
linear property and calculate only
\[ y_1(n) - 2\cos(2\pi k_0/N)y_1(n-1) + y_1(n-2) = x(n) \] (4.18)
at each instant. It requires a multiplication of complex signal with a real
coefficient. It means 2 real multiplications for each instant or 2N in total for
N instants. The resulting output, at the instant \( N - 1 \), is
\[ y(N - 1) = T\{x(N - 1)\} - e^{-j2\pi k_0(N-1)/N}T\{x(N - 1)\} \]
\[ = y_1(N) - e^{j2\pi k_0}y_1(N - 1). \]
It requires just one additional complex multiplication for the last instant
and for one frequency. The total number of multiplications is \( 2N + 4 \). It
is reduced with respect to the previously needed \( 4N \) real multiplications.
The total number of additions is \( 4N + 2 \). It is increased. However the time
needed for a multiplication is much longer than the time needed for an
addition. Thus, the overall efficiency is improved. The efficiency is even more
improved having in mind that (4.18) is the same for calculation of \( X(k_0) \) and
\( X(-k_0) = X(N - k_0) \).

4.6 RELATION OF THE Z-TRANSFORM TO OTHER TRANSFORMS

By sampling a signal \( x(t) \), the Laplace transform integral can be approxi-
mated by a sum
\[ X(s) = \int_{-\infty}^{\infty} x(t)e^{-st}dt \cong \sum_{n=-\infty}^{\infty} x(n\Delta t)e^{-sn\Delta t}\Delta t = \sum_{n=-\infty}^{\infty} x(n)e^{-sn\Delta t} \]
with \( x(n) = x(n\Delta t)\Delta t \). Comparing this relation with the z-transform defi-
nition we can conclude that the Laplace transform of \( x(t) \) corresponds to the
z-transform of its samples with
\[ z = \exp(s\Delta t), \]
that is,
\[ X(s) \leftrightarrow X(z)|_{z=\exp(s\Delta t)}. \] (4.19)

A point \( s = \sigma + j\Omega \) from the Laplace domain maps into the point
\( z = re^{j\omega} \) with \( r = e^{\sigma\Delta t} \) and \( \omega = \Omega\Delta t \). Points from the left half-plane in the
\( s \) domain, \( \sigma < 0 \), map to the interior of unit circle in the \( z \) domain, \( r < 1 \).
According to the sampling theorem, for the Laplace transform of discrete-time signal holds \( X(s)|_{s=0} = X(j\Omega) = X(j(\Omega + 2k\pi/\Delta t)) \).

The Fourier transform of a discrete-time signal is

\[
X(e^{j\omega}) = X(z)|_{z=e^{j\omega}} = \sum_{n=-\infty}^{\infty} x(n)z^{-n} |_{z=e^{j\omega}}.
\]

**Example 4.14.** A causal discrete-time signal \( x(n) \) has the Fourier transform \( X(e^{j\omega}) \). Write its \( z \)-transform in terms of the Fourier transform of the discrete-time signal, i.e., write the \( z \)-transform value based on its values on the unit circle.

\[
\text{\star} \text{The signal can be expressed in term of its Fourier transform as}
\]

\[
x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega n} d\omega
\]

\[
X(z) = \sum_{n=0}^{\infty} x(n)z^{-n} = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega}) \sum_{n=0}^{\infty} e^{j\omega n} z^{-n} d\omega
\]

\[
= \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{X(e^{j\omega})}{1-e^{\omega}z^{-1}} d\omega,
\]

for \( |z| > 1 \).

The DFT of discrete-time signal with \( N \) nonzero samples is

\[
X(k) = X(e^{j\omega})|_{\omega=2\pi k/N} = X(z)|_{z=e^{2\pi k/N}} = \sum_{n=0}^{N-1} x(n)z^{-n} |_{z=e^{2\pi k/N}}.
\]

**Example 4.15.** Consider a discrete-time signal with \( N \) samples different from zero within \( 0 \leq n \leq N-1 \). Show that all values of \( X(z) \), for any \( z \), can be calculated based on its \( N \) samples on the unit circle in the \( z \)-plane.

\[
\text{\star} \text{If the signal has \( N \) nonzero samples, then it can be expressed in term of its DFT as}
\]

\[
X(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi nk/N} \text{ and } x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k)e^{2\pi nk/N}.
\]

Thus, the \( z \)-transform of \( x(n) \), using only the values of the IDFT where the original signal is nonzero, \( 0 \leq n \leq N-1 \),

\[
X(z) = \frac{1}{N} \sum_{k=0}^{N-1} \sum_{n=0}^{N-1} X(k)e^{2\pi nk/N} z^{-n} = \frac{1}{N} \sum_{k=0}^{N-1} \frac{1-z^{-N}e^{2\pi k}}{1-z^{-1}e^{2\pi k/N}} X(k)
\]
with $X(k) = X(z)$ at $z = \exp(j2\pi k/N), k = 0,1,2,\ldots,N-1$.

For a periodic signal, including all periods in the $z$-transform calculation, holds

$$X(z) = \frac{1}{N} \sum_{k=0}^{N-1} \sum_{n=0}^{\infty} X(k) e^{j2\pi nk/N} z^{-n} = \frac{1}{N} \sum_{k=0}^{N-1} \frac{1}{1 - z^{-1} e^{j2\pi k/N}} X(k).$$

4.7 PROBLEMS

Problem 4.1. Find the $z$-transform and the region of convergence for the following signals:

(a) $x(n) = \delta(n - 2)$,

(b) $x(n) = a^{|n|} u(n)$,

(c) $x(n) = \frac{1}{2\pi} u(n) + \frac{1}{2\pi} u(n)$

Problem 4.2. Find the $z$-transform and the region of convergence for the following signals:

(a) $x(n) = \delta(n + 1) + \delta(n) + \delta(n - 1)$,

(b) $x(n) = \frac{1}{2\pi} [u(n) - u(n-10)]$.

Problem 4.3. Using the $z$-transfrom property that

$$Y(z) = -z \frac{dX(z)}{dz}$$

corresponds to

$$y(n) = nx(n) u(n)$$
in the discrete-time domain, with the same region of convergence for $X(z)$ and $Y(z)$, find a causal signal whose $z$-transform is

(a) $X(z) = e^{a/z}, \ |z| > 0$.
(b) $X(z) = \ln(1 + az^{-1}), \ |z| > |a|$.

**Problem 4.4.** (a) How the $z$-transform of $x(-n)$ is related to the $z$-transform of $x(n)$?
(b) If the signal $x(n)$ is real-valued show that its $z$-transform satisfies $X(z) = X^*(z^*)$.

**Problem 4.5.** If $X(z)$ is the $z$-transform of a signal $x(n)$ find the $z$-transform of

$$y(n) = \sum_{k=\infty}^{\infty} x(k)x(n+k).$$

**Problem 4.6.** Find the inverse $z$-transform of

$$X(z) = \frac{1}{2 - 3z}, \ |z| > \frac{2}{3}.$$ 

**Problem 4.7.** The $z$-transform of a causal signal $x(n)$ is

$$X(z) = \frac{z + 1}{(2z - 1)(3z + 2)}.$$ 

Find the signal $x(n)$.

**Problem 4.8.** The transfer function of a discrete system is

$$H(z) = \frac{3 - \frac{5}{6}z^{-1}}{(1 - \frac{1}{4}z^{-1})(1 - \frac{1}{3}z^{-1})}.$$ 

Find the impulse response if:
(a) System is stable,
(b) Region of convergence is $\frac{1}{4} < |z| < \frac{1}{3}$,
(c) System is anticausal.

**Problem 4.9.** For the $z$-transform

$$H(z) = \frac{1}{(1 - 4z)(\frac{1}{4} - \sqrt{3}z + z^2)}$$

identify possible regions of convergence. In each case comment stability and causality of the system whose transfer function is $H(z)$. What is the output of the stable system to the input $x(n) = 2\cos(n\pi/2)$?
Problem 4.10. Find the impulse response of a causal system whose transfer function is
\[ H(z) = \frac{z + 2}{(z - 2)z^2}. \]

Problem 4.11. Find the inverse z-transform of
\[ X(z) = \frac{z^2}{z^2 + 1}. \]

Problem 4.12. The system is described by a difference equation
\[ y(n) - y(n - 1) + \frac{5}{16}y(n - 2) - \frac{1}{16}y(n - 3) = 3x(n) - \frac{5}{4}x(n - 1) + \frac{3}{16}x(n - 2). \]
Find the impulse response of a causal system.

Problem 4.13. Show that the system defined by
\[ y(n) = x(n) - \frac{3}{4}x(n - 1) + \frac{1}{8}x(n - 2) \]
has a finite output duration for an infinite duration input \( x(n) = 1/4^n u(n) \).

Problem 4.14. A linear time-invariant system has impulse response
\[ h(n) = 1/3^n u(n). \]
Using the z-transform find the output to the input signal \( x(n) = u(n) - u(n - 6) \).

Problem 4.15. Find the output of a causal discrete system
\[ y(n) - \frac{11}{6}y(n - 1) + \frac{1}{2}y(n - 2) = 2x(n) - \frac{3}{2}x(n - 1) \]
if the input signal is \( x(n) = \delta(n) - \frac{3}{2}\delta(n - 1) \).

Problem 4.16. Solve the difference equation using the z-transform
\[ x(n + 2) + 3x(n + 1) + 2x(n) = 0 \]
with the initial condition \( x(0) = 0 \) and \( x(1) = 1 \). Signal \( x(n) \) is causal.
**Problem 4.17.** Solve the difference equation

\[ x(n + 1) = x(n) + a^n \]

using the z-transform with the initial condition \( x(0) = 0 \).

**Problem 4.18.** Find the output of a causal discrete system

\[ y(n) - \frac{\sqrt{2}}{2} y(n - 1) + \frac{1}{4} y(n - 2) = x(n) \quad (4.20) \]

to the input signal \( x(n) = \frac{1}{3^n} u(n) \) by a direct solution of the differential equation in the discrete-time domain and by using the z-transform. The initial conditions are \( y(n) = 0 \) for \( n < 0 \).

**Problem 4.19.** The first backward difference is defined as

\[ \nabla x(n) = x(n) - x(n - 1), \]

and the \( m \)th backward difference is defined by

\[ \nabla^m x(n) = \nabla^{m-1} x(n) - \nabla^{m-1} x(n - 1). \]

The first forward difference is

\[ \Delta x(n) = x(n + 1) - x(n), \]

with the \( m \)th forward difference being

\[ \Delta^m x(n) = \Delta^{m-1} x(n + 1) - \Delta^{m-1} x(n). \]

Find the z-transforms of these differences.

**Problem 4.20.** Based on the poles-zero geometry plot the amplitude of the frequency response of system

\[ y(n) = x(n) - \sqrt{2} x(n - 1) + x(n - 2) + r \sqrt{2} y(n - 1) - r^2 y(n - 2) \]

for \( r = 0.99 \). Based on the frequency response, find approximative values of the output signal if the input is a continuous-time signal

\[ x(t) = 2 \cos(10 \pi t) - \sin(15 \pi t) + 0.5 e^{20 \pi t} \]

sampled at \( \Delta t = 1/60 \).
Problem 4.21. Plot the frequency response of the discrete system (comb filter)

\[ H(z) = \frac{1 - z^{-N}}{1 - rz^{-N}} \]

with \( r = 0.9999 \) and \( r^{1/N} \approx 1 \). Show that this system has the same transfer function as

\[ H(z) = \frac{(1 - z^{-2})}{(1 - r^2z^{-2})} \left( \frac{N/2-1}{1} \prod_{k=1}^{N/2-1} \frac{1 - 2 \cos(2k\pi/N)z^{-1} + z^{-2}}{1 - 2r \cos(2k\pi/N)z^{-1} + z^{-2}} \right). \]

4.8 SOLUTIONS

Solution 4.1. (a) The \( z \)-transform is

\[ X(z) = \sum_{n=-\infty}^{\infty} \delta(n-2)z^{-n} = z^{-2} \]

for any \( z \neq 0 \).

(b) For this signal

\[ X(z) = \sum_{n=-\infty}^{\infty} a^{|n|}z^{-n} = \sum_{n=-\infty}^{-1} a^{-n}z^{-n} + \sum_{n=0}^{\infty} a^n z^{-n} = \frac{(1-a^2)z}{(1-az)(z-a)} \]

for \( |z| < 1/a \) and \( |z| > a \). If \( |a| < 1 \) then the region of convergence is \( a < |z| < 1/a \).

(c) In this case

\[ X(z) = \sum_{n=0}^{\infty} \frac{1}{2^n}z^{-n} + \sum_{n=0}^{\infty} \frac{1}{3^n}z^{-n} = \frac{1}{1 - \frac{1}{2}z^{-1}} + \frac{1}{1 - \frac{1}{3}z^{-1}} \]

\[ X(z) = \frac{2 - \frac{5}{6}z^{-1}}{(1 - \frac{1}{2}z^{-1})(1 - \frac{1}{3}z^{-1})} = \frac{z(2z - \frac{5}{6})}{(z - \frac{1}{2})(z - \frac{1}{3})} \]

for \( |z| > 1/2 \) and \( |z| > 1/3 \). The region of convergence is \( |z| > 1/2 \).

Solution 4.2. (a) The \( z \)-transform is

\[ X(z) = \sum_{n=-\infty}^{\infty} (\delta(n+1) + \delta(n) + \delta(n-1))z^{-n} = \]

\[ = z + 1 + z^{-1} = z + 1 + \frac{1}{z}. \]
The region of convergence excludes $z = 0$ and $z \to \infty$.

(b) For $x(n) = \frac{1}{2^n} [u(n) - u(n - 10)]$ we know that

$$u(n) - u(n - 10) = \begin{cases} 1, & n = 0, 1, \ldots, 9 \\ 0, & \text{elsewhere.} \end{cases}$$

The $z$-transform is

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n} = \sum_{n=0}^{9} \frac{1}{2^n}z^{-n} = \sum_{n=0}^{9} (2z)^{-n} = \frac{1 - (2z)^{-10}}{1 - (2z)^{-1}} =$$

$$= \frac{z^{-10} \left( \frac{1}{2} \right)^{10}}{z - \frac{1}{2}} = \frac{z^{10} - \left( \frac{1}{2} \right)^{10}}{z \left( z - \frac{1}{2} \right)}$$

The expression for $X(z)$ is written in this way in order to find the region of convergence, observing the zero-pole locations in the $z$-plane, Fig. 4.7. Poles are at $z_{p1} = 0$ and $z_{p2} = 1/2$. Zeros are $z_{0i} = e^{j2\pi i/10}/2$, Fig. 4.7. Since the $z$-transform has a zero at $z_0 = 1/2$, it will cancel out the pole $z_{p2} = 1/2$. The resulting region of convergence will include the whole $z$ plane, except the point at $z = 0$.

**Solution 4.3.** (a) For $X(z) = e^{a/z}$ holds

$$-z \frac{dX(z)}{dz} = z \frac{a}{z^2} e^{a/z} = \frac{a}{z} X(z)$$
The inverse \( z \)-transform of left and right side of this equation is

\[
nx(n)u(n) = ax(n - 1)u(n)
\]
since \( \mathcal{Z}[nx(n)] = -z \frac{dX(z)}{dz} \) and \( z^{-1}X(z) = \mathcal{Z}[x(n - 1)] \). It means that

\[
x(n) = \frac{a}{n}x(n - 1)
\]
for \( n > 0 \). According to the initial value theorem

\[
x(0) = \lim_{z \to \infty} X(z) = 1.
\]
It means that

\[
x(1) = a, \ x(2) = \frac{a^2}{2}, \ x(3) = \frac{a^3}{2 \cdot 3}, \ldots
\]
or

\[
x(n) = \frac{a^n}{n!}u(n).
\]

(b) For \( X(z) = \ln(1 + az^{-1}) \)

\[
Y(z) = -z \frac{dX(z)}{dz} = -z \frac{d\ln(1 + az^{-1})}{dz} = z \frac{az^{-2}}{1 + az^{-1}} = \frac{az^{-1}}{1 + az^{-1}}.
\]

Therefore

\[
\mathcal{Z}[nx(n)] = -z \frac{dX(z)}{dz} = \frac{az^{-1}}{1 + az^{-1}}
\]

\[
nx(n) = a(-a)^{n-1}u(n - 1),
\]
producing

\[
x(n) = \frac{(-a)^n}{n}u(n - 1).
\]

**Solution 4.4.** (a) The \( z \)-transform of signal \( x(-n) \) is

\[
X_1(z) = \sum_{n=\infty}^{\infty} x(-n)z^{-n}.
\]

With a substitution \(-n = m\) it follows

\[
X_1(z) = \sum_{m=\infty}^{\infty} x(n)z^m = X(1/z).
\]
The region of convergence is complementary to the one of the original signal. If the region of convergence for \( x(n) \) is \( |z| > a \), then the region of convergence for \( x(-n) \) is \( |z| < a \).

(b) For a real-valued signal holds \( x^*(n) = x(n) \). Then we can write \( X^*(z^*) \) as

\[
X^*(z^*) = \sum_{n=-\infty}^{\infty} x^*(n) ((z^*)^{-n})^*.
\]

Since \((z^*)^{-n} = (z^{-n})^*\) we get

\[
X^*(z^*) = \sum_{n=-\infty}^{\infty} x^*(n) z^{-n} = \sum_{n=-\infty}^{\infty} x(n) z^{-n} = X(z),
\]

for a real-valued signal \( x(n) \).

**Solution 4.5.** From

\[
Y(z) = \sum_{n=-\infty}^{\infty} y(n) z^{-n} = \sum_{n=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} x(k) x(n + k) z^{-n},
\]

using the substitution \( n + k = m \), follows

\[
Y(z) = X(z) X\left(\frac{1}{z}\right).
\]

**Solution 4.6.** A direct expansion of the given transform into power series, within the region of convergence, will be used. In order to find the signal \( x(n) \) whose \( z \)-transform is \( X(z) = \frac{1}{2 - \frac{3z}{z}} \), it should be written in a form of power series with respect to \( z^{-1} \). Since the condition \(|\frac{3z}{2}| < 1\) does not correspond to the region of convergence given in the problem formulation we have to rewrite \( X(z) \) as

\[
X(z) = -\frac{1}{3z} \frac{1}{1 - \frac{2}{3z}}.
\]

Now the condition \(|\frac{2}{3z}| < 1\), that is \(|z| > \frac{2}{3}\), corresponds to the problem formulation region of convergence. In order to obtain the inverse \( z \)-transform, write

\[
X(z) = -\frac{1}{3z} \frac{1}{1 - \frac{2}{3z}} = -\frac{1}{3z} X_1(z),
\]

where

\[
X_1(z) = \frac{1}{1 - \frac{2}{3z}}.
\]
For $X_1(z)$ holds

$$X_1(z) = \sum_{n=0}^{\infty} \left( \frac{2}{3z} \right)^n = \sum_{n=0}^{\infty} \left( \frac{2}{3} \right)^n z^{-n}.$$ 

It can be concluded that $X(z)$ can be written as

$$X(z) = -\frac{1}{3} \sum_{n=0}^{\infty} \left( \frac{2}{3} \right)^n z^{-n}.$$ 

Comparing the $z$-transform definition

$$X(z) = \sum_{n=-\infty}^{\infty} x(n) z^{-n} \quad (4.21)$$

and the last expression it follows

$$X(z) = -\frac{1}{3} \sum_{n=0}^{\infty} \left( \frac{2}{3} \right)^n z^{-n} z^{-1} =$$

$$= -\frac{1}{3} \sum_{n=0}^{\infty} \left( \frac{2}{3} \right)^n z^{-(n+1)}.$$ 

With the substitution $n \to n + 1$ we get

$$X(z) = -\frac{1}{3} \sum_{n=1}^{\infty} \left( \frac{2}{3} \right)^{n-1} z^{-n}.$$ 

Finally, comparing this result with (4.21) we get

$$x(n) = \begin{cases} 
-\frac{1}{3} \left( \frac{2}{3} \right)^{n-1}, & \text{for } n = 1, 2, ..., \infty \\
0, & \text{elsewhere,}
\end{cases}$$

or

$$x(n) = -\frac{1}{3} \left( \frac{2}{3} \right)^{n-1} u(n-1).$$

**Solution 4.7.** Since the signal is causal the region of convergence is outside the pole with the largest radius (outside the circle passing through this pole).
Poles of the z-transform are
\[ z_{p1} = \frac{1}{2} \quad \text{and} \quad z_{p2} = -\frac{2}{3}. \]

The region of convergence is \(|z| > \frac{2}{3}\).

The z-transform is
\[
X(z) = \frac{z + 1}{(2z - 1)(3z + 2)} = \frac{A}{2z - 1} + \frac{B}{3z + 2}
\]
\[ A = \frac{3}{7}, \quad B = -\frac{1}{7}. \]

The terms in \(X(z)\) should be written in such a way that they represent sums of geometric series for the given region of convergence. From the solution of the previous problem, we conclude that
\[
X(z) = \frac{A}{2z - 1} - \frac{B}{3z + 2}.
\]

Now we can write
\[
\frac{A}{2z - 1} = \sum_{n=0}^{\infty} \left( \frac{1}{2} \right)^n z^{-n} = \frac{A}{2} \sum_{n=0}^{\infty} \left( \frac{1}{2} \right)^n z^{-n-1}, \quad |z| > \frac{1}{2}
\]
and
\[
\frac{B}{3z + 2} = \sum_{n=0}^{\infty} \left( -\frac{2}{3} \right)^n z^{-n} = \frac{B}{3} \sum_{n=0}^{\infty} \left( -\frac{2}{3} \right)^n z^{-n-1}, \quad |z| > \frac{2}{3}.
\]

The z-transform, with \(m = n + 1\), assumes the form
\[
X(z) = \frac{A}{2} \sum_{m=1}^{\infty} \left( \frac{1}{2} \right)^{m-1} z^{-m} + \frac{B}{3} \sum_{m=1}^{\infty} \left( -\frac{2}{3} \right)^{m-1} z^{-m}.
\]

Replacing the values for \(A\) and \(B\) it follows
\[
X(z) = \frac{3}{7} \sum_{m=1}^{\infty} \left( \frac{1}{2} \right)^{m} z^{-m} + \frac{1}{14} \sum_{m=1}^{\infty} \left( -\frac{2}{3} \right)^{m} z^{-m}.
\]

The signal \(x(n)\) is obtained by comparing this transform with the z-transform definition,
\[
x(n) = \left( \frac{3}{7} \left( \frac{1}{2} \right)^{n} + \frac{1}{14} \left( -\frac{2}{3} \right)^{n} \right) u(n - 1).
\]
Solution 4.8. The transfer function may be written as

\[ H(z) = \frac{3 - 5z^{-1}}{(1 - \frac{1}{4}z^{-1})(1 - \frac{1}{3}z^{-1})} = \frac{A}{1 - \frac{1}{4}z^{-1}} + \frac{B}{1 - \frac{1}{3}z^{-1}} \]

with \( A = 1, \ B = 2. \)

(a) The region of convergence must contain \(|z| = 1\), for a stable system. It is \(|z| > \frac{1}{3}\).

From

\[ H(z) = \frac{1}{1 - \frac{1}{4}z^{-1}} + \frac{2}{1 - \frac{1}{3}z^{-1}} = \sum_{n=0}^{\infty} \left( \frac{1}{4} \right)^n z^{-n} + 2 \sum_{n=0}^{\infty} \left( \frac{1}{3} \right)^n z^{-n}, \quad |z| > \frac{1}{3} \text{ and } |z| > \frac{1}{4} \]

the impulse response is obtained as

\[ h(n) = (4^{-n} + 2 \times 3^{-n})u(n). \]

(b) The region of convergence is \( \frac{1}{4} < |z| < \frac{1}{3} \). The first term in \( H(z) \) is the same as in (a), since it converges for \(|z| > \frac{1}{4}\). It corresponds to the signal \( 4^{-n}u(n) \). The second term must be rewritten in such a way that its geometric series converges for \(|z| < \frac{1}{3}\). Then

\[ \frac{2}{1 - \frac{1}{3}z^{-1}} = -2 \frac{3z}{1 - 3z} = -2 \sum_{n=1}^{\infty} (3z)^n = -2 \sum_{m=-\infty}^{-1} (3z)^{-m} \quad \text{with } |z| < \frac{1}{3}. \]

Signal corresponding to this \( z \)-transform is \(-2 \times 3^{-n}u(-n - 1)\). Then the impulse response of the system with the region of convergence \( \frac{1}{4} < |z| < \frac{1}{3} \) is obtained in the form

\[ h(n) = 4^{-n}u(n) - 2 \times 3^{-n}u(-n - 1). \]

c) For an anticausal system the region of convergence is \(|z| < \frac{1}{4}\). Now the second term in \( H(z) \) is the same as in (b). For \(|z| < \frac{1}{4}\) the first term in \( H(z) \) should be written as:

\[ \frac{1}{1 - \frac{1}{4}z^{-1}} = -4z \frac{1}{1 - 4z} = -\sum_{n=1}^{\infty} (4z)^n = -\sum_{m=-\infty}^{-1} (4z)^{-m} \quad \text{with } |z| < \frac{1}{4}. \]
The signal corresponding to this term is \(-4^{-n}u(-n - 1)\). The impulse response of the antcausal discrete system with given transfer function is

\[ h(n) = -4^{-n}u(-n - 1) - 2 \times 3^{-n}u(-n - 1). \]

**Solution 4.9.** The \(z\)-transform

\[ H(z) = \frac{1}{(1 - 4z)(\frac{1}{4} - \frac{\sqrt{3}}{2}z + z^2)} \]

can be written as

\[ H(z) = \frac{1}{(1 - 4z)(z - \frac{\sqrt{3}}{4} + j\frac{1}{4})(z - \frac{\sqrt{3}}{4} - j\frac{1}{4})} \]

with poles \(z_1 = 1/4, z_2 = \frac{\sqrt{3}}{4} - j\frac{1}{4}, \) and \(z_3 = \frac{\sqrt{3}}{4} + j\frac{1}{4}.\) Since \(|z_2| = |z_3| = 1/2\) possible regions of convergence are: 1) \(|z| < 1/4, 2) 1/4 < |z| < 1/2,\) and 3) \(|z| > 1/2.\) In the first two cases the system is neither causal nor stable, while in the third case the system is causal and stable since \(|z| = 1\) and \(|z| \to \infty\) belong to the region of convergence.

The output to \(x(n) = 2\cos(n\pi/2) = 1 + \cos(n\pi) = 1 + (-1)^n\) is \(y(n) = H(e^{j\omega})_{\omega=0} \times 1 + H(e^{j\omega})_{\omega=\pi} \times (-1)^n = H(z)_{|z|=1} + H(z)_{|z|=-1} (-1)^n = -0.8681 + 0.0945(-1)^n.\)

**Solution 4.10.** The transfer function can be written as

\[ H(z) = \frac{z + 2}{z^2(z - 2)} = \frac{A}{z - 2} + \frac{B}{z} + \frac{C}{z^2}. \]

Multiplying both sides by \(z^2(z - 2)\) yields

\[ Az^2 + Bz(z - 2) + C(z - 2) = z + 2 \]

\[ (A + B)z^2 + (-2B + C) - 2C = z + 2. \]

The coefficients follow from

\[ A + B = 0 \]
\[ -2B + C = 1 \]
\[ -2C = 2, \]

as \(A = 1, B = -1,\) and \(C = -1.\) The transfer function is

\[ H(z) = \frac{z^{-1}}{1 - 2z^{-1}} - \frac{1}{z^2} - \frac{1}{z}. \]
The region of convergence for a causal system is $|z| > 2$. The inverse $z$-transform for a causal system is the system impulse response

$$h(n) = 2^{n-1}u(n-1) - \delta(n-2) - \delta(n-1) = \delta(n-2) + 2^{n-1}u(n-3).$$

The system is not stable.

**Solution 4.11.** The $z$-transform $X(z)$ can be written in the form

$$X(z) = \frac{z^2}{z^2 + 1} = \frac{1}{2z} + \frac{1}{2z}.$$

For the region of convergence defined by $|z| > 1$ the signal is causal and

$$x(n) = \frac{1}{2}[1 + (-1)^n]j^n u(n) = \frac{1}{2}[1 + (-1)^n]e^{j\pi n/2} u(n).$$

For $n = 4k$, where $k \geq 0$ is an integer, $x(n) = 1$, while for $n = 4k + 2$ the signal values are $x(n) = -1$. For other $n$ the signal is $x(n) = 0$.

For $|z| < 1$ the inverse $z$-transform is

$$x(n) = -\frac{1}{2}[1 + (-1)^n]j^n u(-n - 1).$$

**Solution 4.12.** The transfer function of this system is

$$H(z) = \frac{3 - \frac{5}{4}z^{-1} + \frac{3}{16}z^{-2}}{1 - z^{-1} + \frac{5}{16}z^{-2} - \frac{1}{32}z^{-3}} = \frac{\frac{3}{4}z^{-1} + \frac{3}{16}z^{-2}}{(1 - \frac{1}{2}z^{-1} + \frac{1}{16}z^{-2})(1 - \frac{1}{2}z^{-1})} = \frac{1}{1 - \frac{1}{4}z^{-1}} + \frac{1}{(1 - \frac{1}{4}z^{-1})^2} + \frac{1}{(1 - \frac{1}{2}z^{-1})}.$$

For a causal system the region of convergence is outside of the pole $z = 1/2$, that is $|z| > 1/2$. Since

$$\frac{1}{(1 - \frac{1}{4}z^{-1})^2} = \frac{d}{da} \left( \frac{z}{1 - az^{-1}} \right) \bigg|_{a=1/4}$$

$$= \frac{d}{da} \sum_{n=0}^{\infty} a^n z^{-(n-1)} \bigg|_{a=1/4} = \sum_{n=0}^{\infty} na^{n-1} z^{-(n-1)} \bigg|_{a=1/4} = \sum_{n=0}^{\infty} (n + 1) \frac{1}{4^n} z^{-n},$$

the inverse $z$-transform is

$$h(n) = \frac{1}{4^n} u(n) + (n + 1) \frac{1}{4^n} u(n) + \frac{1}{2^n} u(n).$$
**Solution 4.13.** The transfer function of the system defined by

\[ y(n) = x(n) - \frac{3}{4} x(n-1) + \frac{1}{8} x(n-2) \]

is

\[ H(z) = 1 - \frac{3}{4} z^{-1} + \frac{1}{8} z^{-2}. \]

The z-transform of the input signal \( x(n) = \frac{1}{4^n} u(n) \) is

\[ X(z) = \frac{1}{1 - \frac{1}{4} z^{-1}}, \]

with the region of convergence \( |z| > 1/4 \). The output signal z-transform is

\[ Y(z) = H(z) X(z) = \frac{(1 - \frac{1}{2} z^{-1})(1 - \frac{1}{4} z^{-1})}{(1 - \frac{1}{4} z^{-1})} = 1 - \frac{1}{2} z^{-1}. \]

Its inverse is a finite duration output signal

\[ y(n) = \delta(n) - \delta(n-1)/2. \]

**Solution 4.14.** The system transfer function is

\[ H(z) = \frac{1}{1 - \frac{1}{3} z^{-1}} \]

and the input signal z-transform is

\[ X(z) = 1 + z^{-1} + z^{-2} + z^{-3} + z^{-4} + z^{-5} = \frac{1 - z^{-6}}{1 - z^{-1}}. \]

The z-transform of the output signal is

\[ Y(z) = \frac{1 - z^{-6}}{(1 - z^{-1})(1 - \frac{1}{3} z^{-1})} = Y_1(z) - Y_1(z)z^{-6} \]

with

\[ Y_1(z) = \frac{1}{(1 - z^{-1})(1 - \frac{1}{3} z^{-1})} = \frac{3/2}{1 - z^{-1}} - \frac{1/2}{1 - \frac{1}{3} z^{-1}}. \]

Its inverse is

\[ y_1(n) = \left[ \frac{3}{2} - \frac{1}{2} \left( \frac{1}{3} \right)^n \right] u(n). \]
Thus the system output is

\[ y(n) = \left[ \frac{3}{2} - \frac{1}{2} \left( \frac{1}{3} \right)^n \right] u(n) - \left[ \frac{3}{2} - \frac{1}{2} \left( \frac{1}{3} \right)^{n-6} \right] u(n-6). \]

**Solution 4.15.** The transfer function is obtained from

\[ Y(z) \left(1 - \frac{11}{6} z^{-1} + \frac{1}{2} z^{-2}\right) = X(z) \left(2 - \frac{3}{2} z^{-1}\right) \]

as

\[ H(z) = \frac{2 - \frac{3}{2} z^{-1}}{1 - \frac{11}{6} z^{-1} + \frac{1}{2} z^{-2}}. \]

The poles are at \( z_{p1} = 1/3 \) and \( z_{p2} = 3/2 \) with the region of convergence \( |z| > 3/2 \). It means that the system is not stable, Fig.4.8.

The \( z \)-transform of the input signal is

\[ X(z) = 1 - \frac{3}{2} z^{-1} \text{ for } |z| > 0. \]

The output signal transform is

\[ Y(z) = \frac{2 - \frac{3}{2} z^{-1}}{1 - \frac{11}{6} z^{-1} + \frac{1}{2} z^{-2}} \left(1 - \frac{3}{2} z^{-1}\right) = \frac{2 - \frac{3}{2} z^{-1}}{1 - \frac{11}{6} z^{-1} + \frac{1}{2} z^{-2}}. \]

The output signal transform does not have a pole \( z = 3/2 \) since this pole is canceled out. The output signal is

\[ y(n) = \frac{1}{3^n} u(n) - \frac{3}{2} \frac{1}{3^{n-1}} u(n-1). \]
Solution 4.16. The $z$-transform of signal $x(n + 2)$ is

$$X_2(z) = z^2X(z) - z^2x(0) - zx(1)$$

while for $x(n + 1)$ the transform is

$$X_1(z) = zX(z) - zx(0).$$

The $z$-transform domain form is

$$z^2X(z) - z^2x(0) - zx(1) + 3zX(z) - 3zx(0) + 2X(z) = 0$$

with

$$X(z) = \frac{z}{z^2 + 3z + 2} = \frac{1}{1 + z^{-1}} - \frac{1}{1 + 2z^{-1}}.$$  

The inverse $z$-transform of $X(z)$ is

$$x(n) = [(-1)^n - (-2)^n]u(n).$$

Solution 4.17. The $z$-transforms of the left and right side of the equation are

$$zX(z) - zx(0) = X(z) + \frac{z}{z - a}$$

$$X(z) = \frac{z}{(z - a)(z - 1)} = \frac{1}{1 - a} \left[ \frac{1}{z - 1} - \frac{a}{z - a} \right].$$

The inverse $z$-transform is

$$x(n) = \frac{1}{1 - a} \left[ u(n - 1) - a^n u(n - 1) \right] = \frac{1 - a^n}{1 - a} u(n - 1)$$

or

$$x(n) = \sum_{k=0}^{n-1} a^k, \ n > 0.$$  

Solution 4.18. For a direct solution in the discrete-time domain we assume

a solution of the homogenous part of the equation

$$y(n) - \frac{\sqrt{2}}{2} y(n - 1) + \frac{1}{4} y(n - 2) = 0 \quad (4.22)$$

in the form $y_i(n) = C_i \lambda_i^n$. The characteristic polynomial is

$$\lambda^2 - \frac{\sqrt{2}}{2} \lambda + \frac{1}{4} = 0$$
with $\lambda_{1,2} = \frac{\sqrt{2}}{4} \pm j\frac{\sqrt{2}}{4}$. The homogenous solution is

$$y_h(n) = C_1\left(\frac{\sqrt{2}}{4} + j\frac{\sqrt{2}}{4}\right)^n + C_2\left(\frac{\sqrt{2}}{4} - j\frac{\sqrt{2}}{4}\right)^n$$

$$= C_1\frac{1}{2^n}e^{j\frac{n\pi}{4}} + C_2\frac{1}{2^n}e^{-j\frac{n\pi}{4}}.$$ 

A particular solution is of the input signal $x(n) = \frac{1}{T}u(n)$ form. It is $y_p(n) = A\frac{1}{T}u(n)$. The constant $A$ is obtained by replacing this signal into (4.20)

$$A\frac{1}{3^n} - \frac{\sqrt{2}}{2}A\frac{1}{3^{n-1}} + \frac{1}{4}A\frac{1}{3^{n-2}} = \frac{1}{3^n}$$

$$A(1 - \frac{3\sqrt{2}}{2} + \frac{9}{4}) = 1.$$ 

Its value is $A = 0.886$. The general solution is

$$y(n) = y_h(n) + y_p(n) = C_1\frac{1}{2^n}e^{j\frac{n\pi}{4}} + C_2\frac{1}{2^n}e^{-j\frac{n\pi}{4}} + 0.886\frac{1}{3^n}.$$ 

Since the system is causal with $y(n) = 0$ for $n < 0$ then the constants $C_1$ and $C_2$ may be obtained from the initial condition following from $y(n) - \frac{\sqrt{2}}{2}y(n-1) + \frac{1}{4}y(n-2) = x(n)$ as $y(0) = x(0) = 1$ and $y(1) = \frac{\sqrt{2}}{2}y(0) + x(1) = \frac{\sqrt{2}}{2} + \frac{1}{3}$,

$$C_1 + C_2 + 0.886 = 1$$

$$C_1\left(\frac{\sqrt{2}}{2} + j\frac{\sqrt{2}}{2}\right)/2 + C_2\left(\frac{\sqrt{2}}{2} - j\frac{\sqrt{2}}{2}\right)/2 + 0.886\frac{1}{3} = \frac{\sqrt{2}}{2} + \frac{1}{3},$$

as $C_1 = 0.057 - j0.9967 = 0.9984\exp(-j1.5137) = C_2^*$. The final solution is

$$y(n) = 2 \times 0.9984\frac{1}{2^n}\cos(n\pi/4 - 1.5137) + 0.886\frac{1}{3^n}.$$ 

For the $z$-domain we write

$$Y(z) - \frac{\sqrt{2}}{2}Y(z)z^{-1} + \frac{1}{4}Y(z)z^{-2} = X(z)$$

with

$$Y(z) = \frac{1}{1 - \frac{\sqrt{2}}{2}z^{-1} + \frac{1}{4}z^{-2}} \frac{1}{1 - \frac{1}{3}z^{-1}}$$
with

\[ Y(z) = \frac{z^3}{(z - (\frac{\sqrt{2}}{4} + j \frac{\sqrt{2}}{4}))(z - (\frac{\sqrt{2}}{4} - j \frac{\sqrt{2}}{4}))(z - \frac{1}{3})} \]

Using, for example, the residual value based inversion of the z-transform,

\[ y(n) = \sum_{z_{1,2,3}} \left\{ [z^{n-1}Y(z)(z - z_i)]_{z = z_i} \right\} \]

\[ = z^{n+2} \left( \frac{\sqrt{2} + j \sqrt{2}}{4} \right)^{n+2} \frac{1}{(z - \frac{\sqrt{2} - j \sqrt{2}}{4})(z - \frac{1}{3})} + z^{n+2} \frac{1}{(z - \frac{\sqrt{2} + j \sqrt{2}}{4})(z - \frac{1}{3})} \]

\[ + z^{n+2} \frac{1}{(z - \frac{\sqrt{2} + j \sqrt{2}}{4})(z - \frac{\sqrt{2} - j \sqrt{2}}{4})} \]

\[ = \frac{1}{j \sqrt{2}} \left( \frac{\sqrt{2} + j \sqrt{2}}{4} \right)^{n+2} \left( \frac{\sqrt{2} - j \sqrt{2}}{4} \right)^{n+2} \frac{1}{\frac{1}{3} - \frac{\sqrt{2}}{2} + \frac{1}{4}} \frac{1}{\sqrt{2} - j \sqrt{2} - \frac{1}{3}} \]

\[ + \frac{1}{3^{n+2}} \left( \frac{1}{\frac{\sqrt{2}}{2} - \frac{1}{3}} - \frac{\sqrt{2}}{2} + \frac{1}{4} \right) \]

\[ = \frac{1}{2^{n+2}} e^{j(n+2)\pi/4} \left( \frac{\sqrt{2} - j \sqrt{2}}{4} \right)^{n+2} \frac{1}{\frac{\sqrt{2}}{2} - \frac{j \sqrt{2}}{4} + \frac{1}{3}} \]

\[ + \frac{1}{2 n^{n+2}} e^{-j(n+2)\pi/4} \left( \frac{\sqrt{2} + j \sqrt{2}}{4} \right)^{n+2} \frac{j \sqrt{2}}{\frac{\sqrt{2}}{2} - j \sqrt{2} - \frac{1}{3}} \]

\[ + 0.886 \frac{1}{3^n} \]

\[ = 2 \times 0.9984 \frac{1}{2^n} \cos(n \pi/4 - 1.5137) + 0.886 \frac{1}{3^n} \]

for \( n \geq 1 \). For \( n = 0 \) there is no additional pole at \( z = 0 \) the previous result holds for \( n \geq 0 \).

**Solution 4.19.** The z-transform of the first backward difference is

\[ \mathcal{Z}[\nabla x(n)] = \mathcal{Z}[x(n)] - \mathcal{Z}[x(n-1)] = (1 - z^{-1}) X(z). \]

The second backward difference may be written as

\[ \nabla^2 x(n) = \nabla[\nabla x(n)] = \nabla [x(n) - x(n-1)] = \nabla x(n) - \nabla x(n-1) \]

\[ = x(n) - 2x(n - 1) + x(n - 2). \]
Its $z$-transform is
\[ Z[\nabla^2 x(n)] = (1 - z^{-1})^2 X(z). \]

In the same way we get
\[ Z[\nabla^m x(n)] = (1 - z^{-1})^m X(z). \]

The $z$-transform of the first forward difference is
\[
Z[\Delta x(n)] = Z[x(n + 1) - x(n)] = zX(z) - zx(0) - X(z) = (z - 1)X(z) - zx(0).
\]

The second forward difference is
\[
Z[\Delta^2 x(n)] = x(n + 2) - 2x(n + 1) + x(n)
\]
with the $z$-transform
\[
Z[\Delta^2 x(n)] = (z - 1)^2 X(z) - z(z - 1)x(0) - z\Delta x(0).
\]

In a recursive way, the $z$-transform of the $m$th forward difference is
\[
Z[\Delta^m x(n)] = (z - 1)^m X(z) - z \sum_{j=0}^{m-1} (z - 1)^{m-j-1} \Delta^j x(0).
\]

**Solution 4.20.** The transfer function of this system is
\[
H(z) = \frac{1 - \sqrt{2}z^{-1} + z^{-2}}{1 - r\sqrt{2}z^{-1} + r^2z^{-2}} = \frac{1 - (\frac{\sqrt{2}}{2} + j\frac{\sqrt{2}}{2})z^{-1}}{1 - r(\frac{\sqrt{2}}{2} + j\frac{\sqrt{2}}{2})z^{-1}}\cdot
\]
\[
\frac{1 - (\frac{\sqrt{2}}{2} - j\frac{\sqrt{2}}{2})z^{-1}}{1 - r(\frac{\sqrt{2}}{2} - j\frac{\sqrt{2}}{2})z^{-1}} = \frac{[z - (\frac{\sqrt{2}}{2} + j\frac{\sqrt{2}}{2})][z - (\frac{\sqrt{2}}{2} - j\frac{\sqrt{2}}{2})]}{[z - r(\frac{\sqrt{2}}{2} + j\frac{\sqrt{2}}{2})][z - r(\frac{\sqrt{2}}{2} - j\frac{\sqrt{2}}{2})]}.
\]

The zeros and poles are $z_{01,02} = \frac{\sqrt{2}}{2} \pm j\frac{\sqrt{2}}{2}$ and $z_{p1,p2} = r\frac{\sqrt{2}}{2} \pm jr\frac{\sqrt{2}}{2}$. They are located as in Fig.4.9.

The amplitude of the frequency response is
\[
|H(e^{j\omega})| = \left| \frac{B_0}{A_0} \right| \frac{T_O \ T_T}{T_P \ T_P} = \frac{T_O \ T_T}{T_P \ T_P}.
\]

The values of $T_P$ and $T_O$, and $T_T$ and $T_P$, are almost the same for any $\omega$ except $\omega = \pm \pi/4$ where the distance to the transfer function zero is
0, while the distance to the corresponding pole is small but finite. Based on this analysis the amplitude of frequency response is presented in Fig. 4.9.

The input discrete-time signal is

\[ x(n) = x(n\Delta t)n\Delta t = [2\cos(\pi n/6) - \sin(\pi n/4) + 0.5e^{j\pi n/3}] / 60. \]

This system will filter out signal components at \( \omega = \pm \pi/4 \). The output discrete-time signal is

\[ y(n) = [2\cos(n\pi/6) + 0.5e^{jn\pi/3}] / 60. \]

Corresponding continuous-time output signal is

\[ y(t) = 2\cos(10\pi t) + 0.5e^{j20\pi t}. \]

**Solution 4.21.** The zeros of the system are

\[ z_o^{-N} = 1 = e^{-j2\pi m} \]

\[ z_{om} = e^{j2\pi m/N}, m = 0, 1, ..., N - 1 \]

Similarly, the poles are \( z_{mp} = r^{1/N}e^{j2\pi m/N}, m = 0, 1, ..., N - 1 \). The frequency response of the comb filter is

\[ H(z) = \prod_{m=0}^{N-1} \frac{z - z_{om}}{z - z_{pm}} = \prod_{m=0}^{N-1} \frac{z - e^{j2\pi m/N}}{z - r^{1/N}e^{j2\pi m/N}}. \]
With \( r = 0.9999 \) and \( r^{1/N} \cong 1 \) follows

\[
|H(e^{j\omega})| \cong 1 \text{ for } z \neq e^{j2\pi m/N},
\]
\[
|H(e^{j\omega})| = 0 \text{ for } z = e^{j2\pi m/N}.
\]

The same holds for

\[
H(z) = \frac{(1 - z^{-1})(1 + z^{-1})}{(1 - rz^{-1})(1 + rz^{-1})} \prod_{k=1}^{N/2-1} \frac{1 - 2\cos(2k\pi/N)z^{-1} + z^{-2}}{1 - 2r\cos(2k\pi/N)z^{-1} + r^2z^{-2}}
\]

since for \( 1 \leq k \leq N/2 - 1 \) we can group the terms

\[
\frac{(1 - e^{2k\pi/N}z^{-1})(1 - e^{2(N-k)\pi/N}z^{-1})}{(1 - re^{2k\pi/N}z^{-1})(1 - re^{2(N-k)\pi/N}z^{-1})} = \frac{1 - 2\cos(2k\pi/N)z^{-1} + z^{-2}}{1 - 2r\cos(2k\pi/N)z^{-1} + r^2z^{-2}}.
\]

### 4.9 Exercise

**Exercise 4.1.** Find the \( z \)-transform and the region of convergence for the following signals:

(a) \( x(n) = \delta(n-3) - \delta(n+3) \),
(b) \( x(n) = u(n) - u(n-20) + 3\delta(n) \),
(c) \( x(n) = 1/3^n x + 1/2^n u(n) \),
(d) \( x(n) = 3^n u(-n) + 2^{-n} u(n) \),
(e) \( x(n) = n(1/3)^n u(n) \),
(f) \( x(n) = \cos(n\pi/2) \).

**Exercise 4.2.** Find the \( z \)-transform and the region of convergence for the signals:

(a) \( x(n) = 3^n u(n) - (-2)^n u(n) + n^2 u(n) \).
(b) \( x(n) = \sum_{k=0}^{n} 2^k 3^{n-k} \).
(c) \( x(n) = \sum_{k=0}^{n} k^k \).

**Exercise 4.3.** Find the inverse \( z \)-transform of:

(a) \( X(z) = \frac{z^{-3}}{z^{-2} + 3} \), if \( X(z) \) is the \( z \)-transform of a causal signal \( x(n) \).
(b) \( X(z) = \frac{z+2}{(z-2)z^2} \), if \( X(z) \) is the \( z \)-transform of a causal signal \( x(n) \).
(c) \( X(z) = \frac{6z^2+3z-2}{6z^2-5z+1} \), if \( X(z) \) is the \( z \)-transform of an unlimited-duration signal \( x(n) \). Find \( \sum_{n=-\infty}^{\infty} x(n) \) in this case.

**Exercise 4.4.** Find the inverse \( z \)-transforms of:

(a) \( X(z) = \frac{z^3(5z-3)}{(3z-1)(2z-4)} \), if \( x(n) \) is causal,
(b) \( Y(z) = X(\frac{z}{2}) \), for a causal signal \( y(n) \),
(c) \( Y(z) = z^{-2}X(z) \), for a causal signal \( y(n) \).

**Exercise 4.5.** Find the inverse z-transforms of \( X(z) = \cosh(az) \) and \( X(z) = \sinh(az) \).

**Exercise 4.6.** If \( X(z) \) is the z-transform of a signal \( x(n) \), with the region of convergence \( |z| > \frac{1}{2} \), find the z-transforms for the following signals:
(a) \( y(n) = x(n) - x(n - 1) \),
(b) \( y(n) = \sum_{k=-\infty}^{\infty} x(n - kN) \), where \( N \) is an integer,
(c) \( y(n) = x(n) * x(-n) \), where \( * \) denotes convolution.
(d) find the signal whose z-transform is \( Y(z) = \frac{d}{dz}X(z) \).

**Exercise 4.7.** If \( X(z) \) is the z-transform of a signal \( x(n) \) find the z-transform of
\[ y(n) = \sum_{k=-\infty}^{\infty} x^{*}(n - k)x(n + k). \]

**Exercise 4.8.** For the z-transform
\[ H(z) = \frac{(2 - z)}{(1 - 4z)(1 - 3z)} \]
identify possible regions of convergence and find the inverse z-transform for each of them. For each case comment stability and causality. What is the output of the stable system to \( x(n) = 1 + (-1)^n \)?

**Exercise 4.9.** Find the output of a causal discrete system
\[ y(n) - \frac{3}{4}y(n - 1) + \frac{1}{8}y(n - 2) = x(n). \] (4.24)
to the input signal \( x(n) = nu(n) \) by:
(a) a direct solution in the time domain.
(b) using the z-transform.
The initial conditions are \( y(n) = 0 \) for \( n < 0 \), that is \( y(0) = x(0) = 0 \) and \( y(1) = 3y(0)/4 + x(1) = 1 \).

**Exercise 4.10.** A causal discrete system is described by the difference equation
\[ y(n) - \frac{5}{6}y(n - 1) + \frac{1}{6}y(n - 2) = x(n). \] (4.25)
If the input signal is \( x(n) = 1/4^n u(n) \) find the output signal if the initial value of the output was \( y(0) = 2 \).
Hint: Since \( y(0) \) does not follow from (4.25) obviously the system output was "preloaded" before the input is applied. This fact can be taken into account by changing the input signal at \( n = 0 \) to produce the initial output. It is \( x(n) = 1/4^n u(n) + \delta(n) \). Now the initial conditions are \( y(0) = 2 \) and \( y(1) = 5/3 + 1/4 = 23/12 \) and we can apply the \( z \)-transform with this new input signal.

**Exercise 4.11.** Solve the difference equation using the \( z \)-transform

\[
x(n + 2) - \frac{1}{2}x(n + 1) + x(n) = 0
\]

with initial condition \( x(0) = 0 \) and \( x(1) = 1/2 \). The signal \( x(n) \) is causal.

**Exercise 4.12.** Using the basic trigonometric transformations show that a real-valued signal \( y(n) = \cos(2\pi k_0 n / N + \phi) \) is a solution of the homogeneous difference equation

\[
y(n) - 2\cos(2\pi k_0 / N)y(n - 1) + y(n - 2) = 0.
\]

with similar conclusions as in the complex-valued signal case.

**Exercise 4.13.** For the system

\[
H(z) = \frac{(1 - z^{-1})(1 + z^{-1})}{(1 - rz^{-1})(1 + rz^{-1})} \prod_{k=1}^{3} \frac{1 - 2\cos(2k\pi / 8)z^{-1} + z^{-2}}{1 - 2r\cos(2k\pi / 8)z^{-1} + z^{-2}}
\]

and \( r = 0.9999 \) plot the amplitude of the frequency response and find the output to the signal

\[
x(n) = \cos(n\pi / 3 + \pi / 4) + \sin(n\pi / 2) + (-1)^n.
\]
Chapter 7

Discrete-Time Random Signals

Random signals cannot be described by simple mathematical functions. Their values are not known in advance. These signals can be described by stochastic tools only. Here we will restrict the analysis to the discrete-time random signals. The first-order and the second-order statistics will be considered.

7.1 BASIC STATISTICAL DEFINITIONS

7.1.1 Expected Value

The first-order statistics is the starting point in describing random signals. The expected value, or the mean value, of a random signal is one of its basic parameters. If we have a set of signal samples,

\[ \{x(n)\}, \quad n = 1, 2, ..., N, \quad (7.1) \]

the mean value of this set of signal values is calculated as

\[ \mu_x = \frac{1}{N} (x(1) + x(2) + ... + x(N)). \]

Example 7.1. Consider a random signal \( x(n) \) whose one realization is given in Table 7.1. Find the mean value of this signal. Find how many samples of the signal are within the intervals \([1, 10],[11, 20],...,[91, 100]\). Plot the number of occurrences of signal \( x(n) \) samples within these intervals as a function of the interval range.

★ The realization of signal \( x(n) \) defined in Table 7.1 is presented in Fig.7.1.
The mean value of all signal samples is

$$
\mu_x = \frac{1}{100} \sum_{n=1}^{100} x(n) = 55.76.
$$
From Table 7.1 or the graph in Fig. 7.1 we can count that, for example, there is no a signal sample whose value is within the interval $[1, 10]$. Within $[11, 20]$ there are two signal samples ($x(42) = 12$ and $x(95) = 11$). In a similar way, the number of signal samples within other intervals are counted and presented in Fig. 7.2. This kind of random signal presentation is called a histogram of $x(n)$, with defined intervals.

Example 7.2. For the signal $x(n)$ from the previous example assume that a new random signal $y(n)$ is formed as

$$y(n) = \text{int} \left\{ \frac{x(n) + 5}{10} \right\},$$

where int $\{ \}$ denotes the nearest integer. It means that $y(n) = 1$ for $1 \leq x(n) \leq 10$, $y(n) = 2$ for $11 \leq x(n) \leq 20$, ..., $y(n) = i$ for $10(i - 1) + 1 \leq x(n) \leq 10i$ up to $i = 10$. Plot the new signal $y(n)$. What is the set of possible values of $y(n)$? Present on a graph how many times each of the possible values of $y(n)$ appeared in this signal realization. Find the mean value of the new signal $y(n)$ and discuss the result.

☆ The signal $y(n)$ is shown in Fig. 7.3. This signal assumes values from the set $\{2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

For the signal $y(n)$, instead of histogram we can plot a diagram of the number of occurrences of each value that $y(n)$ can assume. It is presented in
The mean value of $y(n)$ is

$$\mu_y = \frac{1}{100} \sum_{n=1}^{100} y(n) = 6.13.$$  

The mean value can also be written, by grouping the same values of $y(n)$, as

$$\mu_y = \frac{1}{100} \left( 1 \cdot n_1 + 2 \cdot n_2 + 3 \cdot n_3 + \ldots + 10 \cdot n_{10} \right) = 1 \cdot \frac{n_1}{N} + 2 \cdot \frac{n_2}{N} + 3 \cdot \frac{n_3}{N} + \ldots + 10 \cdot \frac{n_{10}}{N},$$

where $N = 100$ is the total number of signal values and $n_i$ is the number showing how many times each of the values $i$ appeared in $y(n)$. If there is a sufficient number of occurrences for each outcome value $i$ then

$$P_y(i) = \frac{n_i}{N},$$

can be considered as the probability that the value $i$ appears. In that sense

$$\mu_y = 1 \cdot P_y(1) + 2 \cdot P_y(2) + 3 \cdot P_y(3) + \ldots + 10 \cdot P_y(10)$$

$$= \sum_{i=1}^{10} y(i) P_y(i)$$
Figure 7.4 Number of appearances of each possible value of $y(n)$ (left) and the probabilities that the random signal $y(n)$ takes a value $i = 1,2,...,10$ (right).

with

$$\sum_{i=1}^{10} P_y(i) = 1.$$ 

Values of probability $P_y(i)$ are shown in Fig. 7.4.

In general, the mean for each signal sample could be different. For example, if the signal values represent the highest daily temperature during a year then the mean value is highly dependent on the considered sample. In order to calculate the mean value of temperature, we have to have several realizations of these random signals (measurements over $M$ years), denoted by $\{x_i(n)\}$, where argument $n = 1,2,3,...$ is the cardinal number of the day within a year and $i = 1,2,...,M$ is the index of realization (year index). The mean value is then calculated as

$$\mu_x(n) = \frac{1}{M} (x_1(n) + x_2(n) + ... + x_M(n)) = \frac{1}{M} \sum_{i=1}^{M} x_i(n), \quad (7.2)$$

for each $n$. In this case we have a set (a signal) of mean values $\{\mu_x(n)\}$, for $n = 1,2,...,365$.

**Example 7.3.** Consider a signal $x(n)$ with realizations given in Table 7.2. Its values are equal to the monthly average of maximal daily temperatures in a city measured from year 2001 to 2015. Find the mean temperature for each month over the considered period of years. What is the mean value of temperature over all months and years? What is the mean temperature for each year?
Table 7.2
Average of maximal temperatures value within months over 15 years, 2001-2015.

<table>
<thead>
<tr>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>4</td>
<td>18</td>
<td>17</td>
<td>22</td>
<td>29</td>
<td>30</td>
<td>28</td>
<td>27</td>
<td>17</td>
<td>17</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>11</td>
<td>23</td>
<td>22</td>
<td>32</td>
<td>35</td>
<td>33</td>
<td>22</td>
<td>26</td>
<td>22</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>10</td>
<td>16</td>
<td>21</td>
<td>26</td>
<td>32</td>
<td>31</td>
<td>23</td>
<td>19</td>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>13</td>
<td>19</td>
<td>22</td>
<td>26</td>
<td>34</td>
<td>29</td>
<td>26</td>
<td>22</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>13</td>
<td>21</td>
<td>27</td>
<td>29</td>
<td>30</td>
<td>34</td>
<td>24</td>
<td>20</td>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>11</td>
<td>17</td>
<td>17</td>
<td>27</td>
<td>25</td>
<td>37</td>
<td>34</td>
<td>33</td>
<td>22</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>13</td>
<td>19</td>
<td>23</td>
<td>32</td>
<td>34</td>
<td>38</td>
<td>21</td>
<td>21</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>9</td>
<td>20</td>
<td>21</td>
<td>37</td>
<td>34</td>
<td>34</td>
<td>27</td>
<td>22</td>
<td>20</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>13</td>
<td>23</td>
<td>27</td>
<td>33</td>
<td>29</td>
<td>31</td>
<td>25</td>
<td>21</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>10</td>
<td>17</td>
<td>27</td>
<td>33</td>
<td>38</td>
<td>32</td>
<td>23</td>
<td>20</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>13</td>
<td>24</td>
<td>23</td>
<td>33</td>
<td>33</td>
<td>31</td>
<td>27</td>
<td>21</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>15</td>
<td>18</td>
<td>25</td>
<td>26</td>
<td>27</td>
<td>33</td>
<td>23</td>
<td>23</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>16</td>
<td>17</td>
<td>27</td>
<td>28</td>
<td>30</td>
<td>32</td>
<td>29</td>
<td>24</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>12</td>
<td>14</td>
<td>18</td>
<td>22</td>
<td>29</td>
<td>34</td>
<td>34</td>
<td>23</td>
<td>21</td>
<td>20</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>13</td>
<td>8</td>
<td>22</td>
<td>22</td>
<td>29</td>
<td>30</td>
<td>34</td>
<td>23</td>
<td>18</td>
<td>15</td>
<td>8</td>
</tr>
</tbody>
</table>

★The signal for years 2001 to 2007 is presented in Fig. 7.5. The mean temperature for the $n$th month, over the considered years, is

$$\mu_x(n) = \frac{1}{15} \sum_{i=1}^{15} x_{20i}(n),$$

where the notation $20i$ is symbolic in the sense, 2001, 2002, ..., 2015, for $i = 01, 02, ..., 15$. The mean-value signal $\mu_x(n)$ is presented in the last subplot of Fig. 7.5. The mean value over all months and years is

$$\mu_x = \frac{1}{15 \cdot 12} \sum_{n=1}^{12} \sum_{i=1}^{15} x_{20i}(n) = 19.84.$$  

The mean value for each of the considered years is

$$\mu_x(20i) = \frac{1}{12} \sum_{n=1}^{12} x_{20i}(n).$$

\[\square\]
Figure 7.5 Several realizations of a random signal $x_{20}(n)$, for $i = 01, 02, \ldots, 07$ and the mean value $\mu_x(n)$ for each sample (month) over 15 available realizations.
7.1.2 Probability and Probability Density Function

If the probabilistic description of a random signal is known, then we can calculate the mean value and other parameters of random signals. For the first-order statistics calculation, it is sufficient to know the probabilities or the probability density function.

If a random signal assumes only discrete values in amplitude \( \{ \xi_1, \xi_2, \ldots \} \), then we deal with probabilities,

\[
\text{Probability} \left\{ x(n) = \xi_i \right\} = P_{x(n)}(\xi_i). \tag{7.3}
\]

Probability function \( P_{x(n)}(\xi) \) satisfies the following properties:

1) \( 0 \leq P_{x(n)}(\xi) \leq 1 \) for any \( \xi \).
2) For the events \( x(n) = \xi_i \) and \( x(n) = \xi_j, \ i \neq j \), which exclude each other

\[
\text{Probability} \left\{ x(n) = \xi_i \text{ or } x(n) = \xi_j \right\} = P_{x(n)}(\xi_i) + P_{x(n)}(\xi_j).
\]

3) The sum of probabilities that \( x(n) \) takes any value \( \xi_i \) over the set \( A \) of all possible values of \( \xi \) is a certain event. Its probability is 1,

\[
\sum_{\xi \in A} P_{x(n)}(\xi) = 1.
\]

An impossible event has the probability 0. 
If \( x(n) \) and \( x(m) \) are statistically independent random samples then

\[
\text{Probability} \left\{ x(n) = \xi_i \text{ and } x(m) = \xi_j \right\} = P_{x(n)}(\xi_i)P_{x(m)}(\xi_j).
\]

An example of a signal when the probabilities are calculated after the experiment (a posteriori) is already presented within the first example. A posteriori probability that the signal \( x(n) \) assumes value \( \xi_i \) is defined as a ratio of the number \( N_{\xi_i} \) of appearances of the event \( x(n) = \xi_i \) and the total number of signal values (experiments) \( N \)

\[
P_{x(n)}(\xi_i) = \frac{N_{\xi_i}}{N}
\]

for a sufficiently large \( N \) and \( N_{\xi_i} \).

In some cases it is possible to find the probability of an event before the experiment is performed. For example, if a signal is equal to the numbers appearing in die tossing, then the signal may assume one of the values from the set \( \xi_i \in \{1, 2, 3, 4, 5, 6\} \). In this case, the probability of each event is known in advance (a priori). It is \( P(\xi_i) = 1/6 \). 


Example 7.4. Consider a random signal whose values are equal to the numbers appearing in a die tossing. The set of possible signal values is \( \xi \in \{1,2,3,4,5,6\} \).

Find

\[
\text{Probability} \{x(n) = 2 \text{ or } x(n) = 5\}
\]

and

\[
\text{Probability} \{x(n) = 2 \text{ and } x(n + 1) = 5\}.
\]

★ Events that \( x(n) = 2 \) and \( x(n) = 5 \) are obviously mutually exclusive. Thus,

\[
\text{Probability} \{x(n) = 2 \text{ or } x(n) = 5\} = P_{x(n)}(2) + P_{x(n)}(5) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.
\]

The events that \( x(n) = 2 \) and \( x(n + 1) = 5 \) are statistically independent. In this case

\[
\text{Probability} \{x(n) = 2 \text{ and } x(n + 1) = 5\} = P_{x(n)}(2)P_{x(n)}(5) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}.
\]

Example 7.5. Assume that a signal \( x(n) \) length is \( N \) and that the number of samples disturbed by an extremely high noise is \( I \). The observation set of signal samples is taken as a set of \( M < N \) randomly positioned signal samples. What is the probability that within \( M \) randomly selected signal samples there are no samples affected by the high noise? If \( N = 128 \), \( I = 16 \), and \( M = 32 \) find how many sets of \( M \) samples without high noise can be expected in 1000 realizations (trials).

★ Probability that the first randomly chosen sample is not affected by the high noise could be calculated as a priori probability,

\[
P(1) = \frac{N-I}{N}
\]

since there are \( N \) samples in total and \( N-I \) of them are noise-free. Probability that the first randomly chosen sample is not affected by high noise and that, at the same time, the second randomly chosen sample is not affected by high noise is equal to to the product of their probabilities,

\[
P(2) = \frac{N-I}{N} \cdot \frac{N-I-1}{N-1}.
\]

Here we used so called conditional probability property stating that the probability that both events \( A \) and \( B \) occur is

\[
\text{Probability} \{A \text{ and } B\} = P(A)P(B/A),
\]

where \( P(A) \) is the probability that event \( A \) occurs, while \( P(B/A) \) denotes the probability that event \( B \) occurs subject to the condition that event \( A \) already occurred.
Then we continue the process of random samples selection. In the same way we can calculate the probability that all of \( M \) randomly chosen samples are not affected by the high noise as

\[
P(M) = \prod_{i=0}^{N-1-i} \frac{N-I-i}{N-i}.
\]

For \( N = 128, I = 16, \) and \( M = 32 \) we get

\[
P(32) = 0.0112.
\]

It means that if we repeat the whole procedure 1000 times (1000 realizations) we can expect

\[
P(32) \times 1000 = 11.2,
\]

i.e., about 11 realizations when none of \( M \) signal samples is disturbed by the high noise.

The mean value is calculated as a sum over the set of possible amplitudes, weighted by the corresponding probabilities,

\[
\mu_x(n) = \mathbb{E}\{x(n)\} = \sum_{i=1}^{\infty} \xi_i P_{\xi(n)}(\xi_i).
\]  

(7.4)

If a random signal can assume continuous values in amplitude then we cannot define a probability that one exact signal amplitude value is assumed. In that case the probability density function \( p_{\xi(n)}(\xi) \) is used. It defines the probability that the \( n \)th signal sample \( x(n) \) takes a value within an infinitesimally small interval \( d\xi \) around \( \xi \),

\[
\text{Probability } \{\xi \leq x(n) < \xi + d\xi\} = p_{\xi(n)}(\xi)d\xi.
\]  

(7.5)

Properties of the probability density function are:

1) It is nonnegative, \( p_{\xi(n)}(\xi) \geq 0 \) for any \( \xi \)
2) Since Probability \( \{-\infty < x(n) < \infty\} = 1 \), then

\[
\int_{-\infty}^{\infty} p_{\xi(n)}(\xi)d\xi = 1.
\]

The probability of an event that a value of signal \( x(n) \) is within \( a \leq x(n) < b \) is

\[
\text{Probability } \{a \leq x(n) < b\} = \int_{b}^{a} p_{\xi(n)}(\xi)d\xi.
\]
Cumulative distribution $F(\chi)$ function is the probability that a signal $x(n)$ value is lower than $\chi$,

$$F(\chi) = \text{Probability} \{ x(n) < \chi \} = \int_{-\infty}^{\chi} p_{x(n)}(\xi) d\xi.$$  

Obviously $\lim_{\chi \to -\infty} F(\chi) = 0$, $\lim_{\chi \to +\infty} F(\chi) = 1$, and $F(a) \geq F(b)$ if $a > b$. Note that

$$p_{x(n)}(\xi) = \frac{dF(\xi)}{d\xi}.$$  

The expected value of a random variable $x(n)$ in terms of the probability density function, is

$$\mu_{x(n)} = E\{x(n)\} = \int_{-\infty}^{\infty} \xi p_{x(n)}(\xi) d\xi.$$  \hspace{1cm} (7.6)

7.1.3 Median

In addition to the mean value, a median is used for description of a set of random values. The median is a value in the middle of the set, after the members of the set are sorted. If we denote the sorted values of $x(n)$ as $s(n)$

$$s(n) = \text{sort} \{x(n)\}, \hspace{0.2cm} n = 1, 2, ..., N$$

then the median value is

$$\text{median} \{x(n)\} = s \left( \frac{N + 1}{2} \right), \text{for an odd } N.$$  

If $N$ is an even number then the median is defined as the mean value of two samples nearest to $(N - 1)/2$,

$$\text{median} \{x(n)\} = \frac{s \left( \frac{N}{2} \right) + s \left( \frac{N}{2} + 1 \right)}{2}, \text{for an even } N.$$  

The median will not be influenced by a possible small number of big outliers (signal values being significantly different from the values of the rest of data).
Example 7.6. Find the median of sets
(a) $A = \{-1,1,-2,4,6,-9,0\}$, (b) $B = \{-1,1,-1367,4,35,-9,0\}$, and (c)
of the signal $x(n)$ from Example 7.1.

★ (a) After sorting the values in set $A$ we get $A = \{-9,-2,-1,0,1,4,6\}$. Thus, median($A$) = 0. (b) In a similar way median($B$) = 0. The mean values of these data would significantly differ. (c) The sorted values of $x(n)$ are presented in Fig. 7.6. Since the number of samples of signal $x(n)$ is $N = 100$ there is no single sample in the middle of the sorted sequence. The middle is between sorted samples 50 and 51. Thus the median in this situation is defined as the mean value of the 50th and 51st sorted sample.

In some cases the number of big outliers is small. Thus the median will neglect many signal values that could produce a good estimate of the mean value. In that cases, the best choice would be to use not only the mid-value in the sorted signal, but several samples of the signal around its median and to calculate their mean, for odd $N$, as

$$\text{LSmean}\{x(n)\} = \frac{1}{2L+1} \sum_{i=-L}^{L} s \left( \frac{N+1}{2} + i \right).$$

With $L = (N - 1)/2$ all signal values are used and LSmean\{x(n)\} is the standard mean of a signal. With $L = 0$ the value of LSmean\{x(n)\} is the
standard median. In general, this way of signal parameters estimation is the L-statistics based estimation.

7.1.4 Variance

For random signals that take values from a discrete set, with known probabilities, the variance is defined as

\[
\sigma^2_x(n) = E\{ |x(n) - \mu_x(n)|^2 \} = \sum_\xi |\xi - \mu_x(n)|^2 P_{x(n)}(\xi).
\]

For a random signal \(x(n)\) whose values are available in \(M\) realizations the variance can be estimated as a mean square deviation of the signal values from their corresponding mean values \(\mu_x(n)\),

\[
\sigma^2_x(n) = \frac{1}{M} \left( |x_1(n) - \mu_x(n)|^2 + ... + |x_M(n) - \mu_x(n)|^2 \right).
\]

The standard deviation is a square root of the variance. The standard deviation can be estimated as a square root of the mean of squares of the centered data,

\[
\sigma_x(n) = \sqrt{\frac{1}{M} \left( |x_1(n) - \mu_x(n)|^2 + ... + |x_M(n) - \mu_x(n)|^2 \right)}.
\] (7.7)

For a small number of samples, this estimate tends to produce lower values of the standard deviation. Thus, an adjusted version, the sample standard deviation, is also used. It reads

\[
\sigma_x(n) = \sqrt{\frac{1}{M-1} \left( |(x_1(n) - \mu_x(n))|^2 + ... + |x_M(n) - \mu_x(n)|^2 \right)}.
\]

This form confirms the fact that in the case when only one sample is available, \(M = 1\), we should not be able to estimate the standard deviation.

For the case of random signals whose amplitude is continuous the variance, in terms of the probability density function \(p_{x(n)}(\xi)\), is

\[
\sigma^2_x(n) = \int_{-\infty}^{\infty} \left| \xi - \mu_x(n) \right|^2 p_{x(n)}(\xi) d\xi.
\]
Table 7.3
Random signal $z(n)$

<table>
<thead>
<tr>
<th>55</th>
<th>57</th>
<th>56</th>
<th>54</th>
<th>59</th>
<th>52</th>
<th>66</th>
<th>54</th>
<th>56</th>
<th>56</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>55</td>
<td>51</td>
<td>56</td>
<td>48</td>
<td>59</td>
<td>63</td>
<td>52</td>
<td>59</td>
<td>59</td>
</tr>
<tr>
<td>47</td>
<td>48</td>
<td>58</td>
<td>53</td>
<td>58</td>
<td>59</td>
<td>59</td>
<td>61</td>
<td>58</td>
<td>58</td>
</tr>
<tr>
<td>49</td>
<td>55</td>
<td>54</td>
<td>47</td>
<td>56</td>
<td>50</td>
<td>62</td>
<td>51</td>
<td>58</td>
<td>58</td>
</tr>
<tr>
<td>50</td>
<td>44</td>
<td>55</td>
<td>50</td>
<td>58</td>
<td>58</td>
<td>63</td>
<td>58</td>
<td>52</td>
<td>52</td>
</tr>
<tr>
<td>50</td>
<td>55</td>
<td>55</td>
<td>55</td>
<td>53</td>
<td>60</td>
<td>46</td>
<td>57</td>
<td>59</td>
<td>59</td>
</tr>
<tr>
<td>58</td>
<td>55</td>
<td>58</td>
<td>58</td>
<td>54</td>
<td>53</td>
<td>54</td>
<td>48</td>
<td>54</td>
<td>54</td>
</tr>
<tr>
<td>57</td>
<td>62</td>
<td>53</td>
<td>58</td>
<td>59</td>
<td>60</td>
<td>50</td>
<td>56</td>
<td>56</td>
<td>56</td>
</tr>
<tr>
<td>51</td>
<td>60</td>
<td>54</td>
<td>57</td>
<td>55</td>
<td>52</td>
<td>52</td>
<td>57</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>58</td>
<td>51</td>
<td>54</td>
<td>49</td>
<td>44</td>
<td>60</td>
<td>52</td>
<td>57</td>
<td>56</td>
<td>56</td>
</tr>
</tbody>
</table>

Figure 7.7  Random signal $z(n)$.

Example 7.7. For the signal $x(n)$ from Example 7.1 calculate the mean and variance. Compare it with the mean and variance of the signal $z(n)$ given in Table 7.3.
The mean value and variance for signal $x(n)$ are $\mu_x = 55.76$ and $\sigma^2_x = 314.3863$. The standard deviation is $\sigma_x = 17.7309$. It is a measure of signal value deviations from the mean value. For the signal $z(n)$ the mean value is $\mu_z = 55.14$ (very close to $\mu_x$), while the variance is $\sigma^2_z = 18.7277$ and the standard deviation is $\sigma_z = 4.3275$. Deviations of $z(n)$ from the mean value are much smaller. If signals $x(n)$ and $z(n)$ were measurements of the same physical value, then the individual measurements from $z(n)$ would be much more reliable than the individual measurements from $x(n)$.

Example 7.8. A random signal $x(n)$ can take values from the set $\{0,1,2,3,4,5\}$. It is known that for $k = 1,2,3,4$ the probability of $x(n) = k$ is twice higher than the probability of $x(n) = k + 1$. Find the probabilities $P\{x(n) = k\}$. Find the mean value and variance of signal.

Assume that $P\{x(n) = 5\} = A$. Then the probabilities that $x(n)$ takes a value $k$ are

<table>
<thead>
<tr>
<th>$k$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P{x(n) = k}$</td>
<td>$32A$</td>
<td>$16A$</td>
<td>$8A$</td>
<td>$4A$</td>
<td>$2A$</td>
<td>$A$</td>
</tr>
</tbody>
</table>

Constant $A$ can be found from $\sum_k P\{x(n) = k\} = 1$. It results in $A = 1/63$. Now we have

$$\mu_{x(n)} = \sum_k k P\{x(n) = k\} = \frac{19}{21}$$

$$\sigma^2_{x(n)} = \sum_k \left(k - \frac{19}{21}\right)^2 P\{x(n) = k\} = \frac{626}{441}.$$

Example 7.9. Consider a real-valued random signal $x(n)$ with samples whose values are uniformly distributed over interval $-1 \leq x(n) \leq 1$. a) Find the mean value and variance of the signal samples. b) Signal $y(n)$ is obtained as $y(n) = x^2(n)$. Find the mean value and variance of signal $y(n)$.

Since the random signal $x(n)$ is uniformly distributed, its probability density function is of the form

$$p_{x(n)}(\xi) = \begin{cases} A & \text{for } |\xi| \leq 1 \\ 0 & \text{for } |\xi| > 1 \end{cases}.$$

Constant $A = 1/2$ is obtained from $\int_{-\infty}^{\infty} p_{x(n)}(\xi) d\xi = 1$. Now we have

$$\mu_{x(n)} = \int_{-\infty}^{\infty} \xi p_{x(n)}(\xi) d\xi = \int_{-1}^{1} \frac{1}{2} \xi d\xi = 0$$

$$\sigma^2_{x(n)} = \int_{-\infty}^{\infty} (\xi - \mu_{x(n)})^2 p_{x(n)}(\xi) d\xi = \int_{-1}^{1} \frac{1}{2} \xi^2 d\xi = \frac{1}{3}.$$
The probability that \( y(n) \) is not higher than \( \xi \) is
\[
F_y(\xi) = P\{y(n) \leq \xi\} = P\{x^2(n) \leq \xi\} = P\{-\sqrt{\xi} < x(n) \leq \sqrt{\xi}\} = \left\{
\begin{array}{ll}
0 & \text{for } \xi \leq 0 \\
\int_{-\sqrt{\xi}}^{\sqrt{\xi}} p_x(\xi) \, d\xi & \text{for } 0 < \xi < 1 \\
1 & \text{for } \xi \geq 1
\end{array}\right.
\]

since \( y(n) \leq \xi \) when \( x^2(n) \leq \xi \). The probability density function is
\[
p_{y(n)}(\xi) = \frac{dF(\xi)}{d\xi} = \left\{
\begin{array}{ll}
\frac{1}{2\sqrt{\xi}} & \text{for } 0 < \xi \leq 1 \\
0 & \text{otherwise.}
\end{array}\right.
\]

The mean value and variance of signal \( y(n) \) are
\[
\mu_{y(n)} = \int_{0}^{1} \frac{\xi}{2\sqrt{\xi}} \, d\xi = \frac{1}{3}
\]
\[
\sigma_{y(n)}^2 = \int_{0}^{1} (\xi - \frac{1}{3})^2 \frac{1}{2\sqrt{\xi}} \, d\xi = \frac{4}{45}.
\]

Note: Generalize for \( z(n) = f(x(n)) \). \( \Box \)

As an introduction to the second-order statistics consider two signals \( x(n) \) and \( y(n) \) with continuous amplitude values. Probability that the \( n \)th signal sample \( x(n) \) takes a value within \( \xi < x(n) < \xi + d\xi \) and that \( y(m) \) takes a value within \( \zeta < y(m) < \zeta + d\zeta \) is
\[
\text{Probability}\{\xi < x(n) < \xi + d\xi, \zeta < y(m) < \zeta + d\zeta\} = p_{x(n),y(m)}(\xi,\zeta) \, d\xi \, d\zeta,
\]
where \( p_{x(n),y(m)}(\xi,\zeta) \) is the joint probability density function. The probability of an event \( a < x(n) < b \) and \( c < y(m) < d \) is
\[
\text{Probability}\{a < x(n) < b, c < y(m) < d\} = \int_{a}^{b} \int_{c}^{d} p_{x(n),y(m)}(\xi,\zeta) \, d\xi \, d\zeta.
\]

For mutually independent signals \( p_{x(n),y(m)}(\xi,\zeta) = p_{x(n)}(\xi) \, p_{y(m)}(\zeta) \). A special case of the previous relations is obtained when \( y(m) = x(m) \).

**Example 7.10.** Signal \( x(n) \) is defined as \( x(n) = a(n) + b(n) + c(n) \) where \( a(n) \), \( b(n) \), and \( c(n) \) are mutually independent random signals with a uniform probability density function over the range \([-1,1]\). Find the probability density function of signal \( x(n) \), its mean \( \mu_x \), and variance \( \sigma_x^2 \).
Consider a sum of two independent random signals \( s(n) = a(n) + b(n) \). The probability that \( s(n) = a(n) + b(n) < \theta \) can be calculated from the joint probability distribution of \( a(n) \) and \( b(n) \) as

\[
F(\theta) = P\{s(n) < \theta\} = \text{Probability}\{-\infty < a(n) < \infty, -\infty < a(n) + b(n) \leq a < \theta\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\theta-\zeta} p_{a(n),b(n)}(\xi,\zeta) \, d\xi \, d\zeta = \int_{-\infty}^{\infty} p_{b(n)}(\zeta) \int_{-\infty}^{\theta-\zeta} p_{a(n)}(\xi) \, d\xi \, d\zeta.
\]

Now we can calculate the probability density function of \( s(n) \) as a derivative

\[
p_{s(n)}(\theta) = \frac{dF(\theta)}{d\theta} = \int_{-\infty}^{\infty} p_{b(n)}(\zeta) \frac{d}{d\theta} \int_{-\infty}^{\theta-\zeta} p_{a(n)}(\xi) \, d\xi \, d\zeta = \int_{-\infty}^{\infty} p_{b(n)}(\zeta) p_{a(n)}(\theta - \zeta) \, d\zeta = p_{b(n)}(\theta) \ast_\theta p_{a(n)}(\theta),
\]

meaning that the probability density function of a sum of two independent random variables is a convolution of the individual probability density functions. In a similar way we can include the third signal and obtain

\[
p_{x(n)}(\theta) = p_{c(n)}(\theta) \ast_\theta p_{b(n)}(\theta) \ast_\theta p_{a(n)}(\theta),
\]

\[
p_{x(n)}(\theta) = \begin{cases} \frac{(\theta+3)^2}{16} & \text{for } -3 \leq \theta \leq -1 \\ \frac{3-\theta^2}{16} & \text{for } -1 < \theta \leq 1 \\ \frac{(\theta-3)^2}{16} & \text{for } 1 < \theta \leq 3 \\ 0 & \text{for } |\theta| > 3 \end{cases}.
\]

The mean value and variance can be calculated from \( p_{x(n)}(\theta) \), or in direct way, as

\[
\mu_x = E\{x(n)\} = E\{a(n)\} + E\{b(n)\} + E\{c(n)\} = 0
\]

\[
\sigma_x^2 = E\{(x(n) - \mu_x)^2\} = E\{(a(n) + b(n) + c(n))^2\} = E\{a(n)^2\} + E\{b(n)^2\} + E\{c(n)^2\} + 2(\mu_a\mu_b + \mu_a\mu_c + \mu_b\mu_c) = \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1.
\]
7.2 SECOND-ORDER STATISTICS

7.2.1 Correlation and Covariance

Second-order statistics deals with two samples of random signals. For a signal \( \{ x_i(n) \} \), \( n = 1, 2, ..., N \) and \( i = 1, 2, ..., M \), being the number of realizations of this signal, the autocorrelation function is defined by

\[
r_{xx}(n, m) = E\{x(n)x^*(m)\} = \frac{1}{M} \sum_{i=1}^{M} x_i(n)x_i^*(m).
\] (7.8)

If the probability that a real-valued random signal \( x(n) \) assumes a value \( \xi_1 \) and that \( x(m) \) assumes \( \xi_2 \) is \( P_{x(n),x(m)}(\xi_1, \xi_2) \) then

\[
r_{xx}(n, m) = \sum_{\xi_1} \sum_{\xi_2} \xi_1 \xi_2 P_{x(n),x(m)}(\xi_1, \xi_2). \] (7.9)

For a real-valued random signal with continuous amplitudes and the second-order probability density function \( p_{x(n),x(m)}(\xi_1, \xi_2) \), the autocorrelation is

\[
r_{xx}(n, m) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \xi_1 \xi_2 p_{x(n),x(m)}(\xi_1, \xi_2) d\xi_1 d\xi_2. \] (7.10)

If the real-valued random variables \( x(n) \) and \( x(m) \) are statistically independent, then \( p_{x(n),x(m)}(\xi_1, \xi_2) = p_{x(n)}(\xi_1) p_{x(m)}(\xi_2) \) and \( r_{xx}(n, m) = \mu_x(n) \mu_x(m) \).

The autocovariance function is defined by

\[
c_{xx}(n, m) = E\{ (x(n) - \mu_x(n))(x(m) - \mu_x(m))^* \}
= \frac{1}{M} \sum_{i=1}^{M} (x_i(n) - \mu_x(n))(x_i(m) - \mu_x(m))^*.
\] (7.11)

It may be easily shown that

\[
c_{xx}(n, m) = E\{ (x(n) - \mu_x(n))(x(m) - \mu_x(m))^* \} = r_{xx}(n, m) - \mu_x(n) \mu_x^*(m).
\]

Value of the autocovariance for \( m = n \) is the variance

\[
c_{xx}^2(n) = E\{ |x(n) - \mu_x(n)|^2 \} = r_{xx}(n, n) - |\mu_x(n)|^2.
\] (7.12)
The cross-correlation and the cross-covariance of two signals \( x(n) \) and \( y(n) \) are defined as

\[
\rho_{xy}(n,m) = \mathbb{E}\{x(n)y^*(m)\}
\]

and

\[
\sigma_{xy}(n,m) = \mathbb{E}\{ (x(n) - \mu_x(n))(y(m) - \mu_y(m))^* \} = \rho_{xy}(n,m) - \mu_x(n)\mu_y^*(m).
\] (7.13)

### 7.2.2 Stationarity and Ergodicity

Signals whose first-order and second-order statistics are invariant to a shift in time are called wide sense stationary (WSS) signals. For the WSS signals holds

\[
\mu_x(n) = \mathbb{E}\{x(n)\} = \mu_x
\]

\[
r_{xx}(n,m) = \mathbb{E}\{x(n)x^*(m)\} = r_{xx}(n-m).
\] (7.14)

A signal is stationary in the strict sense (SSS) if all order statistics are invariant to a shift in time. The relations introduced for the second-order statistics may be extended to the higher-order statistics. For example, the third-order moment of a signal \( x(n) \) is defined by

\[
M_{xxx}(n,m,l) = \mathbb{E}\{x(n)x^*(m)x^*(l)\}.
\] (7.15)

For stationary signals it assumes the form

\[
M_{xxx}(m,l) = \mathbb{E}\{x(n)x^*(n-m)x^*(n-l)\}.
\]

In order to calculate the third-order moment we should know the third-order statistics, like the third-order probability \( P_{x(n),x(m),x(l)}(\xi_1,\xi_2,\xi_3) \) or probability density function.

For a random process, as collection of all realizations of a random signal along with its probabilistic description, we say that it is ergodic if its parameters can be estimated by averaging over time instead of over realizations. The process is ergodic in parameter \( \beta \) if that particular parameter can be estimated by averaging over time instead of over realizations. If a random signal \( x(n) \) is a realization of a process ergodic in mean then

\[
\mu_x(n) = \lim_{M\to\infty} \frac{1}{M} (x_1(n) + x_2(n) + ... + x_M(n))
\]

\[
= \lim_{N\to\infty} \frac{1}{N} (x_i(n) + x_i(n-1) + ... + x_i(n-N+1)).
\]
7.2.3 Power Spectral Density

For stationary signals the autocorrelation function is
\[ r_{xx}(n) = E\{x(n + m)x^*(m)\} = r_{xx}(n). \]

The Fourier transform of the autocorrelation function of a WSS signal is the power spectral density
\[ S_{xx}(e^{j\omega}) = \sum_{n=-\infty}^{\infty} r_{xx}(n)e^{-j\omega n} \quad (7.16) \]
\[ r_{xx}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_{xx}(e^{j\omega})e^{j\omega n}d\omega. \quad (7.17) \]

Integral of \( S_{xx}(e^{j\omega}) \) over frequency,
\[ \frac{1}{2\pi} \int_{-\pi}^{\pi} S_{xx}(e^{j\omega})d\omega = r_{xx}(0) = E\{|x(n)|^2\}, \quad (7.18) \]
is equal to the average power of the random signal.

**Example 7.11.** Find the mean, autocorrelation, and power spectral density of the random signal
\[ x(n) = \sum_{k=1}^{K} a_k e^{j(\omega_k n + \theta_k)}, \]
where \( \theta_k \) are random variables uniformly distributed over \(-\pi < \theta_k \leq \pi\). All random variables are statistically independent. Frequencies \( \omega_k \) are \(-\pi < \omega_k \leq \pi\) for each \( k \).

\[ \star \] The mean value is
\[ \mu_x = \sum_{k=1}^{K} a_k E\{e^{j(\omega_k n + \theta_k)}\} = \sum_{k=1}^{K} a_k \int_{-\pi}^{\pi} e^{j(\omega_k n + \theta_k)}d\theta_k = 0. \]

The autocorrelation is
\[ r_{xx}(n) = E\{\sum_{k=1}^{K} a_k e^{j(\omega_k n + \theta_k)} \sum_{k=1}^{K} a_k e^{-j(\omega_k m + \theta_k)}\} = \sum_{k=1}^{K} a_k^2 e^{j\omega_k n}, \]
while the power spectral density for \(-\pi < \omega \leq \pi\) is
\[ S_{xx}(e^{j\omega}) = FT\{r_{xx}(n)\} = 2\pi \sum_{k=1}^{K} a_k^2 \delta(\omega - \omega_k). \]
we can write double sum are functions of \( L_j \) when the summation line is the main diagonal of area performed along the lines where domain defined by Double summation is performed within a square in the two-dimensional For a stationary signal \( (f_{\text{in}}) \) of power spectral density, will not produce the same result, in general. This relation leads to another definition of the power spectral density of random discrete-time signals

\[
P_{xx}(e^{j\omega}) = \lim_{N \to \infty} \frac{1}{2N + 1} \mathbb{E} \left\{ \left| X_N(e^{j\omega}) \right|^2 \right\} \tag{7.19}
\]

\[
= \lim_{N \to \infty} \frac{1}{2N + 1} \mathbb{E} \left\{ \sum_{n=-N}^{N} x(n)e^{-j\omega n} \right\}^2.
\]

Different notation is used since the previous two definitions, (7.16) and (7.19) of power spectral density, will not produce the same result, in general. We can write

\[
P_{xx}(e^{j\omega}) = \lim_{N \to \infty} \frac{1}{2N + 1} \mathbb{E} \left\{ \sum_{m=-N}^{N} \sum_{n=-N}^{N} x(m)x^*(n)e^{-j\omega(m-n)} \right\}.
\]

For a stationary signal

\[
P_{xx}(e^{j\omega}) = \lim_{N \to \infty} \frac{1}{2N + 1} \sum_{m=-N}^{N} \sum_{n=-N}^{N} r_{xx}(m-n)e^{-j\omega(m-n)}.
\]

Double summation is performed within a square in the two-dimensional domain defined by \(-N \leq m \leq N\), \(-N \leq n \leq N\). Since the terms within double sum are functions of \((m-n)\) only, then the summation could be performed along the lines where \((m-n) = k\) is constant. For \((m-n) = k = 0\) the summation line is the main diagonal of area \(-N \leq m \leq N\), \(-N \leq n \leq N\). Along this diagonal there are 2\(N + 1\) points where \(r_{xx}(m-n)e^{-j\omega(m-n)} = r_{xx}(0)\). For the nearest subdiagonals of \(-N \leq m \leq N\), \(-N \leq n \leq N\) when \((m-n) = k = \pm 1\) there are 2\(N\) points where \(r_{xx}(m-n)e^{-j\omega(m-n)} = r_{xx}(\pm 1)e^\pm j\omega\). For arbitrary lines \((m-n) = \pm k\), with \(|k| \leq 2N\), there are 2\(N + 1 - |k|\) terms with \(r_{xx}(m-n)e^{-j\omega(m-n)} = r_{xx}(\pm k)e^\pm jk\omega\). It means that we can write

\[
P_{xx}(e^{j\omega}) = \lim_{N \to \infty} \frac{1}{2N + 1} \sum_{k=-2N}^{2N} (2N + 1 - |k|)r_{xx}(k)e^{-j\omega k}
\]

\[
= \lim_{N \to \infty} \sum_{k=-2N}^{2N} \left(1 - \frac{|k|}{2N + 1}\right)r_{xx}(k)e^{-j\omega k} = \lim_{N \to \infty} \sum_{k=-2N}^{2N} w_B(k)r_{xx}(k)e^{-j\omega k}.
\]
Function \( w_B(k) \) corresponds to a Bartlett window over the calculation interval. If the values of autocorrelation function \( r_{xx}(k) \) are such that the second part of the sum \( \sum_k |k|/(2N + 1)r_{xx}(k)e^{-j\omega k} \) is negligible as compared to \( \sum_k r_{xx}(k)e^{-j\omega k} \) then

\[
P_{xx}(e^{j\omega}) = \lim_{N \to \infty} \sum_{k=-2N}^{2N} r_{xx}(k)e^{-j\omega k} = \text{FT}\{r_{xx}(n)\} = S_{xx}(e^{j\omega}).
\]

This is true for \( r_{xx}(k) = C\delta(k) \). Otherwise \( P_{xx}(e^{j\omega}) \) is a smoothed version of \( S_{xx}(e^{j\omega}) \). Note that \( P_{xx}(e^{j\omega}) \) is always nonnegative, by definition (for a numeric illustration see Example 7.23).

### 7.3 NOISE

In many applications, the desired signal is disturbed by various forms of random signals, caused by numerous factors in the signal sensing, transmission, and/or processing. Often, a cumulative influence of these factors, disturbing useful signal, is described by an equivalent random signal, called noise. In most cases we will use a notation \( \varepsilon(n) \) for these kinds of signals. They model a random, multiple source, disturbance.

A noise is said to be **white** if its values are uncorrelated

\[
r_{\varepsilon\varepsilon}(n,m) = \sigma_{\varepsilon}^2 \delta(n-m) \tag{7.20}
\]

\[
S_{\varepsilon\varepsilon}(e^{j\omega}) = \text{FT}\{r_{xx}(n)\} = \sigma_{\varepsilon}^2.
\]

Spectral density of this kind of noise is constant (like it is the case in the white light). If this property is not satisfied, then the power spectral density is not constant. Such a noise is referred to as **colored**.

Regarding to the distribution of noise \( \varepsilon(n) \) amplitudes the most common types of noise in signal processing are: uniform, binary, Gaussian, and impulsive noise.

#### 7.3.1 Uniform Noise

The uniform noise is a signal with the probability density function

\[
p_{\varepsilon(n)}(\xi) = \frac{1}{\Delta}, \text{ for } -\Delta/2 \leq \xi < \Delta/2 \tag{7.21}
\]
and $p_{\varepsilon(n)}(\xi) = 0$ elsewhere, Fig. 7.8. Its variance is

$$\sigma_{\varepsilon}^2 = \frac{\Delta^2}{12}.$$

This kind of noise is used to model rounding errors in the amplitude quantization of a signal. It indicates that all errors within $-\Delta/2 \leq \xi < \Delta/2$ are equally probable.

### 7.3.2 Binary Noise

Random binary sequence, or binary noise, is a stochastic signal which randomly assumes one of two fixed signal values. Assume that the noise $\varepsilon(n)$ values are, for example, $\{-1, 1\}$ and that the probability that $\varepsilon(n)$ assumes value 1 is $p$. The mean of this noise is

$$\mu_{\varepsilon} = \sum_{\xi=-1,1} \xi P_{\varepsilon}(\xi) = (-1)(1 - p) + 1 \cdot p = 2p - 1.$$

The variance is

$$\sigma_{\varepsilon}^2 = \sum_{\xi=-1,1} (\xi - \mu_{\varepsilon})^2 P_{\varepsilon}(\xi) = 4p(1 - p).$$

A special case is when the values from the set $\{-1, 1\}$ are equally probable, that is when $p = 1/2$. Then we get $\mu_{\varepsilon} = 0$ and $\sigma_{\varepsilon}^2 = 1$. 
Example 7.12. Consider a set of $N \to \infty$ balls. Equal number of balls is marked with 1 (or white) and 0 (or black). A random signal $x(n)$ corresponds to drawing of four balls in a row. It has four values $x(0), x(1), x(2), \text{and } x(3)$. Signal values $x(n)$ are equal to the marks on the drawn balls. Write all possible realizations of $x(n)$. If $k$ is the number of appearances of value 1 in the signal, write the probabilities for each value of $k$.

★ Signal realizations, with the number $k$ being the number of appearances of digit 1 in each signal realization, are given in the next table.

<table>
<thead>
<tr>
<th>$x(0)$</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x(1)$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$x(2)$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$x(2)$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Possible values of $k$ are 0, 1, 2, 3, 4 with corresponding probabilities

\[
P(0) = 1 \cdot \left( \frac{1}{1} \right)^4, \quad P(1) = 4 \cdot \left( \frac{1}{2} \right)^3 \left( \frac{1}{2} \right), \quad P(2) = 6 \cdot \left( \frac{1}{2} \right)^2 \left( \frac{1}{2} \right)^2, \quad P(3) = 4 \cdot \left( \frac{1}{2} \right) \left( \frac{1}{2} \right)^3, \quad \text{and} \quad P(4) = 1 \cdot \left( \frac{1}{2} \right)^4.
\]

These probabilities can be considered as the terms of a binomial expression

\[
(a + b)^4 = \binom{4}{0} a^4 + \binom{4}{1} a^3 b + \binom{4}{2} a^2 b^2 + \binom{4}{3} a b^3 + \binom{4}{4} b^4
\]

with $a = 1/2$ and $b = 1/2$. For the case when $N$ is a finite number see Problem 7.6.

An interesting form of the random variable that can assume only two possible values \{-1, 1\} or \{No, Yes\} or \{A, B\} is the binomial random variable. It has been introduced through the previous simple example. In general, if a signal $x(n)$ assumes value $B$ from the set \{A, B\} with probability $p$, then the probability that there is exactly $k$ values of $B$ in a sequence of $N$ samples of $x(n)$ is

\[
P(k) = \binom{N}{k} p^k (1 - p)^{N-k} = \frac{N!}{k!(N-k)!} p^k (1 - p)^{N-k}.
\]

This is a binomial coefficients form.
The expected value of the number of appearances of event $B$ in $N$ samples, denoted by $y$, is

$$\mu_y = E\{y\} = \sum_{k=0}^{N} kP(k)$$

$$= \sum_{k=0}^{N} k \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k}.$$ 

Since the first term in summation is 0 we will shift the summation for one and reindex it to

$$\mu_y = E\{y\} = \sum_{k=0}^{N-1} (k+1) \frac{N(N-1)!}{(k+1)!((N-(k+1))!} p^{k+1} (1-p)^{N-(k+1)}$$

$$= Np \sum_{k=0}^{N-1} \frac{(N-1)!}{k!((N-1)-k)!} p^k (1-p)^{(N-1)-k}.$$ 

The sum in the last expression is equal to

$$1 = (p + (1 - p))^{N-1} = \sum_{k=0}^{N-1} \binom{N-1}{k} p^k (1-p)^{(N-1)-k}$$

$$= \sum_{k=0}^{N-1} \frac{(N-1)!}{k!((N-1)-k)!} p^k (1-p)^{(N-1)-k}$$

resulting, with $p + (1 - p) = 1$, into

$$\mu_y = E\{y\} = Np.$$ 

As we could write from the beginning, the expected value of the number of appearances of an event $B$, whose probability is $p$, in $N$ realizations is $E\{y\} = Np$. This derivation was performed not only to prove this fact, but it will lead us to the next step in deriving the variance of the event $y$, by using the expected value of the product of $y$ and $y-1$,

$$E\{y(y-1)\} = \sum_{k=0}^{N} k(k-1)P(k)$$

$$= \sum_{k=0}^{N} k(k-1) \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k}.$$ 

Since the first two terms are 0 we can reindex the summation into
\[
E\{y(y - 1)\} = \sum_{k=0}^{N-2} (k+2)(k+1) \frac{N!}{(k+2)!(N-2-k)!} p^{k+2}(1-p)^{N-2-k}
= N(N-1)p^2 \sum_{k=0}^{N-2} \frac{(N-2)!}{k!(N-2-k)!} p^k(1-p)^{N-2-k}.
\]

The relation
\[
\sum_{k=0}^{N-2} \frac{(N-2)!}{k!(N-2-k)!} p^k(1-p)^{N-2-k} = (p + (1-p))^{N-2} = 1
\]
is used to get
\[
E\{y(y - 1)\} = N(N-1)p^2.
\]
The variance of \(y\) follows from
\[
\sigma_y^2 = E\{y^2\} - (E\{y\})^2
= E\{y(y - 1)\} + E\{y\} - (E\{y\})^2
= Np(1-p).
\]

Therefore, in a sequence of \(N\) values of signal \(x(n)\) that can assume values \(\{A,B\}\) the mean value and variance of appearances of \(B\) divided by \(N\) will be
\[
\mu_y = \frac{Np}{N} = p,
\]
\[
\sigma_y^2 = \frac{Np(1-p)}{N^2} = \frac{p(1-p)}{N}.
\]

Increasing the number of the total values \(N\) the variance will be lower and a finite set \(x(n)\) will produce a more reliable mean value \(p\).

### 7.3.3 Gaussian Noise

The Gaussian (normal) noise is used to model a disturbance caused by many small independent factors. Namely, the central limit theorem states that a sum of a large number of statistically independent random variables, with any distribution, obeys to the Gaussian (normal) distribution.

The Gaussian zero-mean noise has the probability density function
\[
p_{\varepsilon(n)}(\xi) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\xi^2/(2\sigma^2)}.
\]
Variance of this noise is $\sigma^2_{\epsilon}$. It is left to reader to prove this by evaluating corresponding integral. For the Gaussian noise with mean $\mu$ and variance $\sigma^2_{\epsilon}$ we can use notation $\mathcal{N}(\mu, \sigma^2_{\epsilon})$.

The probability that the amplitude of a zero-mean Gaussian random variable takes a value smaller than $\lambda$ is

$$\text{Probability}\{|\epsilon(n)| < \lambda\} = \frac{1}{\sigma_{\epsilon}\sqrt{2\pi}} \int_{-\lambda}^{\lambda} e^{-\xi^2/(2\sigma^2_{\epsilon})} d\xi = \text{erf}\left(\frac{\lambda}{\sqrt{2}\sigma_{\epsilon}}\right) \tag{7.23}$$

where

$$\text{erf}(\lambda) = \frac{2}{\sqrt{\pi}} \int_{0}^{\lambda} e^{-\xi^2} d\xi$$

is the error function.

Commonly used probabilities that the absolute value of the noise is within the standard deviation, two standard deviations (two-sigma rule), or three standard deviations are:

$$\text{Probability}\{-\sigma_{\epsilon} < \epsilon(n) < \sigma_{\epsilon}\} = \text{erf}(1/\sqrt{2}) = 0.6827, \tag{7.24}$$

$$\text{Probability}\{-2\sigma_{\epsilon} < \epsilon(n) < 2\sigma_{\epsilon}\} = \text{erf}(\sqrt{2}) = 0.9545,$$

$$\text{Probability}\{-3\sigma_{\epsilon} < \epsilon(n) < 3\sigma_{\epsilon}\} = \text{erf}(3/\sqrt{2}) = 0.9973.$$

**Example 7.13.** Given 12 measurements of a Gaussian zero-mean noise \{-0.7519, 1.5163, -0.0326, -0.4251, 0.5894, -0.0628, -2.0220, -0.9821, 0.6125, -0.0549, -1.1187, 1.6360\}, estimate the probability that the absolute value of this noise will be smaller than 2.5.
The standard deviation of this noise could be estimated by using (7.7) with \( \mu = 0 \) and \( N = 12 \). It is \( \sigma = 1.031 \). Thus, the absolute value of this noise will be smaller than 2.5 with probability

\[
P = \frac{1}{1.031\sqrt{2\pi}} \int_{-2.5}^{2.5} e^{-\xi^2/(2\cdot1.031^2)} d\xi = \text{erf}(2.5/(\sqrt{2} \cdot 1.031)) = 0.9847.
\]

Example 7.14. Consider a signal \( s(n) = A\delta(n - n_0) \) and a zero-mean Gaussian noise \( \epsilon(n) \) with variance \( \sigma^2 \) within the interval \( 0 \leq n \leq N - 1 \), where \( n_0 \) is a constant integer within \( 0 \leq n_0 \leq N - 1 \). Find the probability of event \( A \) that a maximum value of \( x(n) = s(n) + \epsilon(n) \) is obtained at \( n = n_0 \).

\[
\text{Probability density function for any sample } x(n), n \neq n_0, \text{ is}
\]

\[
p_{x(n),n \neq n_0}(\xi) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\xi^2/(2\sigma^2)}.
\]

The probability that any of these samples is smaller than a value of \( \lambda \) could be defined by using (7.23)

\[
P^- (\lambda) = \text{Probability}\{ x(n) < \lambda, n \neq n_0 \}
\]

\[
= \text{Probability}\{ x(n) < 0, n \neq n_0 \} + \text{Probability}\{ 0 \leq x(n) < \lambda, n \neq n_0 \}
\]

\[
= 0.5 + 0.5 \text{erf}(\lambda/(\sqrt{2} \sigma)).
\]

Since the random variables \( x(n), 0 \leq n \leq N - 1, n \neq n_0 \), are statistically independent, then the probability that all of them are smaller than a value
of $\lambda$ is

$$P_{N-1}^- (\lambda) = \text{Probability}\{\text{All } N-1 \text{ values of } x(n) < \lambda, n \neq n_0\}$$

$$= \left[0.5 + 0.5 \text{erf}(\lambda/ (\sqrt{2}\sigma))\right]^{N-1}.$$ 

The probability density function of the sample $x(n_0)$ is a Gaussian function with the mean value $A$,

$$p_x(x(n_0))(\xi) = \frac{1}{\sigma \sqrt{2\pi}} e^{-((\xi-A)^2)/(2\sigma^2)}.$$ 

The probability that the random variable $x(n_0)$ takes a value around $\lambda$,

$$P_{n_0}^+ (\lambda) = \text{Probability}\{\lambda < x(n_0) < \lambda + d\lambda\} = \frac{1}{\sigma \sqrt{2\pi}} e^{-((\xi-A)^2)/(2\sigma^2)} d\lambda$$ 

$$(7.25)$$ 

The probability that all values of $x(n), 0 \leq n \leq N-1, n \neq n_0$ are smaller than $\lambda$ and that, at the same time, $\lambda < x(n_0) < \lambda + d\lambda$ is

$$P_A (\lambda) = P_{N-1}^- (\lambda) P_{n_0}^+ (\lambda) = \left[0.5 + 0.5 \text{erf}(\lambda/ (\sqrt{2}\sigma))\right]^{N-1} \frac{1}{\sigma \sqrt{2\pi}} e^{-((\xi-A)^2)/(2\sigma^2)} d\lambda,$$

while the total probability that all $x(n), 0 \leq n \leq N-1, n \neq n_0$ are bellow $x(n_0)$ is an integral over all possible values of $\lambda$

$$P_A = \int_{-\infty}^{\infty} P_A (\lambda) = \int_{-\infty}^{\infty} \left[0.5 + 0.5 \text{erf}(\lambda/ (\sqrt{2}\sigma))\right]^{N-1} \frac{1}{\sigma \sqrt{2\pi}} e^{-((\xi-A)^2)/(2\sigma^2)} d\lambda.$$ 

$$(7.26)$$ 

Example 7.15. Random signal $x(n)$ is a Gaussian noise with the mean $\mu_x = 1$ and variance $\sigma_x^2 = 1$. A random sequence $y(n)$ is obtained by omitting samples from signal $x(n)$ that are either negative or higher than 1. Find the probability density function of sequence $y(n)$. Find its $\mu_y$ and $\sigma_y$.

★The probability density function for the sequence $y(n)$ is

$$p_y(y(n))(\xi) = \begin{cases} 
B \frac{1}{\sqrt{2\pi}} e^{-((\xi-1)^2)/(2\sigma^2)} & \text{for } 0 < \xi \leq 1 \\
0 & \text{otherwise}
\end{cases}$$

Constant $B$ can be calculated from $\int_{-\infty}^{\infty} p_y(y(n))(\xi) d\xi = 1$, resulting in

$$B = 2 / \text{erf}(\frac{1}{\sqrt{2}}).$$
Now we have
\[
\mu_{y(n)} = \int_0^\zeta \frac{2}{\text{erf}(\frac{1}{\sqrt{2}})} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\zeta-1)^2}{2}} d\zeta = 1 - \frac{\sqrt{2}(1 - e^{-1/2})}{\sqrt{\pi \text{erf}(\frac{1}{\sqrt{2}})}} \approx 0.54
\]
\[
\sigma^2_{y(n)} = \int_0^\zeta (\zeta - \mu_{y(n)})^2 \frac{2}{\text{erf}(\sqrt{2})} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\zeta-1)^2}{2}} d\zeta \approx 0.08.
\]

**Example 7.16.** Consider a random signal \( x(n) \) that can assume values \{No, Yes\} with probabilities \( 1-p \) and \( p \). If a random realization of this signal is available with \( N = 1000 \) samples and we obtained that the event Yes appeared 555 times find the interval where the true \( p \) will be with probability of 0.95. Denote by \( y \) the number of observed Yes values divided by \( N \). We can assume that the mean value estimates for various realizations are Gaussian distributed.

\[\star\text{This is a binomial random variable with the mean } p \text{ and the variance }\]
\[\sigma^2_y = \frac{p(1-p)}{N} \approx \frac{555}{1000}(1 - \frac{555}{1000}) = \frac{0.2470}{1000}\]
\[\sigma_y = 0.0157.\]

Therefore the estimated value
\[\hat{p} = \frac{555}{1000}\]
is within the range
\[\hat{p} = 0.555 \in [p - 2\sigma_y, p + 2\sigma_y]\]
\[= [p - 0.0314, p + 0.0314]\]
with probability 0.95, i.e.,
\[-0.0314 \leq 0.555 - p \leq 0.0314\]
\[|0.555 - p| \leq 0.0314.\]

with the same probability. The true value is around 55.5% within 3.14% range (from 52.36% to 58.64%) with probability 0.95. By increasing the value of \( N \) we can reduce the margin of estimation error. However, about 1000 values are commonly used for various opinion poll estimations. \( \Box \)
7.3.4 Complex Gaussian Noise and Rayleigh Distribution

In many applications the complex-valued Gaussian noise is used as a model for disturbance. Its form is

\[ \varepsilon(n) = \varepsilon_r(n) + j\varepsilon_i(n) \]

where \( \varepsilon_r(n) \) and \( \varepsilon_i(n) \) are real-valued Gaussian noises. Commonly it is assumed that they are zero-mean, independent, with identical distributions (i.i.d.), and variance \( \sigma^2/2 \).

The mean value of this noise is

\[ \mu_\varepsilon = E\{\varepsilon(n)\} = E\{\varepsilon_r(n)\} + jE\{\varepsilon_i(n)\} = 0 + j0. \]

The variance is

\[ \sigma^2_\varepsilon = E\{|\varepsilon(n)|^2\} = E\{\varepsilon(n)\varepsilon^*(n)\} = E\{\varepsilon_r(n)\varepsilon_r(n)\} + E\{\varepsilon_i(n)\varepsilon_i(n)\} = \sigma^2. \]

The amplitude of Gaussian noise \( |\varepsilon(n)| \) is an important parameter in many detection problems. The probability density function of the complex-Gaussian noise amplitude is

\[ p_{|\varepsilon(n)|}(\xi) = \frac{2\xi}{\sigma^2} e^{-\xi^2/\sigma^2} u(\xi). \]

The probability density function \( p_{|\varepsilon(n)|}(\xi) \) is called the Rayleigh distribution.

In order to prove the previous relation consider the probability density function of \( \varepsilon_r(n) \) and \( \varepsilon_i(n) \). Since they are independent and equally distributed then

\[ p_{\varepsilon_r,\varepsilon_i}(\xi,\zeta) = p_{\varepsilon_r}(\xi)p_{\varepsilon_i}(\zeta) = \frac{1}{\sigma^2\pi} e^{-\frac{\xi^2+\zeta^2}{\sigma^2}}. \]

The probability that \( |\varepsilon(n)| = \sqrt{\varepsilon_r^2(n) + \varepsilon_i^2(n)} < \chi \) is

\[ P\{\sqrt{\varepsilon_r^2(n) + \varepsilon_i^2(n)} < \chi\} = \int\int_{\xi^2+\zeta^2<\chi^2} p_{\varepsilon_r,\varepsilon_i}(\xi,\zeta)d\xi d\zeta \]

\[ = \frac{1}{\sigma^2\pi} \int\int_{\xi^2+\zeta^2<\chi^2} e^{-\frac{\xi^2+\zeta^2}{\sigma^2}}d\xi d\zeta. \]
With $\xi = \rho \cos \alpha$ and $\zeta = \rho \cos \alpha$ (the Jacobian of the polar coordinate transformation is $J = |\rho|$) we get

$$
P\{\sqrt{\varepsilon_1^2(n) + \varepsilon_1^2(n)} < \chi\} = \frac{1}{\sigma^2 \pi} \int_0^{2\pi} e^{-\rho^2/\sigma^2} \rho d\rho d\alpha
$$

$$
= \frac{2}{\sigma^2} \int_0^\chi e^{-\rho^2/\sigma^2} \rho d\rho d\alpha = \int_0^{\chi^2/\sigma^2} e^{-\lambda} d\lambda = (1 - e^{-\chi^2/\sigma^2}) u(\chi) = F_{|\varepsilon(n)|}(\chi).
$$

The probability density function is

$$
p_{|\varepsilon(n)|}(\xi) = \frac{dF_{|\varepsilon(n)|}(\xi)}{d\xi} = \frac{2\xi}{\sigma^2} e^{-\xi^2/\sigma^2} u(\xi). \quad (7.27)
$$

\[ \blacksquare \]

**Example 7.17.** A random signal is defined as $y(n) = |\varepsilon(n)|$, where $\varepsilon(n)$ is the Gaussian complex zero-mean i.i.d. noise with variance $\sigma^2$. What is the probability that $y(n) \geq A$? Calculate this probability for $A = 2$ and $\sigma^2 = 1$.

★The probability density function for sequence $y(n)$ is

$$
p_y(x) = \frac{2\xi}{\sigma^2} e^{-\xi^2/\sigma^2} u(\xi)
$$

The probability that $y(n) \geq A$ is

$$
P\{\xi > A\} = 1 - P\{\xi \leq A\} = e^{-A^2/\sigma^2}.
$$

For $A = 2$ and $\sigma^2 = 1$ we get $P\{\xi > A\} \approx 0.0183$. \[ \blacksquare \]

### 7.3.5 Impulsive Noises

This noise is used to model disturbances when strong impulses occur more often than in the case of a Gaussian noise. Due to possible stronger pulses, their probability density function decay toward $\pm \infty$ is slower than in the case of Gaussian noise.

The Laplacian noise has the probability density function

$$
p_{\varepsilon(n)}(\xi) = \frac{1}{2\alpha} e^{-|\xi|/\alpha}.
$$
It decays much slower as $|\xi|$ increases than in the Gaussian noise case.
The Laplacian noise can be generated as

$$\varepsilon(n) = \varepsilon_1(n)\varepsilon_2(n) + \varepsilon_3(n)\varepsilon_4(n)$$

where $\varepsilon_i(n)$, $i = 1,2,3,4$ are real-valued Gaussian independent zero-mean noises, Fig.7.11 (for variance see Problem 7.13).

The impulsive noise could be distributed in other ways, like, for example, the Cauchy distributed noise, whose probability density function is

$$p_{\varepsilon(n)}(\xi) = \frac{1}{\pi(1 + \xi^2)}.$$

The Cauchy distributed noise $\varepsilon(n)$ is a random signal that can be obtained as a ratio of two independent Gaussian random signals $\varepsilon_1(n)$ and $\varepsilon_2(n)$, i.e.,
as
\[ \epsilon(n) = \frac{\epsilon_1(n)}{\epsilon_2(n)}. \]

### 7.3.6 Noisy Signals

In the case of noisy signals the noise could added to the signal \( s(n) \). Then we have
\[ x(n) = s(n) + \epsilon(n). \]
This is an additive noise. For a deterministic signal \( s(n) \)
\[
\begin{align*}
E\{x(n)\} &= E\{s(n) + \epsilon(n)\} = s(n) + \mu_\epsilon(n), \\
E\{|x(n) - \mu_\epsilon(n)|^2\} &= \sigma_\epsilon^2(n).
\end{align*}
\]
Noise can also be multiplicative, when
\[ x(n) = (1 + \epsilon(n))s(n). \]
In this case
\[
\begin{align*}
E\{x(n)\} &= E\{s(n) + \epsilon(n)s(n)\} = s(n)(1 + \mu_\epsilon(n)), \\
E\{|x(n) - \mu_\epsilon(n)|^2\} &= |s(n)|^2 \sigma_\epsilon^2(n).
\end{align*}
\]
Both the mean and the variance are signal dependent in the case of multiplicative noise.

### 7.4 DISCRETE FOURIER TRANSFORM OF NOISY SIGNALS

Consider a noisy signal
\[ x(n) = s(n) + \epsilon(n) \tag{7.28} \]
where \( s(n) \) is a deterministic useful signal and \( \epsilon(n) \) is an additive noise. The DFT of this signal is
\[
X(k) = \sum_{n=0}^{N-1} (s(n) + \epsilon(n)) e^{-j2\pi kn/N} = S(k) + \Xi(k). \tag{7.29}
\]
The mean value of \( X(k) \) is
\[
E\{X(k)\} = \sum_{n=0}^{N-1} s(n) e^{-j2\pi kn/N} + \sum_{n=0}^{N-1} E\{\epsilon(n)\} e^{-j2\pi kn/N} = S(k) + \text{DFT}\{\mu_\epsilon(n)\}.
\]
In the case of a zero-mean noise $\epsilon(n)$, when $\mu_\epsilon(n) = 0$, follows

$$
\mu_X(k) = \mathbb{E}\{X(k)\} = S(k). \tag{7.30}
$$

The variance of $X(k)$, for a zero-mean noise, is

$$
\sigma_X^2(k) = \mathbb{E}\{(X(k) - \mu_X(k))^2\} = \mathbb{E}\{X(k)X^*(k) - S(k)S^*(k)\} = \sum_{n_1=0}^{N-1} \sum_{n_2=0}^{N-1} \mathbb{E}\{(s(n_1) + \epsilon(n_1))(s^*(n_2) + \epsilon^*(n_2))\} e^{-j2\pi k(n_1 - n_2)/N} - \sum_{n_1=0}^{N-1} \sum_{n_2=0}^{N-1} s(n_1)s^*(n_2)e^{-j2\pi k(n_1 - n_2)/N}
$$

$$
= \sum_{n_1=0}^{N-1} \sum_{n_2=0}^{N-1} \mathbb{E}\{\epsilon(n_1)\epsilon^*(n_2)\} e^{-j2\pi k(n_1 - n_2)/N}. \tag{7.31}
$$

For a white noise, with the autocorrelation

$$
r_{\epsilon\epsilon}(n_1,n_2) = \mathbb{E}\{\epsilon(n_1)\epsilon^*(n_2)\} = \sigma_\epsilon^2 \delta(n_1 - n_2),
$$

we get

$$
\sigma_X^2(k) = \sigma_\epsilon^2 N. \tag{7.32}
$$

If the deterministic signal is a complex sinusoid,

$$
s(n) = Ae^{j2\pi k_0 n/N}, \tag{7.33}
$$

with a frequency adjusted to the grid $\omega_0 = 2\pi k_0 / N$, then its DFT is

$$
S(k) = AN\delta(k - k_0).
$$

Peak signal-to-noise ratio, being relevant parameter for the DFT based estimation of frequency, is

$$
PSNR_{out} = \frac{\max_k |S(k)|^2}{\sigma_X^2} = \frac{A^2 N^2}{\sigma_\epsilon^2 N} = \frac{A^2}{\sigma_\epsilon^2} N. \tag{7.34}
$$

It increases as $N$ increases. We have expected this result since the signal values are added in phase, increasing the DFT amplitude $N$ times (its power $N^2$ times), while the noise values are added in power. Noise influence to the DFT of a real-valued sinusoid $s(n) = A\cos(2\pi k_0 n/N) = (Ae^{j2\pi k_0 n/N} + Ae^{-j2\pi k_0 n/N})/2$ is illustrated in Fig. 7.12.
The input signal-to-noise ratio (SNR) for signal 7.33 is

$$SNR_{in} = \frac{E_x}{E_\epsilon} = \frac{\sum_{n=0}^{N-1} |x(n)|^2}{\sum_{n=0}^{N-1} E \{|\epsilon(n)|^2\}} = \frac{N A^2}{N \sigma^2_\epsilon} = \frac{A^2}{\sigma^2_\epsilon}. \quad (7.35)$$

If the maximal DFT value is detected then only its value could be used for the signal reconstruction (equivalent to the notch filter at $k = k_0$ being used). The DFT of output signal is then

$$Y(k) = X(k)\delta(k - k_0).$$

The output signal in the discrete-time domain is

$$y(n) = \frac{1}{N} \sum_{n=0}^{N-1} Y(k) e^{j2\pi kn/N} = \frac{1}{N} X(k_0) e^{j2\pi k_0 n/N}.$$

Since $X(k_0) = AN + \Xi(k_0)$, according to (7.29) and (7.32), where $\Xi(k)$ is a noise with variance $\sigma^2_\epsilon N$, we get

$$y(n) = Ae^{j2\pi k_0 n/N} + \frac{\Xi(k_0)}{N} e^{j2\pi k_0 n/N} = x(n) + \epsilon X(n).$$
The output signal-to-noise ratio is

\[
\text{SNR}_{\text{out}} = \frac{E_x}{E_{\varepsilon}} = \frac{\sum_{n=0}^{N-1} |x(n)|^2}{\sum_{n=0}^{N-1} \mathbb{E}\left\{ \left| \frac{\Xi(k_0)}{N} e^{j2\pi k_0 n/N} \right|^2 \right\}}
\]

\[
= \frac{NA^2}{N N^2 \sigma_\varepsilon^2} = \frac{A^2}{\sigma_\varepsilon^2} = N \cdot \text{SNR}_{\text{in}}.
\]

Taking \(10\log(\circ)\) of both sides we get the signal-to-noise ratio relation in dB,

\[
\text{SNR}_{\text{out}}[\text{dB}] = 10\log N + \text{SNR}_{\text{in}}[\text{dB}].
\] (7.36)

**Example 7.18.** If the DFT of a noisy signal \(s(n) + \varepsilon(n)\) is calculated using a window function \(w(n)\), find its mean and variance. Noise is white, \(\varepsilon = \sigma_\varepsilon^2 \delta(n)\), with zero-mean.

Here,

\[
X(k) = \sum_{n=0}^{N-1} w(n) [s(n) + \varepsilon(n)] e^{-j2\pi kn/N}.
\]

For this DFT, the mean value is

\[
\mu_X(k) = \mathbb{E}\{X(k)\} = \sum_{n=0}^{N-1} w(n) s(n) e^{-j2\pi kn/N} = W(k) * k S(k)
\]

where \(W(k) = \text{DFT}\{w(n)\}\).

The variance of \(X(k)\) is

\[
\sigma_X^2 = \sum_{n_1=0}^{N-1} \sum_{n_2=0}^{N-1} w(n_1) w^*(n_2) \sigma_\varepsilon^2 \delta(n_1 - n_2) e^{-j2\pi k(n_1-n_2)/N}
\]

\[
= \sigma_\varepsilon^2 \sum_{n=0}^{N-1} |w(n)|^2 = \sigma_\varepsilon^2 E_w,
\] (7.37)

where \(E_w\) is the window energy.

**Example 7.19.** The DFT definition, for a given frequency index \(k\), can be understood as

\[
X(k) = \sum_{n=0}^{N-1} (s(n) + \varepsilon(n)) e^{-j2\pi kn/N}
\]

\[
= N \ \text{mean}_{n=0,1,...,N-1} \ \{(s(n) + \varepsilon(n)) e^{-j2\pi kn/N}\}
\] (7.38)
Based on the definition of median, discuss when the DFT estimation

\[ X_R(k) = N \text{ median}_{n=0,1,...,N-1} \Re \left\{ (s(n) + \varepsilon(n))e^{-j2\pi kn/N} \right\} + jN \text{ median}_{n=0,1,...,N-1} \Im \left\{ (s(n) + \varepsilon(n))e^{-j2\pi kn/N} \right\} \]  

(7.39)

can produce better results than (7.38). Calculate the value \( X(0) \) using (7.38) and estimate it by (7.39) for \( s(n) = \exp(j4\pi n/N) \) with \( N = 8 \) and noise \( \varepsilon(n) = 2001\delta(n) - 204\delta(n-3) \). Which one is closer to the noise-free DFT value?

\* If we can expect strong impulsive noise then the mean value will be highly sensitive to this noise. The median based calculation is less sensitive to strong impulsive noise. For the given signal

\[ s(n) = \exp(j\pi n/2) = [1, j, -1, -j, 1, j, -1, -j] \]

and noise \( \varepsilon(n) \) the value of \( X(0) \) is

\[ X(0) = 0 + 2001 - 204 = 805. \]

The median-based estimation is

\[ X_R(0) = 8 \text{ median}_{n=0,1,...,N-1} \{2002, 0, -1, -204, 1, 0, -1, 0\} + j8 \text{ median}_{n=0,1,...,N-1} \{0, 1, 0, -1, 0, 1, 0, -1\} = 0 + j0. \]

Obviously the median-based estimate is not influenced by this impulsive noise. In this case it produced better estimate of the noise-free DFT.

\[ \Box \]

### 7.4.1 Detection of a Sinusoidal Signal Frequency

Consider a set of data \( x(n) \), for \( 0 \leq n \leq N - 1 \). Assume that this set of data are noisy samples of signal \( s(n) = Ae^{j2\pi kn/n} \). Additive noise \( \varepsilon(n) \) is white, complex-valued Gaussian with zero-mean independent real and imaginary parts and variance \( \sigma^2_\varepsilon \). The aim is to find the signal \( s(n) \) parameters from the noisy observations \( x(n) \). Since the signal form is known we look for a solution of the same form, using the model \( be^{j2\pi kn/N} \) where \( b \) and \( k \) are parameters that have to determined, and \( \alpha = \{b, k\} \) is the set of parameters. Parameter \( b \) is complex-valued. It includes amplitude and initial phase of the model. For each value of \( x(n) \) we may define an error as a difference of the given value \( x(n) \) and the assumed model, at the considered instant \( n \),

\[ e(n, \alpha) = x(n) - be^{j2\pi kn/N}. \]  

(7.41)
Since the noise is Gaussian, the probability density function of the error is
\[
p(e(n, \alpha)) = \frac{1}{2\pi\sigma^2} e^{-|e(n, \alpha)|^2/(2\sigma^2)}.
\]

The joint probability density function for all samples from the data set is equal to the product of individual probability density functions
\[
p_e(e(0, \alpha), e(1, \alpha), \ldots, e(N-1, \alpha)) = \frac{1}{(2\pi\sigma^2)^N} e^{-\sum_{n=0}^{N-1}|e(n, \alpha)|^2/(2\sigma^2)}.
\]

The maximum-likelihood solution for parameters \( \alpha = \{b, k\} \) in obtained by maximizing this function for given values of \( x(n) \). Maximization of \( p_e(e(0, \alpha), e(1, \alpha), \ldots, e(N-1, \alpha)) \) is the same as the minimization of the total square error,
\[
\epsilon(\alpha) = \sum_{n=0}^{N-1} |e(n, \alpha)|^2 = \sum_{n=0}^{N-1} \left| x(n) - be^{j2\pi kn/N} \right|^2. \tag{7.42}
\]

The solution of this problem is obtained from \( \partial \epsilon(\alpha) / \partial b = 0 \) (see Example 1.3). It is in the form of a standard DFT of signal \( x(n) \),
\[
b = \frac{1}{N} \sum_{n=0}^{N-1} x(n)e^{-j2\pi kn/N} = \text{mean} \left\{ x(n)e^{-j2\pi kn/N} \right\} = \frac{1}{N} X(k).
\]

A specific value of parameter \( k \) that minimizes \( \epsilon(\alpha) \) and gives the estimate of the signal frequency index \( k_0 \) is obtained by replacing the obtained \( b \) back into relation (7.42) defining \( \epsilon(\alpha) \),
\[
\epsilon(\alpha) = \sum_{n=0}^{N-1} \left| x(n) - be^{j2\pi kn/N} \right|^2 = \left( \sum_{n=0}^{N-1} |x(n)|^2 \right) - N|b|^2.
\]

Minimal value of \( \epsilon(\alpha) \) is achieved when \( |b|^2 \) (or \( |X(k)|^2 \)) is maximal,
\[
\hat{k}_0 = \text{arg} \{ \max |X(k)|^2 \} = \text{arg} \{ \max |X(k)| \}.
\]

If there is no noise \( |x(n)| = A, \hat{k}_0 = k_0, b = A \) or \( X(k_0) = NA \), and \( \epsilon(k_0) = 0 \). The same approach can be used for a signal \( s(n) = A e^{j\omega_0 n} \). Assuming the solution in the form \( be^{j\omega n} \), the Fourier transform of discrete-time signals would follow.
If the additive noise were, for example, Laplacian then the probability density function would be

\[ p(e(n, \alpha)) = \frac{1}{2\sigma_e} e^{-|e(n, \alpha)|/\sigma_e} \]

and the solution of

\[ e(\alpha) = \sum_{n=0}^{N-1} |e(n, \alpha)| \]

minimization would follow from

\[ X(k) = N\text{median}\left\{ x(n)e^{-j2\pi kn/N} \right\}. \]

Note that the absolute value of error can be written as

\[ |e(n, \alpha)| = \left| x(n) - be^{j2\pi kn/N} \right| = \left| x(n)e^{-j2\pi kn/N} - b \right|. \]

Minimization of a sum of this kind of terms is discussed in (10.72).

Now we will analyze the signal frequency estimation for a single component sinusoidal signal \( s(n) \) with unknown discrete frequency \( \omega_0 = 2\pi k_0/N \) using the DFT. Since a frequency on the frequency grid is assumed this case can be understood as a frequency position detection. Available observations of the signal are

\[ x(n) = s(n) + \epsilon(n), \text{ for } 0 \leq n \leq N - 1, \]

where \( \epsilon(n) \) is a complex zero mean Gaussian white noise with independent real and imaginary parts, with variance \( \sigma^2_{\epsilon} \). Its DFT is

\[ X(k) = \sum_{n=0}^{N-1} (s(n) + \epsilon(n))e^{-j2\pi kn/N} = N\epsilon(k - k_0) + \Xi(k), \]

with \( \sigma^2_X(k) = \sigma^2_{\epsilon} N \) and \( \mathbb{E}\{\Xi(k)\} = 0 \). The real and imaginary parts of the DFT \( X(k_0) \) at the signal position \( k = k_0 \) are Gaussian random variables, with total variance \( \sigma^2_{\epsilon} N \), or

\[ \mathcal{N}(NA, \sigma^2_{\epsilon} N/2), \quad \mathcal{N}(0, \sigma^2_{\epsilon} N/2), \quad (7.43) \]

respectively, where a real-valued \( A \) is assumed without any loss of generality.

Real and imaginary parts of the noise only DFT values \( X(k) \) for \( k \neq k_0 \) are zero-mean random variables with the same variance

\[ \mathcal{N}(0, \sigma^2_{\epsilon} N/2). \]

Next, we will find the probability that a DFT value of noise at any \( k \neq k_0 \) is higher than the signal DFT value at \( k = k_0 \). This case corresponds
to a false detection of the signal frequency position, resulting in an arbitrary large and uniform estimation error (within the considered frequency range).

The probability density function for the absolute DFT values outside the signal frequency is Rayleigh-distributed (7.27)

$$q(\xi) = \frac{2\xi}{\sigma^2 N} e^{-\xi^2/(\sigma^2 N)}, \xi \geq 0.$$  

The DFT at a noise only position takes a value greater than $\Xi$, with probability

$$Q(\Xi) = \int_{\Xi}^{\infty} \frac{2\xi}{\sigma^2 N} e^{-\xi^2/(\sigma^2 N)} d\xi = \exp\left(-\frac{\Xi^2}{\sigma^2 N}\right).$$  

The probability that a DFT of noise only is lower than $\Xi$ is

$$1 - Q(\Xi).$$

The total number of noise only points in the DFT is $M = N - 1$. The probability that $M$ independent DFT noise only values are lower than $\Xi$ is $[1 - Q(\Xi)]^M$. Probability that at least one of $M$ DFT noise only values is greater than $\Xi$, is

$$G(\Xi) = 1 - [1 - Q(\Xi)]^M.$$  

The probability density function for the absolute DFT values at the position of the signal (whose real and imaginary parts are described by (7.43)) is Rice-distributed

$$p(\xi) = \frac{2\xi}{\sigma^2 N} e^{-\xi^2/(\sigma^2 N)} I_0\left(2NA\xi/(\sigma^2 N)\right), \xi \geq 0,$$  

where $I_0(\xi)$ is the zero-order modified Bessel function (for $A = 0$, when $I_0(0) = 1$ the Rayleigh distribution is obtained).

When a noise only DFT value surpasses the DFT signal value, then an error in estimation occurs. To calculate this probability, consider the absolute DFT value of a signal at and around $\xi$. The DFT value at the signal position is within $\xi$ and $\xi + d\xi$ with the probability $p(\xi)d\xi$, where $p(\xi)$ is defined by (7.46). The probability that at least one of $M$ DFT noise only values is above $\xi$ in amplitude is $G(\xi) = 1 - [1 - Q(\xi)]^M$. Thus, the probability that the absolute DFT signal component value is within $\xi$ and $\xi + d\xi$ and that at least one of the absolute DFT noise only values exceeds the DFT signal value is $G(\xi)p(\xi)d\xi$. Considering all possible values of $\xi$, from (7.44) and (7.45), it
follows that the probability of the wrong signal frequency detection is

\[
P_E = \int_0^\infty G(\xi)p(\xi)d\xi = \int_0^\infty \left(1 - \exp\left(-\frac{\xi^2}{\sigma^2 N}\right)\right)^M d\xi = \int_0^\infty \frac{2\xi}{\sigma^2 N} e^{-\left(\xi^2 + N^2 A^2\right)/\left(\sigma^2 N\right)} I_0\left(2NA \xi / \left(\sigma^2 N\right)\right) d\xi.
\]

Approximation of this expression can be calculated by assuming that the DFT of the signal component is not random and that it is equal to \(NA\) (positioned at the mean value of the signals DFT). The form of error probability is then very simple

\[
P_E \approx 1 - \left[1 - \exp\left(-\frac{NA^2}{\sigma^2 N}\right)\right]^M.
\]

This expression can be used for a simple rough approximative analysis.

Analysis can easily be generalized to the case with \(K\) signal components,

\[
s(n) = \sum_{k=1}^K A_k e^{j\omega_k n}.
\]

In many cases, the discrete frequency of the deterministic signal does not satisfy the relation \(\omega_0 = 2\pi k_0 / N\), where \(k_0\) is an integer. In these cases, when \(\omega_0 \neq 2\pi k_0 / N\), the frequency estimation result can be improved, for example, by zero-padding before the Fourier transform calculation or using finer grid around the detected maximum. Comments on the estimation of signal frequency outside the grid are given in Chapter III as well.

### 7.5 LINEAR SYSTEMS AND RANDOM SIGNALS

If a random signal \(x(n)\) passes through a linear time-invariant system, with an impulse response \(h(n)\), then the mean value of the output signal \(y(n)\) is

\[
\mu_y(n) = \mathbb{E}\{y(n)\} = \sum_{k=-\infty}^{\infty} h(k) \mathbb{E}\{x(n-k)\} \quad (7.49)
\]

\[
= \sum_{k=-\infty}^{\infty} h(k) \mu_x(n-k) = h(n) *_n \mu_x(n). \quad (7.50)
\]

For a stationary signal

\[
\mu_y = \mu_x \sum_{k=-\infty}^{\infty} h(k) = \mu_x H(e^{j0}). \quad (7.51)
\]
The cross-correlation of the output and input signal is
\[ r_{yx}(n,m) = E\{y(n)x^*(m)\} = \sum_{k=-\infty}^{\infty} E\{x(k)x^*(m)\}h(n-k) \]
\[ = \sum_{k=-\infty}^{\infty} r_{xx}(k,m)h(n-k). \] (7.52)

For a stationary signal, with \( n - m = l \) and \( k - m = p \), we get
\[ r_{yx}(l) = \sum_{p=-\infty}^{\infty} r_{xx}(p)h(l-p) = r_{xx}(l) \ast_l h(l). \]

The \( z \)-transform of both sides gives
\[ R_{yx}(z) = R_{xx}(z)H(z). \]

The cross-correlation of the input and output signal is
\[ r_{xy}(n,m) = E\{x(n)y^*(m)\} = \sum_{k=-\infty}^{\infty} E\{x(n)x^*(k)\}h^*(m-k) \]
\[ = \sum_{k=-\infty}^{\infty} r_{xx}(n,k)h^*(m-k). \] (7.53)

For a stationary signal, with \( n - m = l \) and \( n - k = p \), we get
\[ r_{xy}(l) = \sum_{p=-\infty}^{\infty} r_{xx}(p)h^*(p-l). \]

The \( z \)-transform of both sides are
\[ \sum_{l=-\infty}^{\infty} r_{xy}(l)z^{-l} = \sum_{l=-\infty}^{\infty} \sum_{p=-\infty}^{\infty} r_{xx}(p)h^*(p-l)z^{-l} \]
\[ = \sum_{k=-\infty}^{\infty} \sum_{p=-\infty}^{\infty} r_{xx}(p)h^*(k)z^{-p} \left(\frac{1}{z^{-1}}\right)^{-k} \]
\[ R_{xy}(z) = R_{xx}(z)H^*\left(\frac{1}{z^*}\right). \]

If we calculate the Fourier transform of both sides, we get
\[ S_{xy}(e^{j\omega}) = S_{xx}(e^{j\omega})H^*(e^{j\omega}). \] (7.54)
In a vector notation

\[ G(n) = U(n) + \sum_{k=1}^{L} b_k(n) G(n-k). \]

Generalization of this algorithm is obtained if different steps \( \mu \) are used for different coordinates of the error vector. Then instead of step \( \mu \) a diagonal matrix of steps is used

\[ M = \text{diag}(\mu_1, \mu_2, \ldots, \mu_{N+L}). \]

Special attention in the adaptive recursive systems has to be paid to the system stability. It requires additional constraints on coefficients \( b_k(n) \).

**Example 8.26.** Consider identification of system from Example 8.13 (page 456). Here this system will be identified by using an adaptive recursive system with \( N = 2 \) and \( L = 2 \). Step \( \mu \) for the coefficients in the numerator of the transfer function is 0.025 while its value is 0.005 for the coefficients in denominator. Other parameters are the same as in Example 8.13.

★ The simulation results are presented in Figure 8.31.

8.9 **FROM THE LMS ALGORITHM TO THE KALMAN FILTERS**

Consider the problem of unknown system identification, where the unknown system has the same input signal \( x(n) \) as the adaptive system, and the output of the unknown system is used as a reference signal \( d(n) = X^T(n)H^* + \nu(n) \) in the adaptive system. Here \( \nu(n) \) denotes zero-mean Gaussian measurement noise, with variance \( \sigma^2 \), and it is assumed that it is not correlated with other variables. It has been shown that the cost function

\[ J_{\text{MSE}}(n) = E[e^2(n)] \]

can be used to define the deviation of the error signal from the ideal case. Minimization of this function provides the optimal system parameters in form of the Wiener optimal filter. It is also proved that LMS algorithm converges "in mean" toward the optimal system coefficient values \( H^* \). In the unknown system identification framework, the optimal coefficient values are equal to the coefficients of the unknown system. In each

---

time instant the adaptive system coefficients are changed following the rule

\[ H(n+1) = H(n) + \mu(n)e(n)X(n). \]  

(8.22)

by changing previous values of system coefficients in the direction of input signal vector \( X(n) \). Since the LMS algorithm employs stochastic gradient descent to minimize the cost function \( J_{MSE} \), it performs locally optimal steps, but not the globally optimal shortest path to the solution, which especially slows the convergence of algorithm in the case of correlated data. The step \( \mu(n) \) which can be in general time-dependent (class of variable step LMS algorithms) controls the magnitude of adaptation steps, but not the direction. In order to be able to follow the shortest path to the optimal solution, i.e. to control the direction among the amplitude of adaptation steps, the scalar step size \( \mu(n) \) can be replaced by a positive definite matrix \( G(n) \), which introduces more degrees of freedom in adaptation steps. This
is the first step towards Kalman filters, and (8.22) now becomes
\[ H(n + 1) = H(n) + G(n)e(n)X(n) = H(n) + g(n)e(n). \] (8.23)

Previous recursion is also known as generalized LMS algorithm. Since the unknown system identification framework is considered, instead of error \( e(n) = d(n) - y(n) = d(n) - X^T(n)H(n) \) the weight error vector defined as the deviation of adaptive coefficients \( \hat{H}(n) \) from the optimal coefficients \( H^* \)
\[ \hat{H}(n) = H^* - H(n) \] (8.24)
can be introduced, and based on it, we can define a measure of how closely adaptive system coefficients \( H(n) \) approach the optimal solution \( H^* \). This measure is the mean square deviation (MSD) and it is given with
\[ J_{MSD}(n) = E \left\{ \left\| \hat{H}(n) \right\|^2 \right\} \] (8.25)

Note that it is assumed that the unknown system is deterministic and non-stationary. Since the weight error vector can be related with the system output error \( e(n) \) with:
\[ e(n) = X^T(n)H^* + v(n) - X^T(n)H(n) = X^T(n)\hat{H}(n) + v(n), \] (8.26)
a relation between \( J_{MSE} \) and \( J_{MSD} \) can be found indicating that the minimization of MSD also corresponds to the minimization of MSE. For the simplicity of derivation we will assume that \( X(n) \) is deterministic which is a common assumption in Kalman filtering literature, although it is usually treated as a zero-mean process with autocorrelation matrix \( R \) in the context of adaptive systems. If we introduce the weight error covariance matrix \( P(n) = E \{ \hat{H}(n)\hat{H}^T(n) \} \), in order to perform the minimization of \( J_{MSD} \), starting from (8.23) a recursive relation for the matrix \( P(n) \) is established:
\[
\begin{align*}
H^* - H(n + 1) &= H^* - H(n) - G(n)X(n)(X^T(n)\hat{H}(n) + v(n)) \\
\hat{H}(n + 1) &= \hat{H}(n) - g(n)X^T(n)\hat{H}(n) - g(n)v(n)
\end{align*}
\]
\[
\begin{align*}
\hat{H}(n + 1)\hat{H}^T(n + 1) &= \left( \hat{H}(n) - g(n)X^T(n)\hat{H}(n) - g(n)v(n) \right) \\
&\quad \left( \hat{H}(n) - g(n)X^T(n)\hat{H}(n) - g(n)v(n) \right)^T \\
P(n + 1) &= P(n) - \left( P(n)X(n)g^T(n) + g(n)X^T(n)P(n) \right) \\
&\quad + g(n)g^T(n) \left( X^T(n)P(n)X(n) + v(n) \right).
\end{align*}
\]
By taking expectations of both sides, previously multiplied with their transposes, and having in mind that 

\[ \text{tr} \{ P(n)X(n)g^T(n) \} = g^T(n)P(n)X(n) \]

this leads us to the MSD recursion of the form

\[ J_{\text{MSD}}(n+1) = J_{\text{MSD}}(n) - 2g^T(n)P(n)X(n) + \| g(n) \|^2 (X^T(n)P(n)X(n) + \sigma_v^2). \]

The optimal learning gain vector \( g(n) \) which provides the control over both direction and amplitude of adaptation steps in (8.23) is obtained by solving \( \partial J_{\text{MSD}}(n+1)/\partial g(n) = 0 \) as

\[ g(n) = G(n)e(n) = \frac{P(n)X(n)}{X^T(n)P(n)X(n) + \sigma_v^2}, \tag{8.27} \]

which is known as the Kalman gain. Besides the calculation of (8.27), the Kalman filter which estimates the optimal time-invariant and deterministic coefficients for each time instant also includes the coefficients adjustment

\[ H(n+1) = H(n) + g(n)(d(n) - X^T(n)H(n)), \]

as well as the weight error covariance matrix update

\[ P(n+1) = P(n) - g(n)X^T(n)P(n). \tag{8.28} \]

Note that previous algorithm steps for \( \sigma_v^2 = 1 \) can be related with the RLS algorithm equations.

A generalization of the previous approach assumes time-varying and stochastic weight vector \( H^*(n) \)

\[ H^*(n+1) = F(n)H^*(n) + q(n), \tag{8.29} \]

\[ d(n) = X^T(n)H^*(n) + v(n) \tag{8.30} \]

with \( q(n) \) being a zero-mean Gaussian stochastic process with covariance matrix \( Q = E\{ q(n)q^T(n) \} \), and \( F(n) \) being a known matrix which describes the system changes over time (state-transition matrix). It is assumed that the measurement noise \( v(n) \) is also uncorrelated with \( q(n) \). In the framework of the general Kalman filter, the coefficient vector is updated using the current state estimate denoted with \( H(n|n) \), while the prediction of its next state is denoted with \( H(n+1|n) \). The prediction step is needed for tracking the time-varying error surface. The coefficients are updated by

\[ H(n|n) = H(n|n-1) + g(n)(d(n) - X^T(n)H(n|n)) \tag{8.31} \]
while the coefficients prediction is obtained with

\[ H(n+1|n) = F(n)H(n|n). \]  

(8.32)

Note that the same definition of the weight error vector \( \hat{H}(n|n) = H^* - H(n|n) \) holds, as well as for weight error covariance matrix

\[ P(n|n) = E \left\{ \hat{H}(n|n)\hat{H}^T(n|n) \right\}. \]

The weight error covariance matrix is updated in the same manner as for the time-invariant deterministic case

\[ P(n|n) = P(n|n-1) - g(n)X^T(n)P(n|n-1), \]  

(8.33)

with the respect to the new index notation. The general Kalman filter also includes the prediction step of weight error matrix which easily follows from its definition

\[ P(n+1|n) = E \left\{ \hat{H}(n+1|n)\hat{H}^T(n+1|n) \right\} = F(n)P(n|n)F^T(n) + Q. \]  

(8.34)

Similarly to the time-invariant deterministic case, the Kalman gain which minimizes MSD is obtained in the following form

\[ g(n) = G(n)e(n) = \frac{P(n|n-1)X(n)}{X^T(n)P(n|n-1)X(n) + \sigma_w^2}. \]  

(8.35)

**Example 8.27.** Consider the problem of identification of unknown time-invariant deterministic system with two coefficients \( h_0 = 3 \) and \( h_1 = 4 \) using the standard LMS algorithm and Kalman filter (for stationary system identification), with \( N = 2 \). The input signal is colored noise \( x(n) = 5w(n) + 3.4w(n-1) + w(n-2) \), where \( w(n) \) is a zero mean white noise with variance \( \sigma_w^2 = 1 \). The step \( \mu = 0.0005 \) is used for the LMS algorithm. Show the convergence paths on the MSE contour plot. After how many iterations the Kalman filter approaches the optimal solution?

⭐The convergence paths on the MSE contour are shown in Fig. 8.32. Numbers on the Kalman filter path indicate that the optimal solution is obtained after only two iterations.

\[ \square \]
8.10 NEURAL NETWORKS

Artificial neural networks, or just neural networks, represent a simplified mathematical model of biological systems. In such systems a distributed parallel data processing is performed, in contrast to the common engineering systems that are designed for a sequential data processing. Common computer systems are based on well defined algorithms that are executed in a sequential order, while in the neural networks a learning period is required to achieve their satisfactory response to the input data. Correspondence with biological systems, that also require learning, is evident. Commonly, a trained network continues to learn and adapt to the new situations during the exploitation. It means that the process of learning does not end.
Neural network can be defined as an artificial cell system capable of accepting, memorizing and applying empirical knowledge. The knowledge here means that the neural network can respond to an input from the environment in an appropriate way. Neural network is connected to the environment in two ways: through the inputs where the environment influences the network and through the outputs where the network responses to environment, as it is illustrated in Figure 8.33.

The basic element in a neural network is neuron. It is the elementary unit for a distributed signal processing in a neural network. A full functionality of neural networks is achieved using large number of interconnected neurons. Connections among neurons are one-directional (the outputs from one neuron can be used as inputs to the other neuron). They are called synapses, in analogy with the biological systems.

Possible applications of neural networks include almost all aspects of modern life, text and speech recognition, optimization of a communication channel, financial forecasts, detection of a fraud credit card usage, are just a few examples.

Of course, there are many situations when a usage of neural networks is not justified. In many cases our knowledge about the system, that we want to control or observe, is sufficient and complete so the problem can be solved using classical algorithms, with sequential processing on common computers.

An ideal system for neural networks realization would use independent systems for hardware realization of each neuron. Then the distributed processing would be most efficient. In the cases of monoprocessor computers, high efficiency is achieved by using very fast sequential data processing. Typical examples are computer programs for recognition of a scanned text.
8.10.1 Neuron

The first step in a neuron design is to define its inputs and outputs. In biological systems the input and output signals to a neuron are electric potential that can be modelled by real numbers. The same principle is used in artificial neurons. Illustration of a neuron is given in Figure 8.34(a) for the case when it has \( N \) inputs \((x_1(n), x_2(n), \ldots, x_N(n))\) and one output \( y(n) \). Index \( n \) may be a time index, but it can also be understood as a cardinal number that identifies the input and output index of a neuron.

Neuron represents and algorithm that transforms \( N \) input data into one output signal. It is common to split this algorithm into two parts: 1) combinatorial process that transforms \( N \) input data to one output value \( u(n) \) and 2) the process that produces output signals \( y(n) \) based on the value of \( u(n) \). This two-phase model of a neuron is presented in Figure 8.34(b). The algorithm/rule to produce \( u(n) \) is called the network function, while the second part which determines the output value is the activation function.

Neuron knowledge is accumulated and contained in the way how the input data are combined, i.e., in the network function.

8.10.2 Network Function

The basic task of the network function is to combine the input data. The simplest way of combining \( N \) input signals is in their linear weighed combination with coefficients \( w_i, i = 1, 2, \ldots, N \). This is a linear network function. Because of it simplicity, this type of function is commonly used in neurons.
Examples of network function are given in the table.

<table>
<thead>
<tr>
<th>Name</th>
<th>Network function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear form</td>
<td>( u(n) = \sum_{i=1}^{N} w_i x_i(n) + \theta )</td>
</tr>
<tr>
<td>Linear form (II order)</td>
<td>( u(n) = \sum_{i=1}^{N} \sum_{k=1}^{N} w_{ik} x_i(n) x_k(n) + \theta )</td>
</tr>
<tr>
<td>Product form</td>
<td>( u(n) = \prod_{i=1}^{N} x_i^{w_i}(n) )</td>
</tr>
</tbody>
</table>

The values of network function commonly depends not only on the input data, but also on the current state of the neuron. This state is modeled by a real variable \( \theta \), called bias or threshold. Neuron model with a nonzero bias and a linear activation function may be simplified if we introduce an additional input \( x_0 \) in addition to the existing \( N \) inputs. It will be assumed that this additional input is always set to \( x_0 = 1 \). Then the bias can be modelled with a coefficient \( w_0 \) and the neuron considered as a zero-bias neuron. This kind of simplification will be used in the presentation that follows.

The knowledge (as a way to transform input data to output signal) is contained in the values of the coefficients \( w_i \) (or \( w_{ik} \)) of the network function.

### 8.10.3 Activation Function

The activation function transform the output value from the network function to an acceptable output value. A common requirement is that the output values have limited range. Thus, most of the activation functions have a bounded interval of real numbers as its codomain, like for example, \([0,1]\) or \([-1,1]\) or a set of binary digits. Forms of commonly used activation functions are presented in table. The most important functions from this set are the unipolar threshold function and the unipolar sigmoid. Some of the activation functions are presented in Figure 8.35 as well.
### Function   |  Formula
---|---
Linear   |  \( f(u) = u \)
Linear with a limiter   |  \( f(u) = \begin{cases} 1 & \text{za } u > 1 \\ za & \text{za } -1 \leq u \leq 1 \\ -1 & \text{za } u < -1 \end{cases} \)
Threshold function (unipolar)   |  \( f(u) = \begin{cases} 1 & \text{za } u > 0 \\ 0 & \text{za } u < 0 \end{cases} \)
Threshold function (bipolar)   |  \( f(u) = \begin{cases} 1 & \text{za } u > 0 \\ -1 & \text{za } u < 0 \end{cases} \)
Sigmoid (unipolar)   |  \( f(u) = \frac{1}{1 + \exp(-u)} \)
Sigmoid (bipolar)   |  \( f(u) = \frac{2}{1 + \exp(-2u)} - 1 \)
Inverse tangent function   |  \( f(u) = \frac{2}{\pi} \arctan(u) \)
Gauss function   |  \( f(u) = \exp\left(\frac{(u-m)^2}{\sigma^2}\right) \)

In literature hyperbole tangent function is used as well. It is equal the bipolar sigmoid.
8.10.4 Neural Network Topology

A number of neurons is connected to each other and to input signal within a neural network in order to get a network output. Configuration of the signal flow can be presented by a graph where the neurons are nodes of the graph and the signal flow is represented by the directed edges. A special category are the input edges (its start is not a neuron but a node that represents acquisition of data outside of the neural network - input node) and the output edges (its end is not a neuron but a node that represents output information forwarded to the environment - output node).

From the point of view of topology the neural networks can be classified in various ways. If a neural network contains a closed loop in the graph then it is a cyclic (recurrent) neural network. The network containing no closed loop is an acyclic (feed-forward) neural network. Examples of cyclic and acyclic networks are presented in Figure 8.36. Cyclic or recursive neural networks are dynamic nonlinear systems (with memory) whose design, learning, and implementation is complex, mainly because of the nonlinear nature of individual neuron activation functions. In acyclic networks there are no closed loops so that the input information pass to the network output through a finite number of neurons.

A special category of the acyclic neural networks are layer networks, where the neurons can be divided into disjunctive subsets (layers). The output data from one layer are the input data to other layer. Neurons from the first layer get the information from the network input, while the neurons from the last layer produce output information from the network. The simplest cases are neural networks with one or two layers of neurons. In multilayer configurations it is assumed that the inputs to the \( m \)th layer are the outputs from the \( (m - 1) \)th layer of neurons. This approach simplifies
mathematical analysis of the neural networks. In situation when the state of one neuron is not influenced by all input values to that layer, the inputs without influence are modelled with zero weighting coefficients in the network function of this neuron.

In layer networks it is common to introduce the zeroth (input) layer of neurons where there is no data processing. Its function is to forward the input data to the first layer. The last layer is the output layer. An example of one-layer neural network is presented in Figure 8.37. An example of neural network with four layers is shown in Figure 8.38. This network consists of three hidden layers (I, II, and III) with 3, 2 and 4 neurons, respectively, and the output layer with one neuron.
8.10.5 **Network with Supervised Learning**

The network adapts for desired functioning through a process of training. The training is implemented using the following algorithm:

1. Data for the network training are acquired. This data consists of the input-output pairs. The output data are assumed, estimated or obtained through experiments. This set of training data pairs is finite. Denote the number of available input-output pairs by $K$.

2. The network is initiated, commonly by using random parameters of neurons (if an *a priori* information about the range of their values does not exist). After the initialization, the iterative training procedure is implemented as follows:

   (a) One input-output pair of data is considered. The output of the neural network is calculated. The output value is compared with the desired (given) output value. If the output from the neural network is equal to the desired output value (or sufficiently close to this value) then no correction in the network is done. Otherwise when the result of comparison is not satisfactory, the parameters of neural network are corrected to get a better result in the considered case.

   (b) Next pair of the input-output data is considered. The calculation, comparison, and correction process is repeated.

   (c) This cycle of training ends when all data available for training are used. This whole cycle is called one *epoch*.

   (d) The result achieved by the neural network in the previous epoch of training is analyzed. If the output of the neural network has been satisfactory for each individual pair of the data then the training is finished. If the result is not satisfactory, then another epoch of training is needed. Go back to 2a and repeat all previous steps.

3. Neural network is ready for use. A testing of the trained neural network can now be done. A common way of testing is in omitting some pairs of the available input-output data in the training procedure and in using them after the training process is completed, to test the neural network accuracy.
The iterative algorithm for training of neural network (steps 2a–2d) does not necessarily converge. Algorithm is usually implemented by imposing a maximal number of epochs. In the case that the result is not satisfactory (neural network does not produce accurate results after training) the training process may be repeated with new initialization of the network parameters. This is one of the reasons why the random initial values of the parameters are used.

One of the main problems in the neural networks training is the way how to modify the parameters when we are not satisfied with the results. Several networks will be analyzed next, where this problem will be addressed. In some cases just a small random correction of parameters can improve results, when the process of parameter changes is locked in a local minimum.

8.10.6 One-Layer Network with Binary Output - Perceptron

Perceptron as the oldest simple form of neural networks. Perceptron has to establish if the input data have a desired property or not. It produces a binary output. For example, the input data may be scanned pixels of one letter in the text. From neural network we expect a simple answer is the scanned letter, for example letter “A” or not. Since the output is a logical variable, an obvious choice for the neuron activation function is a function with binary output (bipolar or unipolar threshold function). Assume that unipolar function is used as the activation function

\[
f(u) = \begin{cases} 
1 & u > 0 \\
0 & u < 0 
\end{cases}.
\]

The neuron network function, in the case of perceptron, is a linear form

\[
u(n) = \sum_{k=1}^{N} w_k x_k(n) \tag{8.36}
\]

where it has been assumed that the neuron has \(N\) input data. The weighting coefficients \(w_k\) represent “knowledge” that the network should get through the training procedure. This knowledge will be then used in real situations.

The vector notation is

\[
X(n) = \begin{bmatrix} x_1(n) \\ x_2(n) \\ \vdots \\ x_N(n) \end{bmatrix}_{N \times 1} \quad W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix}_{N \times 1}.
\]
The network function can be written as

\[ u(n) = W^T X(n) = X^T(n) W. \]

The neuron output is

\[ y(n) = f(u(n)) = f \left( W^T X(n) \right). \]

Perceptron basic form consists of one neuron only. In topological sense it is acyclic network. It is one layer neural network. In a similar way it is possible to define perceptron with more than one neuron. In that case we have a neural network with \( M \) output signals. Network can decide if the input data contain some out of \( M \) properties that we are interested in. An example of such a network would be a perceptron which, based on the pixels of a scanned letter, decides which of the letters „A”, „B” or „C” was scanned. Note than in this case we may obtain more than one 1 at the output (network, for example has recognized that the input data correspond to the letter „A” and letter „B”). In the considered case it means that the network was not able to recognize the letter. However, in some cases the output data can be defined in such a way that more than one property are satisfied for one set of input data. For example, in letter recognition, the results may be: „scanned letter contains horizontal line”, „scanned letter contains vertical line” and „scanned letter contains oblique” and to train the neural network to recognize these properties of a scanned letter.

The network output can be written as a column vector with \( M \) elements,

\[ Y(n) = \begin{bmatrix} y_1(n) \\ y_2(n) \\ \vdots \\ y_M(n) \end{bmatrix}_{M \times 1}. \]

The weighting coefficients of individual neurons \( W_l, l = 1,2,\ldots,M \) can be written in a matrix form

\[ W = \begin{bmatrix} W_1 & W_2 & \cdots & W_M \end{bmatrix}_{N \times M} = \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{M1} \\ w_{12} & w_{22} & \cdots & w_{M2} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1N} & w_{2N} & \cdots & w_{MN} \end{bmatrix}_{N \times M}. \]
The relation between output and input can be written as

\[ Y(n) = f(W^T X(n)). \]

Consider again the simple case of a perceptron with one neuron. In the training process, when for a given input \( X(n) \) the network produces output \( y(n) \), it has to be compared with desired value \( d(n) \). Possible cases are \( y(n) = d(n) \), when no correction of the network coefficients \( W \) is done, and \( y(n) \neq d(n) \) when the coefficients are modified as

\[ W_{\text{new}} = W_{\text{old}} + \Delta W. \]

The correction \( \Delta W \) should be done in such a way that it increases the possibility to get a desired output for the considered input data. In this case the output may take only one of two binary values: 0 for \( u(n) = W^T X(n) < 0 \) or 1 for \( u(n) = W^T X(n) > 0 \). Assume that the desired value of the output is 1 and that we obtained 0. It means \( W^T X(n) < 0 \) holds, while it should be \( W^T X(n) > 0 \). The value of \( W^T X(n) \) should be increased to increase the possibility of getting the desired output. In the case that the desired output is 0 and the output from the network is 1, using a similar reasoning we can conclude that \( W^T X(n) \) has to be decreased. A way to implement the desired coefficients modification is

\[ W_{\text{new}} = W_{\text{old}} + \mu(d(n) - y(n))X(n). \]

In the case when \( d(n) = 1 \) and \( y(n) = 0 \) it means that \( W_{\text{new}} = W_{\text{old}} + \mu X(n) \) or

\[ W_{\text{new}}^T X(n) = W_{\text{old}}^T X(n) + \mu X^T(n)X(n) \]
\[ = W_{\text{old}}^T X(n) + \mu ||X(n)||^2_2, \]

where \( ||X(n)||^2_2 \) is the squared norm two of vector \( X(n) \) (sum of its squared elements). The value of \( W^T X(n) \) is increased for \( \mu ||X(n)||^2_2 \), what was the aim. If \( d(n) = 0 \) and \( y(n) = 1 \) then \( W_{\text{new}} = W_{\text{old}} - \mu X(n) \) holds, meaning that \( W^T X(n) \) is reduced for \( \mu ||X(n)||^2_2 \).

The coefficient \( \mu \) is the learning coefficient. It is positive. The choice of parameter \( \mu \) value is of great importance for the rate of convergence and learning process of the network. Larger values may reduce the learning period, but also may influence the convergence of the training process.

Example 8.28. Consider a one-neuron neural network. Assume that the activation function of the neuron is unipolar threshold function and that the neuron
is biased. The network has three inputs and one output. Set of data for the neural network training is

\[
X = \begin{bmatrix}
1 & 1 & 0 & 0 & 1 \\
1 & 0 & 1 & 1 & 0 \\
0 & 1 & 1 & 0 & 1 \\
\end{bmatrix}
\]

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

where matrix \( X \) contains the input data and vector \( D \) consists of desired outputs from the neural network for the considered input data values. Train the neural network with \( \mu = 0.5 \).

Since the neuron is biased one more input will be introduced. Its input value is always 1. After this modification the matrix of input data is

\[
X = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 \\
1 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 \\
0 & 1 & 1 & 0 & 1 \\
\end{bmatrix}
\]

Initial values of weighting coefficients are random, for example,

\[
W = \begin{bmatrix}
w_0 \\
w_1 \\
w_2 \\
w_3 \\
\end{bmatrix} = \begin{bmatrix}
-1 \\
1 \\
1 \\
0 \\
\end{bmatrix}
\]

Now we can start the first epoch of training process. We will use all input-output data pairs and calculate the output \( y(n) \) from the neural network. The output \( y(n) \) will be compared with the desired value \( d(n) \) and the coefficients \( W \) will be appropriately modified for each pair of data.

For the first pair of data we have

\[
y(1) = f \left( W^T X(1) \right) = f(\begin{bmatrix}
-1 \\
1 \\
1 \\
0 \\
\end{bmatrix}) = 1.
\]

Since \( d(1) = 1 \) the error \( d(n) - y(n) \) is 0 and the coefficients are not modified.

For the second pair of data

\[
y(2) = f \left( W^T X(2) \right) = 0.
\]

The desired value is \( d(2) = 1 \). Since the error is not zero, the coefficients should be modified as

\[
W_{new} = W_{old} + \mu(d(2) - y(2))X(2) = \begin{bmatrix}
-1 \\
1 \\
1 \\
0 \\
\end{bmatrix} + 0.5 \begin{bmatrix}
1 \\
1 \\
0 \\
1 \\
\end{bmatrix} = \begin{bmatrix}
-0.5 \\
1.5 \\
1 \\
0.5 \\
\end{bmatrix}.
\]
Next pair of input-output data is used. After all data pairs are used, the first epoch of training is finished. Nonzero error appeared in three out of six data pairs. The final value of the coefficients, after the first training epoch, is

\[ W_{\text{epoch } 1} = [-1.5 \, 1 \, 0.5 \, 0]^T. \]

With this initial value, the second epoch of training is completed, using the same input-output pairs of data. After the second epoch nonzero error appeared two times. The final values of the coefficients, after the second epoch, are

\[ W_{\text{epoch } 2} = [-1.5 \, 1 \, 1 \, 0]^T. \]

The process is continued in the third epoch. In the fifth epoch we came to the situation that the neural network has made no error. It means that the training is completed and that more epochs are not needed. The final values of the coefficients are

\[ W = [-1.5 \, 1.5 \, 0.5 \, 0.5]^T. \]

### 8.10.7 One-Layer Neural Network with Continuous Output

In this kind of neural networks the output signal is not binary, but a real number (usually within the interval from 0 to 1). It may be interpreted as a probability that the input data contain or do not contain certain property. In general any interval of real numbers can be a codomain of the output function. The main difference from the perceptron is that we do not require that the neural network achieves an exact precision \( y(n) - d(n) = 0 \). In this case the aim to get a small error in the processing of input results.

Since the output variable is continuous, the activation function should have such a property as well. Consider, for example, the unipolar sigmoid activation function

\[ f(u) = \frac{1}{1 + e^{-u}}. \]

A simple way to quantify the difference of the output signal from the desired signal is to use the square error

\[ \epsilon(n) = \frac{1}{2} (d(n) - y(n))^2 \]

where constant 1/2 is introduced to simplify the notation in the period of the neural network training process. The goal is to minimize the square error. This minimization can be done in various ways: using steepest descent
method, conjugate gradient method, Newton method, are some of these methods. We will use the steepest descend method in the correction of the neural network coefficients.

Consider a network with \( N \) inputs. The input data vector is \( \mathbf{X}(n) \) and the desired output is \( d(n) \). The network output signal is obtained as

\[
y(n) = f(u(n)) = f\left(\mathbf{W}^T \mathbf{X}(n)\right) = f\left(\sum_{k=1}^{n} w_k x_k(n)\right)
\]

with the square error

\[
\varepsilon(n) = \frac{1}{2} (d(n) - y(n))^2 = \frac{1}{2} \left( d(n) - f\left(\sum_{k=1}^{n} w_k x_k(n)\right) \right)^2.
\]

This error is a function of the network coefficients \( w_k \). Using the steepest descend method the coefficients modification is done as

\[
w_{k,\text{new}} = w_{k,\text{old}} - \mu \frac{\partial \varepsilon(n)}{\partial w_k}
\]

or

\[
\mathbf{W}_{\text{new}} = \mathbf{W}_{\text{old}} - \mu \frac{\partial \varepsilon(n)}{\partial \mathbf{W}}
\]

where \( \frac{\partial \varepsilon(n)}{\partial \mathbf{W}} \) is the gradient of error function. Derivatives can be calculated from (8.37) as

\[
\frac{\partial \varepsilon(n)}{\partial w_k} = -(d(n) - y(n)) \frac{\partial y(n)}{\partial w_k} = -(d(n) - y(n)) f'(\sum_{k=1}^{n} w_k x_k(n)) x_k(n).
\]

For the unipolar sigmoid activation functions we have

\[
f'(u) = \frac{d}{du} \frac{1}{1 + e^{-u}} = -\frac{-e^{-u}}{(1 + e^{-u})^2} = \frac{1}{1 + e^{-u}} \frac{e^{-u}}{1 + e^{-u}} =
\]

\[
= f(u) \left(1 - \frac{1}{1 + e^{-u}}\right) = f(u) (1 - f(u)).
\]

Therefore

\[
\frac{\partial \varepsilon(n)}{\partial w_k} = -(d(n) - y(n)) y(n) (1 - y(n)) x_k(n),
\]

where

\[
f\left(\sum_{k=1}^{n} w_k x_k(n)\right) = y(n)
\]
is used. The training rule (correction of the coefficients) is

\[ w_{k,(new)} = w_{k,(old)} + \mu (d(n) - y(n)) y(n) (1 - y(n)) x_k(n) \]

or in vector form

\[ \mathbf{W}_{new} = \mathbf{W}_{old} + \mu (d(n) - y(n)) y(n) (1 - y(n)) \mathbf{X}(n). \]

In is common to denote \((d(n) - y(n)) y(n) (1 - y(n))\) by \(\delta_n\) so that the training rule can be written as

\[ \mathbf{W}_{new} = \mathbf{W}_{old} + \mu \delta_n \mathbf{X}(n). \]

This rule is called delta-rule. Note that the letter \(\delta\) is also used for Dirac delta pulse in some chapters of the book. These two values do not have anything in common.

For the activation function in the form of bipolar sigmoid

\[ f(u) = \frac{2}{1 + e^{-2u}} - 1 = \frac{1 - e^{-2u}}{1 + e^{-2u}} \]

we would have

\[ f'(u) = \frac{4e^{-2u}}{(1 + e^{-2u})^2} = \frac{(1 + e^{-2u})^2 + 4e^{-2u} - (1 + e^{-2u})^2}{(1 + e^{-2u})^2} \]

\[ = 1 - \frac{(1 + 2e^{-2u} + e^{-4u}) - 4e^{-u}}{(1 + e^{-2u})^2} \]

\[ = 1 - \frac{(1 - e^{-2u})^2}{(1 + e^{-u})^2} = 1 - f^2(u) \]

and the value of \(\delta_n\) would be

\[ \delta_n = (d(n) - y(n))(1 - y^2(n)). \]

**Example 8.29.** Neural network consists of one unbiased neuron with two input signals and a sigmoid activation function. Input values are random numbers from the interval \([0,1]\). Available are \(K = 30\) input-output pairs of data. Training of the neural network should be done in 30 epochs with \(\mu = 2\).

Data for network training are obtained as a set of 30 input values of \(x_1\) and \(x_2\). They are assumed as random numbers from the interval from 0 to 1 with a uniform probability density function. For each training pair of random
numbers \( x_1 \) and \( x_2 \) the desired output data is calculated using the formula

\[
d = \frac{1}{2} + \frac{x_1 - 2x_2}{3 + x_1^2 + 3x_2^2}.
\]

Find the total square error after the first, second, fifth and thirtieth epoch. What are the coefficient values at the end of training process? If the input values \( x_1 = 0.1 \) and \( x_2 = 0.8 \) are applied to the network after the training process is completed find the output value \( y \) and compare it with the desired result \( d \) calculated using the formula.

★ Coefficients of the neuron are \( w_1 \) and \( w_2 \). With the sigmoid activation function the coefficient corrections are

\[
\begin{bmatrix}
w_1 \\
w_2
\end{bmatrix}_{\text{new}} = \begin{bmatrix}
w_1 \\
w_2
\end{bmatrix}_{\text{old}} + \mu (d(n) - y(n))y(n)(1 - y(n)) \begin{bmatrix}x_1(n) \\
x_2(n)
\end{bmatrix},
\]

where index \( n \) assumes values from 1 to 30 within one epoch. It denotes the index of the input-output pair of data. The output \( y \) is calculated using

\[
y(n) = f \left( W^T x(n) \right) = f (w_1 x_1(n) + w_2 x_2(n)) .
\]

Initial coefficient values are randomly chosen. The training process is implemented on a computer and the following results are obtained: Total square error after the first epoch of training is 0.4266. After the second training epoch the total error is reduced to 0.1062. The total error after the fifth epoch is 0.0336, while its value at the end of the training process (after 30 epochs) is 0.0170. The final values of the neuron coefficients are

\[
w_1 = 1.0455 \quad w_2 = -1.9401 .
\]

For the input data \( x_1 = 0.1 \) and \( x_2 = 0.8 \) we get

\[
y = f (w_1 x_1 + w_2 x_2) = 0.1904
\]
\[
d = \frac{1}{2} + \frac{x_1 - 2x_2}{3 + x_1^2 + 3x_2^2} = 0.1957 .
\]

The error is small. The task for neural network in this example was to find a complex, nonlinear relation between the input and output data.

\[\Box\]

8.10.8 Multilayer Neural Networks

The multilayer neural networks are characterized by at least one hidden layer, the layer whose values do not appear as the network output. During the training process available are the input-output pairs of data, while the data for hidden layers are not known. On of the approaches to train
this kind of neural network is so called „error backpropagation learning” method. In this method the known error at the output layer of neurons is transformed into the error at the output of the previous neuron layer.

This algorithm will be illustrated on an example of a neural network with two layers and one output. The considered neural network has \( N \) inputs, \( M \) neurons in the hidden layer and one neuron in output layer. The unipolar sigmoid is assumed as the activation function, while the network function is a linear combination of the inputs. Consider one element from the training data set pairs \((X(n), d(n))\). The outputs of the hidden layer are denoted by \( U(n) \) and calculated as

\[
U(n) = f \left( W^T X(n) \right),
\]

where \( W_{N \times M} \) is the matrix of neuron weighting coefficients in the hidden layer, and \( U(n) \) is a vector column with \( M \) output values of the hidden layer. The neural network output is

\[
y(n) = f \left( V^T U(n) \right) = f \left( V^T f \left( W^T X(n) \right) \right)
\]

where \( V_{M \times 1} \) is the vector of weighting coefficients of the output neuron. The square error is

\[
\varepsilon(n) = \frac{1}{2} (d(n) - y(n))^2.
\]

The desired output value is \( d(n) \). Coefficients \( V \) are modified in the same way as in the previous case with the one-layer neural network with continuous output,

\[
v_{m,(new)} = v_{m,(old)} - \mu \frac{\partial \varepsilon(n)}{\partial v_m}
\]

where

\[
\frac{\partial \varepsilon(n)}{\partial v_m} = -(d(n) - y(n)) f'(V^T U(n)) u_m(n)
\]

\[
= -(d(n) - y(n)) y(n) (1 - y(n)) u_m(n).
\]

The \( m \)th element of vector \( U(n) \) is denoted by \( u_m(n) \).

The final modification relations are

\[
V_{new} = V_{old} + \mu (d(n) - y(n)) y(n) (1 - y(n)) U(n)
\]

\[
V_{new} = V_{old} + \mu \delta_n U(n)
\]
where $\delta_n = (d(n) - y(n)) y(n) (1 - y(n))$ is the training rule.

Consider now $k$th neuron in the hidden layer. Coefficients of this neuron are the elements of $k$th column of matrix $W$, denoted by $W_k$. Coefficients of this neuron are modified as

$$w_{pk,(new)} = w_{pk,(old)} - \mu \frac{\partial \varepsilon(n)}{\partial w_{pk}}$$

where

$$\frac{\partial \varepsilon(n)}{\partial w_{pk}} = -(d(n) - y(n)) f'(W^T X(n)) v_k f'(W_k^T x_p(n)).$$

The $p$th element of vector $X(n)$ is denoted by $x_p(n)$, while the $k$th element of vector $V$ is $v_k$. Taking into account that $u_k(n) = f(W_k^T X(n))$ we get

$$\frac{\partial \varepsilon(n)}{\partial w_{pk}} = -(d(n) - y(n)) y(n) (1 - y(n)) v_k [u_k(n)(1 - u_k(n))] x_p(n).$$

Coefficients modification rule for this neuron is

$$w_{pk,(new)} = w_{pk,(old)} + \mu (d(n) - y(n)) y(n) (1 - y(n)) v_k [u_k(n)(1 - u_k(n))] x_p(n)$$

$$\times x_p(n) = w_{pk,(old)} + \mu \delta_{n2} v_k [u_k(n)(1 - u_k(n))] x_p(n)$$

where $\delta_{n2}$ denotes the learning rule for the considered layer of neurons. In vector form we can write

$$W_{k,(new)} = W_{k,(old)} + \mu \delta_{n2} v_k [u_k(n)(1 - u_k(n))] X(n).$$

This is the modification formula for all coefficients of one neuron in the hidden layer. The modification can be generalized to all neurons in the hidden layer

$$W_{(new)} = W_{(old)} + \mu \delta_{n2} X(n) [V \ast U(n) \ast (1 - U(n))]^T,$$

where $\ast$ denotes the element-by-element multiplication, while $1$ is the vector of the same dimension as $U(n)$ whose elements are equal to 1.

The described procedure can be generalized for neural networks with more than two layers. The basic principle is that based on the error in one layer, the coefficients are modified in this layer and then in all layers before the considered layer. It means that the influence of the output error is transferred in an inverse way (backpropagated) to the correction of the coefficients of the layers of neurons.
Example 8.30. Consider a two-layer neural network with two neurons in the hidden layer and one neuron in the output layer. The activation function for all neurons is the unipolar sigmoid. The task for this neural network is to find unknown relation between the input and output data. Step $\mu = 5$ is used in the training process. The data for the training are formed as in Example 8.29, i.e., as a set of $K = 30$ input data $x_1$ and $x_2$ that are uniformly distributed random numbers from the interval $0 \leq x_1 \leq 1$, $0 \leq x_2 \leq 1$. For each training input value of $x_1$ and $x_2$ the desired signal is calculated as

$$d = \frac{1}{2} + \frac{x_1 - 2x_2}{3 + x_1^2 + 3x_2^2}.$$

Find the total square error after 10th, 100th, and 300th epoch. What are the coefficients of the neurons after the training process? If the values of $x_1 = 0.1$ and $x_2 = 0.8$ input the trained neural network find the output $y$ and compare it with the desired result $d$.

The training process is implemented on a computer and the following results are obtained: Total square error after 10 epochs of training is 0.1503. After 100 epochs the total square error is reduced to 0.0036, while the squared error after 300 epochs is 0.0003. The final coefficient values in the hidden and output layers, $W$ and $V$, are

$$W = \begin{bmatrix} -0.2911 & 1.8297 \\ 3.4435 & -0.6945 \end{bmatrix}, \quad V = \begin{bmatrix} -2.6173 \\ 2.5889 \end{bmatrix}.$$

For the input data $x_1 = 0.1$ and $x_2 = 0.8$ we get

$$y = f \left( V^T f \left( W^T \begin{bmatrix} 0.1 \\ 0.8 \end{bmatrix} \right) \right) = 0.1978$$

$$d = \frac{1}{2} + \frac{x_1 - 2x_2}{3 + x_1^2 + 3x_2^2} = 0.1957.$$ 

The error is very small. As expected this result is better than in the case of one-layer neural network (Example 8.29). However, the calculation process is significantly more demanding.

8.10.9 Neural Networks with Unsupervised Learning

Consider an example of one-layer neural network with $N$ neurons and two input data. The input data will be here interpreted as the coordinates of points in plane. If the input data (coordinates of points) exhibit property of being grouped in certain regions of the two-dimensional plane (regions are defined by straight lines passing through the origin), then we can ask the neural network to find the group to which an arbitrary input data pair
(point) belongs. In an ideal case the number of groups (categories) is known and equal to the number of neurons $N$. The training process for a neural network reduces to the selection of the neuron with the highest output (assume that it is the neuron with index $k$) and to the modification of its coefficients using

$$W_{k,(\text{new})} = W_{k,(\text{old})} + \mu \left( X(n) - W_{k,(\text{old})} \right)$$

After the training process we may expect that each of the neurons recognizes one category (belonging to one group) of the input signals. If an uncategorized input signal appears it means that the estimation of the number of neurons is not good. It should be increased and the training process should be continued. When two neuron adjust to the same category, then they produce the same result and one of them can be eliminated. In this way, we may avoid the assumption that the number of categories (groups) or neurons $N$ is known in advance.

Example 8.31. Consider a neural network with two input data and 3 neurons. The task of neural network is to classify the input data in one of three categories. Each neuron corresponds to one category. The classification decision is made by choosing the neuron with the highest output. Activation function is a bipolar sigmoid.

Simulate the neural network in the case when the input data belongs to one of thee categories with equal probability. Data from the first category are pairs of Gaussian random variables with probability density function whose means are $\bar{x}_1 = 0$ and $\bar{x}_2 = 4$ and variances are $\sigma^2_{x_1} = 4$ and $\sigma^2_{x_2} = 0.25$. For the data from the second category the mean values and variances of Gaussian variables are $\bar{x}_1 = 4$, $\bar{x}_2 = -2$, $\sigma^2_{x_1} = 1$ and $\sigma^2_{x_2} = 4$. In the third category are the input data with $\bar{x}_1 = -4$ and $\bar{x}_2 = -2$, $\sigma^2_{x_1} = 1$ and $\sigma^2_{x_2} = 1$. during the training process the step $\mu = 0.5$ is used.

★ Results achieved by neural network after 10 and 100 pairs of input data are presented in Figure 8.39. The categories are indicated with different colors. Learning process of the neural network in the input classification of data is fast.

8.10.10 Voting Machines

Voting machines are special forms of the application of neural networks. Two basic forms of the voting machines are used: neural network ensemble and mixture of experts.

Voting machines of the neural network ensemble type consists of several neural networks which are independently designed and trained.
Chapter 9

Time-Frequency Analysis

The Fourier transform provides a unique mapping of a signal from the time domain to the frequency domain. The frequency domain representation provides the signal’s spectral content. Although the phase characteristic of the Fourier transform contains information about the time distribution of the spectral content, it is very difficult to use this information. Therefore, one may say that the Fourier transform is practically useless for this purpose, i.e., that the Fourier transform does not provide a time distribution of the spectral components.

Depending on problems encountered in practice, various representations have been proposed to analyze non-stationary signals in order to provide time-varying spectral description. The field of the time-frequency signal analysis deals with these representations of non-stationary signals and their properties. Time-frequency representations may roughly be classified as linear, quadratic, and higher order representations.

Linear time-frequency representations exhibit linearity, i.e., the representation of a linear combination of signals equals the linear combination of the individual representations. From this class, the most important one is the short-time Fourier transform (STFT) and its variations. The energetic version of the STFT is called spectrogram. It is the most frequently used tool in time-frequency signal analysis.

The second class of time-frequency representations are the quadratic ones. The most interesting representations of this class are those which provide a distribution of signal energy in the time-frequency plane. They will be referred to as distributions. The concept of a distribution is borrowed from
the probability theory, although there is a fundamental difference. For example, in time-frequency analysis, distributions may take negative values. Other possible domains for quadratic signal representations are the ambiguity domain, the time-lag domain and the frequency-Doppler frequency domain. In order to improve time-frequency representation various higher-order distributions have been defined as well.

9.1 SHORT-TIME FOURIER TRANSFORM

The idea behind the short-time Fourier transform (STFT) is to apply the Fourier transform to a portion of the original signal, obtained by introducing a sliding window function \( w(t) \) to localize the analyzed signal \( x(t) \). The Fourier transform is calculated for the localized part of the signal. It produces the spectral content of the portion of the analyzed signal within the time interval defined by the width of the window function. The STFT (a time-frequency representation of the signal) is then obtained by sliding the window along the signal. Illustration of the STFT calculation is presented in Fig. 9.1.

Analytic formulation of the STFT is

\[
STFT(t, \Omega) = \int_{-\infty}^{\infty} x(t + \tau) w(\tau) e^{-j\Omega \tau} d\tau.
\]  

\( (9.1) \)

From (9.1) it is apparent that the STFT actually represents the Fourier transform of a signal \( x(t) \), truncated by the window \( w(\tau) \) centered at
instant \( t \) (see Fig. 9.1). From the definition, it is clear that the STFT satisfies properties inherited from the Fourier transform (e.g., linearity).

By denoting \( x_1(\tau) = x(t + \tau) \) we can conclude that the STFT is the Fourier transform of the signal \( x_1(\tau)w(\tau) \),

\[
STFT(t, \Omega) = \mathcal{F}_\tau \{ x_1(\tau)w(\tau) \}.
\]

Another form of the STFT, with the same time-frequency performance, is

\[
STFT_{II}(t, \Omega) = \int_{-\infty}^{\infty} x(\tau)w^*(\tau - t)e^{-j\Omega \tau} d\tau
\]

(9.2)

where \( w^*(t) \) denotes the conjugated window function.

It is obvious that definitions (9.1) and (9.2) differ only in phase, i.e., \( STFT_{II}(t, \Omega) = e^{-j\Omega t} STFT(t, \Omega) \) for real valued windows \( w(\tau) \). We will mainly use the first STFT form.

Example 9.1. To illustrate the STFT application, let us perform the time-frequency analysis of the following signal

\[
x(t) = \delta(t - t_1) + \delta(t - t_2) + e^{j\Omega_1 t} + e^{j\Omega_2 t}.
\]

(9.3)

The STFT of this signal equals

\[
STFT(t, \Omega) = w(t_1 - t)e^{-j\Omega(t_1 - t)} + w(t_2 - t)e^{-j\Omega(t_2 - t)}
+ W(\Omega - \Omega_1)e^{j\Omega_1 t} + W(\Omega - \Omega_2)e^{j\Omega_2 t},
\]

(9.4)

where \( W(\Omega) \) is the Fourier transform of the used window. The STFT is depicted in Fig. 9.2 for various window lengths, along with the ideal representation. A wide window \( w(t) \) in the time domain is characterized by a narrow Fourier transform \( W(\Omega) \) and vice versa. Influence of the window to the results will be studied later.

Example 9.2. The STFT of signal

\[
x(t) = e^{iat}
\]

(9.5)

can be approximately calculated for a large \( a \), by using the method of stationary phase. Find its form and the relation for the optimal window \( w(\tau) \) width, assuming that the window is nonzero for \( |\tau| < T \).
Figure 9.2 Time-frequency representation of the sum of two delta pulses and two sinusoids obtained by using (a) wide window, (b) narrow window (c) medium width window and (d) ideal time-frequency representation.

Applying the stationary phase method (1.62), we get

\[
STFT(t, \Omega) = \int_{-\infty}^{\infty} e^{jat(1+\tau)^2} w(\tau) e^{-j\Omega\tau} d\tau
\]

\[
= e^{jat^2} e^{-j(2at-\Omega)^2/4a} w\left(\frac{\Omega - 2at}{2a}\right) \sqrt{\frac{2\pi j}{a}}
\]

since

\[2a(t + \tau_0) = \Omega.\]
Note that the STFT absolute value reduces to

\[ |STFT(t,\Omega)| \simeq \left| w \left( \frac{\Omega - 2at}{2a} \right) \right| \sqrt{\frac{\pi}{a}}. \quad (9.7) \]

In this case, the width of \(|STFT(t,\Omega)|\) along frequency does not decrease with the increase of the window \(w(\tau)\) width. The width of \(|STFT(\Omega,t)|\) around the central frequency \(\Omega = 2at\) is

\[ D = 4aT, \]

where \(2T\) is the window width in the time domain. Note that this relation holds for a wide window \(w(\tau)\), such that the stationary phase method may be applied. If the window is narrow with respect to the phase variations of the signal, the STFT width is defined by the width of the Fourier transform of window. It is proportional to \(1/T\). Thus, the overall STFT width could be approximated by a sum of the frequency variation caused width and the window’s Fourier transform width, that is,

\[ D_o = 4aT + \frac{2c}{T}, \quad (9.8) \]

where \(c\) is a constant defined by the window shape (by using the main lobe as the window width, it will be shown later that \(c = 2\pi\) for a rectangular window or \(c = 4\pi\) for a Hann(ing) window). This relation corresponds to the STFT calculated as a convolution of an appropriately scaled time domain window whose width is \(|\tau| < 2aT\) and the frequency domain form of window \(W(\Omega)\). The approximation is checked against the exact STFT calculated by definition. The agreement is almost complete, Fig.9.3. Therefore, there is a window width \(T\) producing the narrowest possible STFT for this signal. It is obtained by equating the derivative of the overall width to zero,

\[ 4a - \frac{2c}{T^2} = 0, \]

which results in

\[ T_o = \sqrt{\frac{c}{2a}}. \quad (9.9) \]

As expected, for a sinusoid, \(a \rightarrow 0\), \(T_o \rightarrow \infty\). This is just an approximation of the optimal window, since for narrow windows we may not apply the stationary phase method (the term \(4aT\) is then much smaller than \(2c/T\) and may be neglected anyway).

Note that for \(a = 1/2\), when the instantaneous frequency is a symmetry line for the time and the frequency axis

\[ 2 - \frac{2c}{T^2} = 0 \text{ or } 2T = \frac{2c}{T}, \]

meaning that the optimal window should have the widths equal in the time-domain \(2T\) and in the frequency domain \(2c/T\) (main lobe width).
The STFT can be expressed in terms of the signal’s Fourier transform

\[
STFT(t, \Omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} X(\theta) e^{j(t+\tau)\theta} \omega(\tau) e^{-j\Omega \tau} \, d\theta \, d\tau
\]

\[
= \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\theta) W(\Omega - \theta) e^{j\Omega \theta} \, d\theta = \left[ X(\Omega) e^{j\Omega} \right] \ast_{\Omega} W(\Omega). \tag{9.10}
\]

where \( \ast_{\Omega} \) denotes convolution in \( \Omega \). It may be interpreted as an inverse Fourier transform of the frequency localized version of \( X(\Omega) \), with localization window \( W(\Omega) = \text{FT}\{w(\tau)\} \).

The energetic version of the STFT, called the spectrogram, is defined by

\[
SPEC(t, \Omega) = |STFT(t, \Omega)|^2
\]

\[
= \int_{-\infty}^{\infty} x(\tau) w^*(\tau - t) e^{-j\Omega \tau} \, d\tau = \int_{-\infty}^{\infty} x(t + \tau) w^*(\tau) e^{-j\Omega \tau} \, d\tau.
\]

Obviously, linearity property is lost in the spectrogram.
Example 9.3. For illustration consider two different signals \( x_1(t) \) and \( x_2(t) \) producing the same amplitude of the Fourier transform, Fig. 9.4,

\[
x_1(t) = \sin \left( 122\pi \frac{t}{128} \right) - \cos \left( 42\pi \frac{t}{128} - \frac{16\pi}{11} \left( t - \frac{128}{64} \right)^2 \right) \\
- 1.2 \cos \left( 94\pi \frac{t}{128} - 2\pi \left( t - \frac{128}{64} \right)^2 - \frac{16\pi}{11} \left( t - \frac{120}{64} \right)^3 \right) e^{-\frac{(t-50)^2}{2}} \\
- 1.6 \cos \left( 15\pi \frac{t}{128} - 2\pi \left( t - \frac{50}{64} \right)^2 \right) e^{-\frac{(t-50)^2}{2}}
\]

(9.11)

\[
x_2(t) = x_1(255 - t).
\]

Their spectrograms are presented in Fig.9.5. From the spectrograms we can follow time variations of the spectral content. The signals obviously consist of one constant high frequency component, one linear frequency component (in the first signal with increasing frequency as time progresses, and in the second signal with decreasing frequency), and two chirps (one appearing at different time instants and the other having different frequency variations).

The signal can be obtained from the STFT calculated at an instant \( t_0 \) as

\[
x(t_0 + \tau) = \frac{1}{w(\tau)} \int_{-\infty}^{\infty} STFT(t_0, \Omega) e^{-j\Omega \tau} d\tau.
\]
This relation can be theoretically used for the signal within the region $w(\tau) \neq 0$. In practice it is used within the region of significant window $w(\tau)$ values.

If the window is shifted for $R$, for each next STFT calculation, then a set of values

$$x(t_0 + i R + \tau)w(\tau) = \int_{-\infty}^{\infty} STFT(t_0 + i R, \Omega)e^{-j \Omega \tau} d\tau$$
is obtained. If the value of step $R$ is smaller than the window duration then the same signal value is used within two (several) windows. Using a change of variables $iR + \tau = \lambda$ and summing over all overlapping windows we get

$$x(t_0 + \lambda) \sum_i w(\lambda - iR) = \sum_i \int_{-\infty}^{\infty} STFT(t_0 + iR, \Omega) e^{-j\Omega \lambda} e^{i\Omega iR} d\lambda.$$  

Values of $i$ in the summation are such that for a given $\lambda$ and $R$ the value of $iR - \lambda = \tau$ is within the window $w(\tau)$.

If the sum of shifted versions of the windows is constant (without loss of generality assume equal to 1), $\sum w(\tau - iR) = 1$, then

$$x(t_0 + \lambda) = \sum_i \int_{-\infty}^{\infty} STFT(t_0 + iR, \Omega) e^{-j\Omega \lambda} e^{i\Omega iR} d\lambda$$

for any $\lambda$. Condition $\sum_i w(\tau - iR) = 1$ means that a periodic extension of the window, with period $R$, is constant. Periodic extension of a continuous signal corresponds to the sampling of the window Fourier transform at $\Omega = \frac{2\pi}{R} n$ in the Fourier domain, (1.59). It means that $W(\frac{2\pi}{R} n) = 0$ when $n \neq 0$ for $\sum_i w(\lambda - iR) = 1$.

### 9.2 WINDOWS

The window function plays a crucial role in the localization of the signal in the time-frequency plane. The most commonly used windows will be presented next.

#### 9.2.1 Rectangular Window

The simplest window is the rectangular one, defined by

$$w(\tau) = \begin{cases} 1 & \text{for } |\tau| < T \\ 0 & \text{elsewhere} \end{cases} \quad (9.12)$$

whose Fourier transform is

$$W_R(\Omega) = \int_{-T}^{T} e^{-j\Omega \tau} d\tau = \frac{2\sin(\Omega T)}{\Omega}. \quad (9.13)$$
The rectangular window function has very strong and oscillatory sidelobes in the frequency domain, since the function \( \sin(\Omega T)/\Omega \) converges very slowly, toward zero, in \( \Omega \) as \( \Omega \to \pm \infty \). Slow convergence in the Fourier domain is caused by a significant discontinuity in time domain, at \( t = \pm T \). The mainlobe width of \( W_R(\Omega) \) is \( d_\Omega = 2\pi/T \). In order to enhance signal localization in the frequency domain, other window functions have been introduced.

The discrete-time form of the rectangular window is

\[
w(n) = u(n+N/2) - u(n-N/2)
\]

with the Fourier transform

\[
W(e^{j\omega}) = \sum_{n=-N/2}^{N/2-1} e^{-j\omega n} = \frac{\sin(\omega N/2)}{\sin(\omega/2)}.
\]

### 9.2.2 Triangular (Bartlett) Window

It is defined by

\[
w(\tau) = \begin{cases} 
1 - |\tau/T| & \text{for } |\tau| < T \\
0 & \text{elsewhere.}
\end{cases} \quad (9.14)
\]

It could be considered as a convolution of the rectangular window of duration \( T \) with itself, since

\[
[u(t+T/2) - u(t-T/2)] * [u(t+T/2) - u(t-T/2)] = (1 - |\tau/T|)[u(t+T) - u(t-T)].
\]

The Fourier transform of the triangular window is a product of two Fourier transforms of the rectangular window of the width \( T \),

\[
W_T(\Omega) = \frac{4\sin^2(\Omega T/2)}{\Omega^2}. \quad (9.15)
\]

Convergence of this function toward zero as \( \Omega \to \pm \infty \) is of the \( 1/\Omega^2 \) order. It is a continuous function of time, with discontinuities in the first derivative at \( t = 0 \) and \( t = \pm T \). The mainlobe of this window function is twice wider in the frequency domain than in the rectangular window case. Its width follows from \( \Omega T/2 = \pi \) as \( d_\Omega = 4\pi/T \).

The discrete-time form is

\[
w(n) = \left[ 1 - 2|n|/N \right] [u(n+N/2) - u(n-N/2)].
\]
In the frequency domain its form is
\[ W(e^{j\omega}) = \sum_{n=-N/2}^{N/2-1} \left[ 1 - \frac{2|n|}{N} \right] e^{-j\omega n} = \frac{\sin^2(\omega N/4)}{\sin^2(\omega/2)}. \]

### 9.2.3 Hann(ing) Window

This window is of the form
\[ w(\tau) = \begin{cases} 
0.5(1 + \cos(\pi\tau/T)) & \text{for } |\tau| < T \\
0 & \text{elsewhere.}
\end{cases} \quad (9.16) \]

Since \( \cos(\pi\tau/T) = \frac{\exp(j\pi\tau/T) + \exp(-j\pi\tau/T)}{2} \), the Fourier transform of this window is related to the Fourier transform of the rectangular window of the same width as
\[
W_H(\Omega) = \frac{1}{2} W_R(\Omega) + \frac{1}{4} W_R(\Omega - \pi/T) + \frac{1}{4} W_R(\Omega + \pi/T) \\
= \frac{\pi^2 \sin(\Omega T)}{\Omega(\pi^2 - \Omega^2 T^2)}. \quad (9.17)
\]

The function \( W_H(\Omega) \) decays in frequency as \( \Omega^3 \), much faster than \( W_R(\Omega) \).

The discrete-time domain form is
\[
w(n) = 0.5 \left[ 1 + \cos \left( \frac{2\pi n}{N} \right) \right] \left[ u(n+N/2) - u(n-N/2) \right]
\]
with the DFT of the form
\[ W(k) = \frac{N}{2} \delta(k) + \frac{N}{4} \delta(k+1) + \frac{N}{4} \delta(k-1). \]

If the window is used on the data set from 0 to \( N-1 \) then
\[
w(n) = 0.5 \left[ 1 - \cos \left( \frac{2\pi n}{N} \right) \right] \left[ u(n) - u(n-N) \right]
\]
\[ W(k) = \frac{N}{2} \delta(k) - \frac{N}{4} \delta(k+1) - \frac{N}{4} \delta(k-1). \]

If a signal is multiplied by the Hann(ing) window the previous relation also implies the relationship between the DFTs of the signal \( x(n) \) calculated using the rectangular and Hann(ing) windows. The DFT of windowed
signal is moving average (smoothed) form of the original signal,

\[
\text{DFT}\{x(n)w(n)\} = \frac{1}{N} \text{DFT}\{x(n)\} *_k \text{DFT}\{w(n)\} \\
= \frac{1}{4} X(k + 1) + \frac{1}{2} X(k) + \frac{1}{4} X(k - 1)
\]

**Example 9.4.** Find the window that will correspond to the frequency smoothing \((X(k + 1) + X(k) + X(k - 1))/3\), i.e., to

\[
\text{DFT}\{x(n)w(n)\} = \frac{1}{N} \text{DFT}\{x(n)\} *_k \text{DFT}\{w(n)\} \\
= \frac{1}{3} X(k + 1) + \frac{1}{3} X(k) + \frac{1}{3} X(k - 1).
\]

★The DFT of this window is

\[
W(k) = \frac{N}{3} \delta(k) + \frac{N}{3} \delta(k + 1) + \frac{N}{3} \delta(k - 1).
\]

In the discrete-time domain the window form is

\[
w(n) = \frac{1}{3} \left[ 1 + 2\cos\left(\frac{2\pi n}{N}\right) \right] [u(n) - u(n - N)].
\]

**Example 9.5.** Find the formula to calculate the STFT with a Hann(ing) window, if the STFT calculated with a rectangular window is known.

★From the frequency domain STFT definition

\[
\text{STFT}(t, \Omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\theta)W(\Omega - \theta)e^{j\theta} d\theta
\]

easily follows that, if we use the window,

\[
W_H(\Omega) = \frac{1}{2} W_R(\Omega) + \frac{1}{4} W_R(\Omega - \pi/T) + \frac{1}{4} W_R(\Omega + \pi/T),
\]

then

\[
\text{STFT}_H(t, \Omega) = \frac{1}{2} \text{STFT}_R(t, \Omega) + \frac{1}{4} \text{STFT}_R(t, \Omega - \pi/T) + \frac{1}{4} \text{STFT}_R(t, \Omega + \pi/T). \tag{9.18}
\]

\[
+ \frac{1}{4} \text{STFT}_R(2t, \Omega - \pi/T) + \frac{1}{4} \text{STFT}_R(2t, \Omega + \pi/T). \tag{9.19}
\]

\[\square\]
For the Hann(ing) window \( w(\tau) \) of the width \( 2T \), we may roughly assume that its Fourier transform \( W_H(\Omega) \) is nonzero within the main lattice \( |\Omega| < 2\pi/T \) only, since the sidelobes decay very fast. Then we may write \( d_\Omega = 4\pi/T \). It means that the STFT is nonzero valued in the shaded regions in Fig. 9.2.

We see that the duration in time of the STFT of a delta pulse is equal to the widow width \( d_t = 2T \). The STFTs of two delta pulses (very short duration signals) do not overlap in time-frequency domain if their distance is greater than the window duration \( |t_1 - t_2| > d_t \). Then, these two pulses can be resolved. Thus, the window width is here a measure of time resolution. Since the Fourier transform of the Hann(ing) window converges fast, we can roughly assume that a measure of duration in frequency is the width of its mainlobe, \( d_\Omega = 4\pi/T \). Then we may say that the Fourier transforms of two sinusoidal signals do not overlap in frequency if the condition \( |\Omega_1 - \Omega_2| > d_\Omega \) holds. It is important to observe that the product of the window durations in time and frequency is a constant. In this example, considering time domain duration of the Hann(ing) window and the width of its mainlobe in the frequency domain, this product is \( d_t d_\Omega = 8\pi \). Therefore, if we improve the resolution in the time domain \( d_t \), by decreasing \( T \), we inherently increase the value of \( d_\Omega \) in the frequency domain. This essentially prevents us from achieving the ideal resolution \((d_t = 0 \text{ and } d_\Omega = 0)\) in both domains. A general formulation of this principle, stating that the product of effective window durations in time and in frequency cannot be arbitrarily small, will be presented later.

### 9.2.4 Hamming Window

This window has the form

\[
w(\tau) = \begin{cases} 
0.54 + 0.46\cos(\pi\tau/T) & \text{for } |\tau| < T \\
0 & \text{elsewhere.} 
\end{cases}
\]  

(9.20)

A similar relation between the Hamming and the rectangular window transforms holds, as in the case of Hann(ing) window.

The Hamming window was derived starting from

\[
w(\tau) = a + (1-a)\cos(\pi\tau/T)
\]

within \( |\tau| < T \), with

\[
W(\Omega) = a\frac{2\sin(\Omega T)}{\Omega} + (1-a)\left(\frac{\sin((\Omega - \pi/T)T)}{\Omega - \pi/T} + \frac{\sin((\Omega + \pi/T)T)}{\Omega + \pi/T}\right).
\]
If we choose such a value of $a$ to cancel out the second sidelobe at its maximum (i.e., at $\Omega T \cong 2.5\pi$) then we get

$$0 = \frac{2aT}{2.5\pi} - (1 - a)\left(\frac{T}{1.5\pi} + \frac{T}{3.5\pi}\right)$$

resulting in

$$a = \frac{25}{46} \cong 0.54. \quad (9.21)$$

This window has several sidelobes, next to the mainlobe, lower than the previous two windows. However, since it is not continuous at $t = \pm T$, its decay in frequency, as $\Omega \to \pm \infty$, is not fast. Note that we let the mainlobe to be twice wider than in the rectangular window case, so we cancel out not the first but the second sidelobe, at its maximum.

The discrete-time domain form is

$$w(n) = \left[0.54 + 0.46\cos\left(\frac{2\pi n}{N}\right)\right] [u(n + N/2) - u(n - N/2)]$$

with

$$W(k) = 0.54N\delta(k) + 0.23N\delta(k + 1) + 0.23N\delta(k - 1).$$

9.2.5 Blackman and Kaiser Windows

In some applications it is crucial that the sidelobes are suppressed, as much as possible. This is achieved by using windows of more complicated forms, like the Blackman window. It is defined by

$$w(\tau) = \begin{cases} 
0.42 + 0.5\cos(\pi\tau/T) + 0.08\cos(2\pi\tau/T) & \text{for } |\tau| < T \\
0 & \text{elsewhere.}
\end{cases} \quad (9.22)$$

This window is derived from

$$w(\tau) = a_0 + a_1\cos(\pi\tau/T) + a_2\cos(2\pi\tau/T)$$

with $a_0 + a_1 + a_2 = 1$ and canceling out the Fourier transform values $W(\Omega)$ at the positions of the third and the fourth sidelobe maxima (i.e., at $\Omega T \cong 3.5\pi$ and $\Omega T \cong 4.5\pi$). Here, we let the mainlobe to be three times wider than in the rectangular window case, so we cancel out not the first nor the second but the third and fourth sidelobes, at their maxima.
The discrete-time and frequency domain forms are

\[
\begin{align*}
    w(n) &= \left[ 0.42 + 0.5 \cos \left( \frac{2\pi n}{N} \right) + 0.08 \cos \left( \frac{4\pi n}{N} \right) \right] \left[ u(n + \frac{N}{2}) - u(n - \frac{N}{2}) \right] \\
    W(k) &= \left[ 0.42 \delta(k) + 0.25(\delta(k + 1) + \delta(k - 1)) + 0.04(\delta(k + 2) + \delta(k - 2)) \right] N.
\end{align*}
\]

Further reduction of the sidelobes can be achieved by, for example, the Kaiser (Kaiser-Bessel) window. It is an approximation to a restricted time duration function with minimum energy outside the mainlobe. This window is defined by using the zero-order Bessel functions, with a localization parameter. It has the ability to keep the maximum energy within the mainlobe, while minimizing the sidelobe energy. The sidelobe level can be as low \(-70\) dB, as compared to the mainlobe, and even lower. This kind of window is used in the analysis of signals with significantly different amplitudes, when the sidelobe of one component can be much higher than the amplitude of the mainlobe of other components.

These are just a few of the windows used in signal processing. Some windows, along with the corresponding Fourier transforms, are presented in Fig. 9.6.

**Example 9.6.** Calculate the STFT at \(t = 0\) with a Hamming and Blackman window of the signals \(x_1(t) = 2\cos(4\pi t/T) + 2\cos(12\pi t/T)\) and \(x_2(t) = 2\cos(4\pi t/T) + 0.001\cos(64\pi t/T)\) with \(T = 128\) using in numerical calculation \(\Delta t = 1\). Comment the results.

The STFT at \(t = 0\) is shown in Fig. 9.7. The resolution of close components in \(x_1(t)\) is better with the Hann(ing) than with the Blackman window, since the main lobe of the Blackman window is wider. Small signal in \(x_2(t)\) is visible in the STFT with the Blackman window since its side-lobes are much lower than in the Hamming window.

9.2.6 **Discrete Form and Realizations of the STFT**

In numerical calculations the integral form of the STFT should be discretized. By sampling the signal with sampling interval \(\Delta t\) we get

\[
\begin{align*}
    \text{STFT}(t, \Omega) &= \int_{-\infty}^{\infty} x(t + \tau)w(\tau)e^{-j\Omega \tau} d\tau \\
    &\approx \sum_{m=-\infty}^{\infty} x\left( (n + m)\Delta t \right)w(m\Delta t)e^{-jm\Omega \Delta t} \Delta t.
\end{align*}
\]
Figure 9.6  Windows in the time and frequency domains: rectangular window (first row), triangular (Bartlett) window (second row), Hann(ing) window (third row), Hamming window (fourth row), and Blackman window (fifth row).
By denoting
\[ x(n) = x(n\Delta t)\Delta t \]
and normalizing the frequency \( \Omega \) by \( \Delta t, \omega = \Delta t\Omega \), we get the time-discrete form of the STFT as
\[
\text{STFT}(n, \omega) = \sum_{m=-\infty}^{\infty} w(m)x(n + m)e^{-jm\omega}.
\] (9.23)

We will use the same notation for continuous-time and discrete-time signals, \( x(t) \) and \( x(n) \). However, we hope that this will not cause any confusion since we will use different sets of variables, for example \( t \) and \( \tau \) for continuous time and \( n \) and \( m \) for discrete time. Also, we hope that the context will be always clear, so that there is no doubt what kind of signal is considered.
It is important to note that \( STFT(n, \omega) \) is periodic in frequency with period \( 2\pi \). The relation between the analog and the discrete-time form is

\[
STFT(n, \omega) = \sum_{k=-\infty}^{\infty} STFT(n\Delta t, \Omega + 2k\Omega_0) \quad \text{with} \quad \omega = \Delta t \Omega.
\]

The sampling interval \( \Delta t \) is related to the period in frequency as

\[
\Delta t = \frac{\pi}{\Omega_0}.
\]

According to the sampling theorem, in order to avoid the overlapping of the STFT periods (aliasing), we should take

\[
\Delta t = \frac{\pi}{\Omega_0} \leq \frac{\pi}{\Omega_m}
\]

where \( \Omega_m \) is the maximal frequency in the STFT. Strictly speaking, the windowed signal \( x(t + \tau)w(\tau) \) is time limited, thus it is not frequency limited. Theoretically, there is no maximal frequency since the width of the window’s Fourier transform is infinite. However, in practice we can always assume that the value of spectral content of \( x(t + \tau)w(\tau) \) above frequency \( \Omega_m \), i.e., for \( |\Omega| > \Omega_m \), can be neglected, and that overlapping of the frequency content above \( \Omega_m \) does not degrade the basic frequency period.

The discretization in frequency should be done by a number of samples greater than or equal to the window length \( N \). If we assume that the number of discrete frequency points is equal to the window length, then

\[
STFT(n,k) = STFT(n, \omega) \big|_{\omega = \frac{2\pi k}{N}} = \sum_{m=-N/2}^{N/2-1} w(m)x(n+m)e^{-j2\pi mk/N} \quad (9.24)
\]

and it can be efficiently calculated using the fast DFT routines

\[
STFT(n,k) = DFT_m \{w(m)x(n+m)\},
\]

for a given instant \( n \). When the DFT routines with indices from 0 to \( N - 1 \) are used, then a shifted version of \( w(m)x(n+m) \) should be formed for the calculation for \( N/2 \leq m \leq N - 1 \). It is obtained as \( w(m - N)x(n + m - N) \), since in the DFT calculation periodicity of the signal \( w(m)x(n+m) \), with period \( N \), is inherently assumed.

**Example 9.7.** Consider a signal with \( M = 16 \) samples, \( x(0), x(1), \ldots, x(15) \), write a matrix form for the calculation of a four-sample STFT. Present nonoverlapping and overlapping cases of the STFT calculation.
For the calculation of (9.24) with \( N = 4 \), when \( k = -2, -1, 0, 1 \), for given instant \( n \), the following matrix notation can be used

\[
\begin{bmatrix}
\text{STFT}(n, -2) \\
\text{STFT}(n, -1) \\
\text{STFT}(n, 0) \\
\text{STFT}(n, 1)
\end{bmatrix} = \begin{bmatrix}
W_4^4 & W_4^2 & 1 & W_4^{-2} \\
W_4^2 & W_4^4 & 1 & W_4^{-1} \\
1 & 1 & 1 & 1 \\
W_4^{-2} & W_4^{-1} & 1 & W_4^1
\end{bmatrix} \begin{bmatrix}
x(n - 2) \\
x(n - 1) \\
x(n) \\
x(n + 1)
\end{bmatrix}
\]

or

\[
\text{STFT}(n) = W_4 x(n)
\]

with \( \text{STFT}(n) = [\text{STFT}(n, -2) \ \text{STFT}(n, -1) \ \text{STFT}(n, 0) \ \text{STFT}(n, 1)]^T \), \( x(n) = [x(n - 2) \ x(n - 1) \ x(n) \ x(n + 1)]^T \), and \( W_4 \) is the DFT matrix of order four with elements \( W_4^{m,k} = \exp(-j2\pi mk/N) \). Here a rectangular window is assumed. Including the window function, the previous relation can be written as

\[
\text{STFT}(n) = W_4 H_4 x(n),
\]

with

\[
H_4 = \begin{bmatrix}
w(-2) & 0 & 0 & 0 \\
0 & w(-1) & 0 & 0 \\
0 & 0 & w(0) & 0 \\
0 & 0 & 0 & w(1)
\end{bmatrix}
\]

being a diagonal matrix whose elements are the window values \( w(m) \), \( H_4 = \text{diag}(w(m)) \), \( m = -2, -1, 0, 1 \) and

\[
W_4 H_4 = \begin{bmatrix}
w(-2)W_4^4 & w(-1)W_4^2 & w(0) & w(1)W_4^{-2} \\
w(-2)W_4^2 & w(-1)W_4^4 & w(0) & w(1)W_4^{-1} \\
w(-2) & w(-1) & w(0) & w(1) \\
w(-2)W_4^{-2} & w(-1)W_4^{-1} & w(0) & w(1)W_4^1
\end{bmatrix}.
\]

All STFT values for the nonoverlapping case are obtained as

\[
\text{STFT} = W_4 H_4 \begin{bmatrix}
x(0) \\
x(1) \\
x(2) \\
x(3)
\end{bmatrix} = W_4 H_4 X_{4,4}
\]

where \( \text{STFT} \) is a matrix of the STFT values with columns corresponding to the calculation instants and the rows to the frequencies. This matrix is of the form

\[
\text{STFT} = \begin{bmatrix}
\text{STFT}_M(0) & \text{STFT}_M(1) & \cdots & \text{STFT}_M(N - M)
\end{bmatrix}
\]

\[
= \begin{bmatrix}
\text{STFT}(2, -2) & \text{STFT}(6, -2) & \text{STFT}(10, -2) & \text{STFT}(14, -2) \\
\text{STFT}(2, -1) & \text{STFT}(6, -1) & \text{STFT}(10, -1) & \text{STFT}(14, -1) \\
\text{STFT}(2, 0) & \text{STFT}(6, 0) & \text{STFT}(10, 0) & \text{STFT}(14, 0) \\
\text{STFT}(2, 1) & \text{STFT}(6, 1) & \text{STFT}(10, 1) & \text{STFT}(14, 1)
\end{bmatrix}.
\]
Matrix $X_{4,4}$ is formed of by using four successive signal values in each column. Notation $X_{N,R}$ will be used to denote the signal matrix with columns containing $N$ signal values and the difference of the first signal value indices in the successive columns is $R$. For $R = N$ the nonoverlapping calculation is performed.

For a STFT calculation with overlapping $R < N$, for example with the time step in the STFT calculation $R = 1$, we get

$$\text{STFT} = H_4 W_4 = W_4 H_4 X_{4,1}.$$  

The step $R$ defines the difference of arguments in two neighboring columns. In the first case the difference of arguments in two neighboring columns was 4 (time step in the STFT calculation was $R = 4$ equal to the window width, meaning nonoverlapping calculation). In the second example difference is $R = 1 < 4$, meaning the overlapped STFT calculation. Note that the window function $H_N$ and the DFT matrix $W_N$ remain the same for both cases.

**Example 9.8.** Consider a signal

$$x(t) = e^{-t^2}e^{-j6\pi t^2-j32\pi t} + e^{-4(t-1)^2}e^{j16\pi t^2+j160\pi t}.$$  

Assuming that the values of the signal with amplitudes bellow $1/e^4$ could be neglected, find the sampling rate for the STFT-based analysis of this signal. Write the approximate spectrogram expression for the Hann(ing) window of $N = 32$ samples in the analysis. What signal will be presented in the time-frequency plane, within the basic frequency period, if the signal is sampled at $\Delta t = 1/128$?

★The time interval, with significant signal content, for the first signal component is $-2 \leq t \leq 2$, with the frequency content within $-56\pi \leq \Omega \leq -8\pi$, since the instantaneous frequency is $\Omega(t) = -12\pi t - 32\pi$. For the second component these intervals are $0 \leq t \leq 2$ and $160\pi \leq \Omega \leq 224\pi$. The maximal frequency in the signal is $\Omega_m = 224\pi$. Here we have to take into account possible spreading of the spectrum caused by the lag window. Its width in the time domain is $d_t = 2T = N\Delta t = 32\Delta t$. Width of the mainlobe in frequency domain $d_\omega$ is defined by $32d_\omega\Delta t = 4\pi$, or $\Omega_w = \pi/(8\Delta t)$. Thus, taking the sampling interval $\Delta t = 1/256$, we will satisfy the sampling theorem condition in the worst instant case, since $\pi/(\Omega_m + d_\omega) = 1/256$.

In the case of the Hann(ing) window with $N = 32$ and $\Delta t = 1/256$, the lag interval is $N\Delta t = 1/8$. We will assume that the amplitude variations within the window are small, that is, $w(\tau)e^{-((t+\tau)^2)} \simeq w(\tau)e^{-t^2}$ for $-1/16 < \tau \leq 1/16$. Then, according to the stationary phase method, we can write the
STFT approximation,

$$|STFT(t, \Omega)|^2 = \frac{1}{6} e^{-2t^2} w^2 \left(\frac{\Omega+12\pi t+32\pi}{12\pi}\right) + \frac{1}{32} e^{-8(t-1)^2} w^2 \left(\frac{\Omega-32\pi t-160\pi}{32\pi}\right)$$

with $t = n/256$ and $\Omega = 256\omega$ within $-\pi \leq \omega < \pi$.

In the case of $\Delta t = 1/128$ the signal will be periodically extended with period \(2\Omega_0 = 256\pi\). The basic period will be for $-256\pi < \Omega < 256\pi$. It means that the first component will remain unchanged within the basic period, while the second component is outside the basic period. However, its replica shifted for one period to the left, that is, for $-256\pi$, will be within the basic period. It will be located within $160\pi - 256\pi \leq \Omega \leq 224\pi - 256\pi$, that is, within $-96\pi \leq \Omega \leq -32\pi$. Thus, the signal represented by the STFT in this case will correspond to

$$x_r(t) = e^{-t^2} e^{-j6\pi t^2} - \beta 32\pi t + e^{-4(t-1)^2} e^{j16\pi t^2} + j(160-256)\pi t,$$

with approximation,

$$|STFT(t, \Omega)|^2 = \frac{1}{6} e^{-2t^2} w^2 \left(\frac{\Omega+12\pi t+32\pi}{12\pi}\right) + \frac{1}{32} e^{-8(t-1)^2} w^2 \left(\frac{\Omega-32\pi t-96\pi}{32\pi}\right),$$

with $t = n/128$ and $\Omega = 128\omega$ within $-\pi \leq \omega < \pi$ or $-128\pi \leq \Omega < 128\pi$. □

### 9.2.7 Recursive STFT Realization

For the rectangular window, the STFT values at an instant $n$ can be calculated recursively from the STFT values at $n-1$, as

$$STFT_R(n, k) = [x(n + N/2 - 1) - x(n - N/2 - 1)](-1)^k e^{j2\pi k/N} + STFT_R(n - 1, k) e^{j2\pi k/N}.$$ 

This recursive formula follows easily from the STFT definition (9.24).

For other window forms, the STFT can be obtained from the STFT obtained by using the rectangular window. For example, according to (9.18) the STFT with Hann(ing) window $STFT_H(n, k)$ is related to the STFT with rectangular window $STFT_R(n, k)$ as

$$STFT_H(n, k) = \frac{1}{2} STFT_R(n, k) + \frac{1}{4} STFT_R(n, k - 1) + \frac{1}{4} STFT_R(n, k + 1).$$

This recursive calculation is important for hardware implementation of the STFT and other related time-frequency representations (e.g., the higher order representations implementations based on the STFT).
A system for the recursive implementation of the STFT is shown in Fig. 9.8. The STFT obtained by using the rectangular window is denoted by $STFT_R(n,k)$, Fig. 9.8, while the values of coefficients are

$$ (a_{-1}, a_0, a_1) = \left( \frac{1}{4}, \frac{1}{2}, \frac{1}{4} \right), $$

$$ (a_{-1}, a_0, a_1) = (0.23, 0.54, 0.23), $$

$$ (a_{-2}, a_{-1}, a_0, a_1, a_2) = (0.04, 0.25, 0.42, 0.25, 0.04) $$

for the Hann(ing), Hamming and Blackman windows, respectively.

Note that in general instead of multiplying the signal by the previous window functions, for each calculation instant $n$, the STFT matrix $STFT$ can be calculated without window multiplication (using a rectangular window). The STFT matrix for the Hann(ing) window, for example, is obtained as $STFT_H = 0.5STFT + 0.25STFT_{\downarrow} + 0.25STFT_{\uparrow}$, where $STFT_{\downarrow}$ and $STFT_{\uparrow}$ are the STFT matrices with circularly shifted rows down and up for one position, respectively.

### 9.2.8 Filter Bank STFT Implementation

According to (9.1), the STFT can be written as a convolution
Figure 9.9 Filter bank realization of the STFT

\[ \text{STFT}(t, \Omega) = \int_{-\infty}^{\infty} x(t + \tau) w(\tau) e^{-j\Omega \tau} d\tau \]

where an even, real valued, window function is assumed, \( w(\tau) = w(-\tau) \). For a discrete set of frequencies \( \Omega_k = k\Delta\Omega = 2\pi k / (N\Delta t), k = 0, 1, 2, ..., N-1 \), and discrete values of signal, we get that the discrete STFT, (9.24), is an output of the filter bank with impulse responses

\[ \text{STFT}(n, k) = x(n) *_n \left[ w(n)e^{j2\pi kn/N} \right] = x(n) *_n h_k(n) \]

\[ h_k(n) = w(n)e^{j2\pi kn/N} \]

\[ k = 0, 1, ..., N - 1 \]

what is illustrated in Fig.9.9. The next STFT can be calculated with time step \( R\Delta t \), meaning downsampling in time with factor \( 1 \leq R \leq N \). Two special cases are: no downsampling, \( R = 1 \), and nonoverlapping calculation, \( R = N \). Influence of \( R \) to the signal reconstruction will be discussed later.
9.2.8.1 Overlapping windows

Nonoverlapping cases are important and easy for analysis. They also keep the number of the STFT coefficients equal to the number of the signal samples. However, the STFT is commonly calculated using overlapping windows. There are several reasons for introducing overlapped STFT representations. Rectangular windows have poor localization in the frequency domain. The localization is improved by other window forms. In the case of nonrectangular windows some of the signal samples are weighted in such a way that their contribution to the final representation is small. Then we want to use additional STFT with a window positioned in such a way that these samples contribute more to the STFT calculation. Also, in the parameters estimation and detection the task is to achieve the best possible estimation or detection for each time instant instead of using interpolations for the skipped instants when the STFT with a big step (equal to the window width) is calculated. Commonly, the overlapped STFTs are calculated using, for example, rectangular, Hann(ing), Hamming, Bartlett, Kaiser, or Blackman window of a constant window width \( N \) with steps \( N/2, N/4, N/8, \ldots \) in time. Computational cost is increased in the overlapped STFTs since more STFTs are calculated. A way of composing STFTs calculated with a rectangular window into a STFT with, for example, the Hann(ing), Hamming, or Blackman window, is presented in Fig. 9.8.

If a signal \( x(n) \) is of duration \( M \), in some cases in addition to the overlapping in time, an interpolation in frequency is done, for example up to the DFT grid with \( M \) samples. The overlapped and interpolated STFT of this signal is calculated, using a window \( w(m) \) whose width is \( N \leq M \), as

\[
\text{STFT}_N(n,k) = \sum_{m=-N/2}^{N/2-1} w(m)x(n+m)e^{-j2\pi mk/M}
\]

\( n = N/2 + 1, N/2 + 2, \ldots, M - N/2 \)

\( k = -M/2, -M/2 + 1, \ldots, -1, 0, 1, \ldots, M/2 - 1 \).

**Example 9.9.** The STFT calculation of a signal whose frequency changes linearly is done by using a rectangular window. Signal samples within \( 0 \leq n \leq M - 1 \) with \( M = 64 \) were available. The nonoverlapping STFT of this signal is calculated with a rectangular window of the width \( N = 8 \) and presented in Fig. 9.10. The nonoverlapping STFT values obtained by using the rectangular window are shifted in frequency, scaled, and added up, Fig. 9.11, to produce the STFT with a Hamming window, Fig. 9.12.

The STFT calculation for the same linear FM signal will be repeated for the overlapping STFT with step \( R = 1 \). Results for the rectangular and Hamming window (obtained by a simple matrix calculation from the rectangular
The STFT of a linear FM signal $x(n)$ calculates using a rectangular window of the width $N = 8$. The same procedure is repeated with the windows zero padded up to the widest used window (interpolation in frequency). The results are presented in Fig. 9.14. Note that regarding to the amount of information all these figures do not differ from the basic time-frequency representation presented in Fig. 9.10.
Figure 9.11 The STFT of a linear FM signal calculated using a rectangular window (from the previous figure), along with its frequency shifted versions \(STFT_R(n,k-1)\) and \(STFT_R(n,k)\). Their weighted sum produces the STFT of the same signal with a Hamming window \(STFT_H(n,k)\).

### 9.2.9 Signal Reconstruction from the Discrete STFT

Signal reconstruction from non-overlapping STFT values is obvious for a rectangular window. A simple illustration is presented in Fig.9.15. Windowed signal values are reconstructed from the STFTs by a simple inversion of each STFT

\[
STFT(n) = W_N H_w x(n)
\]

\[
H_w x(n) = \text{IDFT}\{STFT(n)\} = W_N^{-1} STFT(n)
\]

where \(H_w\) is a diagonal matrix with the window values as its elements, \(H_w = \text{diag}(w(m))\).

**Example 9.10.** Consider a signal with \(M = 16\) samples, \(x(0), x(1),..., x(16)\). Write a matrix form for the signal inversion using a four-sample STFT (\(N = 16\)) calculated with the rectangular and a Hann(ing) window: (a) Without overlapping, \(R = 16\). (b) With a time step in the STFT calculation of \(R = 2\).

\[\star\] (a) For the nonoverlapping case the STFT calculation is done according to:

\[
\text{STFT} = W_4 H_4
\]

\[
\begin{bmatrix}
  x(0) & x(4) & x(8) & x(12) \\
  x(1) & x(5) & x(9) & x(13) \\
  x(2) & x(6) & x(10) & x(14) \\
  x(3) & x(7) & x(11) & x(15)
\end{bmatrix}
\]

with \(H_4 = \text{diag}(\{w(-2) w(-1) w(0) w(1]\})\) and \(W_4\) is the corresponding four sample DFT matrix.
The inversion relation is

\[
\begin{bmatrix}
  x(0) & x(4) & x(8) & x(12) \\
  x(1) & x(5) & x(9) & x(13) \\
  x(2) & x(6) & x(10) & x(14) \\
  x(3) & x(7) & x(11) & x(15)
\end{bmatrix} = H_4^{-1} W_4^{-1} \text{STFT}
\]

where the elements of diagonal matrix \( H_4^{-1} \) are proportional to \( 1/w(m) \), \( H_4^{-1} = \text{diag}(1/w(-2) 1/w(-1) 1/w(0) 1/w(1)) \). If a rectangular window is used in the STFT calculation then \( H_4^{-1} = I_4 \) is unity matrix and this kind of
Figure 9.13  Time-frequency analysis of a linear frequency modulated signal with overlapping windows of various widths. Time step in the STFT calculation is $R = 1$. 
Figure 9.14  Time-frequency analysis of a linear frequency modulated signal with overlapping windows of various widths. Time step in the STFT calculation is $R = 1$. For each window width the frequency axis is interpolated (signal in time is zero padded) up to the total number of available signal samples $M = 64$. 
Figure 9.15  Illustration of the signal reconstruction from the STFT with nonoverlapping windows.
calculation can be used. However if a nonrectangular window is used then some of the window values are quite small. The signal value is then obtained by multiplying the inverse DFT with large values 1/w(m). This kind of division with small values is very imprecise, if any noise in the reconstructed signal is expected. In the Hann(ing) window case the ending point is even nonoverlapping windows. It would be after inversion, IDFT

\[ \text{STFT} = W_4H_4 \begin{bmatrix} 0 & x(0) & x(2) & x(4) & x(6) & x(8) & x(10) & x(12) & x(14) \\ 0 & x(1) & x(3) & x(5) & x(7) & x(9) & x(11) & x(13) & x(15) \\ x(0) & x(2) & x(4) & x(6) & x(8) & x(10) & x(12) & x(14) & 0 \\ x(1) & x(3) & x(5) & x(7) & x(9) & x(11) & x(13) & x(15) & 0 \end{bmatrix} \]

The inversion is

\[ W_4^{-1}\text{STFT} = H_4X = \begin{bmatrix} 0 & x(0)w(-2) & x(2)w(-2) & x(4)w(-2) & \ldots & x(14)w(-2) \\ 0 & x(1)w(-1) & x(3)w(-1) & x(5)w(-1) & \ldots & x(15)w(-1) \\ x(0)w(0) & x(2)w(0) & x(4)w(0) & x(6)w(0) & \ldots & 0 \\ x(1)w(1) & x(3)w(1) & x(5)w(1) & x(7)w(1) & \ldots & 0 \end{bmatrix} \]

where X is the matrix with signal elements. The window matrix is left on the right side, since in general it may be not invertible. By calculating \( W_4^{-1}\text{STFT} \) we can then recombine the signal values. For example, the element producing \( x(0)w(0) \) in the first column is combined with the element producing \( x(0)w(-2) \) in the second column to get \( x(0)w(0) + x(0)w(-2) = x(0) \), since for the Hann(ing) window of the width \( N \) holds \( w(n) + w(n - N/2) = 1 \). The same is done for other signal values in the matrix obtained after inversion,

\[
\begin{align*}
  x(0)w(0) + x(0)w(-2) &= x(0) \\
  x(1)w(1) + x(1)w(-1) &= x(1) \\
  x(2)w(0) + x(1)w(-2) &= x(2) \\
  \ldots \\
  x(15)w(1) + x(15)w(-1) &= x(15)
\end{align*}
\]

Note that the same relation would hold for a triangular window, while for a Hamming window a similar relation would hold, with \( w(n) + w(n - N/2) = 1.08 \). The results should be corrected in that case, by a constant factor of 1.08.

Illustration of the STFT calculation for an arbitrary window width \( N \) at \( n = n_0 \) is presented in Fig.9.16. Its inversion produces \( x(n_0 + m)w(m) = \text{IDFT}\{\text{STFT}_N(n_0,k)\} \). Consider the previous STFT value in the case of nonoverlapping windows. It would be \( \text{STFT}_N(n_0 - N,k) \). Its inverse

\[
\text{IDFT}\{\text{STFT}_N(n_0 - N,k)\} = x(n_0 - N + m)w(m)
\]
is also presented in Fig. 9.16. As it can be seen, by combining these two inverse transforms we would get signal with very low values around \( n = n_0 - N/2 \). If one more STFT is calculated at \( n = n_0 - N/2 \) and its inverse combined with previous two it will improve the signal presentation within the overlapping region \( n_0 - N \leq n < n_0 \). In addition for the most of common windows \( w(m - N) + w(m - N/2) + w(m) = 1 \) (or a constant) within \( 0 \leq m < N \) meaning that the sum of overlapped inverse STFTs, as in Fig. 9.16, will give the original signal within \( n_0 - N \leq n < n_0 \).

In general, let us consider the STFT calculation with overlapping windows. Assume that the STFTs are calculated with a step \( 1 \leq R \leq N \) in time. Available STFT values are

\[
\ldots \\
\text{STFT}(n_0 - 2R), \\
\text{STFT}(n_0 - R), \\
\text{STFT}(n_0), \\
\text{STFT}(n_0 + R), \\
\text{STFT}(n_0 + 2R), \\
\ldots
\]

(9.26)

Based on the available STFT values (9.26), the windowed signal values can be reconstructed as

\[
H_w x(n_0 + iR) = \mathbf{W}_N^{-1} \text{STFT}(n_0 + iR), \quad i = \ldots -2, -1, 0, 1, 2, \ldots
\]

For \( m = -N/2, -N/2 + 1, \ldots, N/2 - 1 \) we get signal values \( x(n_0 + iR + m) \)

\[
w(m)x(n_0 + iR + m) = \frac{1}{N} \sum_{k=-N/2}^{N/2-1} \text{STFT}(n_0 + iR, k)e^{j2\pi mk/N}. \quad (9.27)
\]

Since \( R < N \) we we will get the same signal value within different STFT, for different \( i \). For example, for \( N = 8, R = 2 \) and \( n_0 = 0 \) we will get the value \( x(0) \) for \( m = 0 \) and \( i = 0 \), but also for \( m = -2 \) and \( i = 1 \) or \( m = 2 \) and \( i = -1 \), and so on. Then in the reconstruction we should use all these values to get the most reliable reconstruction.

Let us reindex the reconstructed signal values (9.27) by substitution \( m = l - iR \)

\[
w(l - iR)x(n_0 + l) = \frac{1}{N} \sum_{k=-N/2}^{N/2-1} \text{STFT}(n_0 + iR, k)e^{j2\pi lk/N}e^{-j2\pi ik/N}
\]

\[-N/2 \leq l - iR \leq N/2 - 1.
\]
Figure 9.16 Illustration of the STFT calculation with windows overlapping in order to produce an inverse STFT whose sum will give the original signal within $n_0 - N \leq n < n_0$. 
If $R < N$ then a value of signal $x(n_0 + l)$ will be obtained by inverting

$$w(l)x(n_0 + l) = \frac{1}{N} \sum_{k=-N/2}^{N/2-1} \text{STFT}(n_0,k)e^{j2\pi lk/N}$$

but also it will be obtained within the inversions

$$w(l - 2R)x(n_0 + l) = \frac{1}{N} \sum_{k=-N/2}^{N/2-1} \text{STFT}(n_0 + 2R,k)e^{j2\pi lk/N}e^{-j2\pi 2Rk/N}$$

$$w(l - R)x(n_0 + l) = \frac{1}{N} \sum_{k=-N/2}^{N/2-1} \text{STFT}(n_0 + R,k)e^{j2\pi lk/N}e^{-j2\pi Rk/N}$$

$$w(l + R)x(n_0 + l) = \frac{1}{N} \sum_{k=-N/2}^{N/2-1} \text{STFT}(n_0 - R,k)e^{j2\pi lk/N}e^{j2\pi Rk/N}$$

$$w(l + 2R)x(n_0 + l) = \frac{1}{N} \sum_{k=-N/2}^{N/2-1} \text{STFT}(n_0 - 2R,k)e^{j2\pi lk/N}e^{j2\pi 2Rk/N}$$

as far as $w(l - 2iR)$, for $i = 0, \pm 1, \pm 2, \ldots$ is within

$$-N/2 \leq l - 2iR < N/2.$$ 

By summing all reconstructions over $i$ satisfying $-N/2 \leq l - iR \leq N/2 - 1$ we get the reconstructed signal $x(n_0 + l)$. It is undistorted (up to a constant) if

$$c(l) = \sum_i w(l - iR) = \text{const.} = C \quad \text{(9.28)}$$

since

$$\sum_i w(l - iR)x(n_0 + l) = Cx(n_0 + l)$$

for any $n_0$ and $l$. Note that $\sum_i w(l - iR)$ is a periodic extension of $w(l)$ with a period $R$. If $W(e^{j\omega})$ is the Fourier transform of $w(l)$ then the Fourier transform of its periodic extension is equal to the samples of $W(e^{j\omega})$ at $\omega = 2\pi k/R$. The condition (9.28) is equivalent to

$$W(e^{j2\pi k/R}) = CN\delta(k) \text{ for } k = 0, 1, \ldots, R - 1.$$ 

Special cases:
Figure 9.17 Signal reconstruction from the STFT for the case $N = 8$, when the STFT is calculated with step $R = N/2 = 4$ and the window satisfies $w(m) + w(m - N/2) = 1$. This is the case for the rectangular, Hann(ing), Blackman and triangular windows. The same holds for the Hamming window up to a constant scaling factor of 1.08.

1. For $R = N$ (nonoverlapping), relation (9.28) is satisfied for the rectangular window, only.

2. For a half of the overlapping period, $R = N/2$, condition (9.28) is met for the rectangular, Hann(ing), Hamming, and triangular window. Realization in this case for $N = 8$ and $R = N/2 = 4$ is presented in Fig.9.17. Signal values with a delay of $N/2 = 4$ samples are obtained at the exit. The STFT calculation process is repeated after each 4 samples, producing blocks of 4 signal samples at the output.

3. The same holds for $R = N/2, N/4, N/8$, if the values of $R$ are integers.

4. For $R = 1$, (the STFT calculation in each available time instant), any window satisfies the inversion relation. In this case we may also use a
simple reconstruction formula, Fig.9.18

\[
\frac{1}{N} \sum_{k=-N/2}^{N/2-1} STFT(n,k) = \frac{1}{N} \sum_{m=-N/2}^{N/2-1} \left( w(m)x(n + m) \sum_{k=-N/2}^{N/2-1} e^{-j2\pi mk/N} \right)
= w(0)x(n).
\]

Very efficient realizations, for this case, are the recursive ones, instead of the direct DFT calculation, Fig.9.8.

In analysis of non-stationary signals our primary interest is not in signal reconstruction with the fewest number of calculation points. Rather, we are interested in tracking signals’ non-stationary parameters, like for example, instantaneous frequency. These parameters may significantly vary between neighboring time instants \(n\) and \(n + 1\). Quasi-stationarity of signal within \(R\) samples (implicitly assumed when down-sampling by factor of \(R\) is done) in this case is not a good starting point for the analysis. Here, we have to use the time-frequency analysis of signal at each instant \(n\), without any down-sampling.

9.2.10 Time-Varying Windows

In general, varying window widths could be used for different time-frequency points. When \(N_i\) changes with \(n_i\) we have the case of a time-varying window. Assuming a rectangular window we can write,

\[
STFT_{N_i}(n_i,k) = \sum_{m=-N_i/2}^{N_i/2-1} x(n_i + m)e^{-j2\pi mk/N_i}
\tag{9.29}
\]

Notation \(STFT_{N_i}(n,k)\) means that the STFT is calculated using signal samples within the window \([n_i - N_i/2, n_i + N_i/2 - 1]\) for \(-N_i/2 \leq k \leq N_i/2 - 1\), corresponding to an even number of \(N_i\) discrete frequencies from \(-\pi\) to \(\pi\). For an odd \(N_i\), the summation limits are \(\pm(N_i - 1)/2\). Let us restate that a wide window includes signal samples over a wide time interval, losing the possibility to detect fast changes in time, but achieving high frequency resolution. A narrow window in the STFT will track time changes, but with a low resolution in frequency. Two extreme cases are \(N_i = 1\) when

\[
STFT_1(n,k) = x(n)
\]

and \(N_i = M\) when

\[
STFT_M(n,k) = X(k),
\]
where $M$ is the total number of all available signal samples and $X(k) = \text{DFT}\{x(n)\}$.

In vector notation

$$\text{STFT}_{N_i}(n_i) = W_{N_i}x_{N_i}(n_i),$$

where $\text{STFT}_{N_i}(n_i)$ and $x_{N_i}(n_i)$ are column vectors. Their elements are $\text{STFT}_{N_i}(n_i,k), k = -N_i/2,\ldots, N_i/2 - 1$ and $x(n_i + m), m = -N_i/2,\ldots, N_i/2 - 1$, respectively

$$\text{STFT}_{N_i}(n_i) = [\text{STFT}_{N_i}(n_i, -N_i/2) \ldots \text{STFT}_{N_i}(n_i, N_i/2 - 1)]^T$$

$$x_{N_i}(n_i) = [x(n_i - N_i/2) \ldots x(n_i + N_i/2 - 1)]^T.$$ 

Matrix $W_{N_i}$ is an $N_i \times N_i$ DFT matrix with elements

$$W_{N_i}(m,k) = \exp(-j2\pi mk/N_i),$$
where \( m \) is the column index and \( k \) is the row index of the matrix. The STFT value \( STFT_{N_i}(n_i, k) \) is presented as a block in the time-frequency plane of the width \( N_i \) in the time direction, covering all time instants \( [n_i - N_i/2, n_i + N_i/2 - 1] \) used in its calculation. The frequency axis can be labeled with the DFT indices \( p = -M/2, ..., M/2 - 1 \) corresponding to the DFT frequencies \( 2\pi p/M \) (dots in Fig.9.19). With respect to this axis labeling, the block \( STFT_{N_i}(n_i, k) \) will be positioned at the frequency \( 2\pi k/N_i = 2\pi (kM/N_i)/M \), i.e., at \( p = kM/N_i \). The block width in frequency is \( M/N_i \) DFT samples. Therefore the block area in time and DFT frequency is always equal to the number of all available signal samples \( M \) as shown in Fig.9.19 where \( M = 16 \).

**Example 9.11.** Consider a signal \( x(n) \) with \( M = 16 \) samples. Write the expression for calculation of the STFT value \( STFT_{4}(2,1) \) with a rectangular window. Indicate graphically the region of time instants used in the calculation and the frequency range in terms of the DFT frequency values included in the calculation of \( STFT_{4}(2,1) \)?

The STFT value \( STFT_{4}(2,1) \) is:

\[
STFT_{4}(2,1) = \sum_{m=-2}^{1} x(2 + m)e^{-j\frac{2\pi m}{4}}.
\]

It uses discrete-time samples of \( x(n) \) within

\[
-2 \leq 2 + m < 1
\]

\[
0 \leq n \leq 3.
\]

The frequency term is \( \exp(-j\frac{2\pi m}{4}) \). For the DFT of a signal with \( M = 16 \)

\[
X(k) = \sum_{n=0}^{15} x(n)e^{-j\frac{2\pi nk}{16}}
\]

\[
k = -8, -7, ..., -1, 0, 1, ..., 6, 7
\]

this frequency would correspond to the term \( \exp(-j\frac{2\pi 4m}{16}) \). Therefore \( k = 1 \) corresponds to the frequency \( p = 4 \) in the DFT. Since the whole frequency range \( -\pi \leq \omega < \pi \) in the case of \( N_i = 4 \) is covered with 4 STFT values \( STFT_{4}(2,-2), STFT_{4}(2,-1), STFT_{4}(2,0), \) and \( STFT_{4}(2,1) \) and the same frequency range in the DFT has 16 frequency samples, it means that each STFT value calculated with \( N_i = 4 \) corresponds to a range of frequencies corresponding to 4 DFT values,

\[
k = -2, \text{ corresponds to } p = -8, -7, -6, -5
\]

\[
k = -1, \text{ corresponds to } p = -4, -3, -2, -1
\]

\[
k = 0, \text{ corresponds to } p = 0, 1, 2, 3
\]

\[
k = 1, \text{ corresponds to } p = 4, 5, 6, 7.
\]
Figure 9.19 The nonoverlapping STFTs with: (a) constant window of the width $N = 4$, (b) constant window of the width $N = 2$, (c)-(d) time-varying windows. Time index is presented on the horizontal axis, while the DFT frequency index is shown on the vertical axis (the STFT is denoted by $S$ for notation simplicity).

This discrete-time and the DFT frequency region, $0 \leq n \leq 3$ and $4 \leq p \leq 7$, is represented by a square denoted by $S_4(2,1)$ in Fig. 9.19(a).

In a nonoverlapping STFT, covering all signal samples

$$x = [x(0), x(1), ..., x(M - 1)]^T$$

with $\text{STFT}_{N_i}(n_i)$, the STFT should be calculated at $n_0 = N_0/2$, $n_1 = N_0 + N_1/2$, $n_2 = N_0 + N_1 + N_2/2$, $n_3 = M - N_K/2$. A matrix form for all STFT
Time-Frequency Analysis

values is

\[
\text{STFT} = \begin{bmatrix}
    W_{N_0} & 0 & \cdots & 0 \\
    0 & W_{N_1} & \cdots & 0 \\
    \vdots & \vdots & \ddots & \vdots \\
    0 & 0 & \cdots & W_{N_K}
\end{bmatrix} x
\]

\[
\text{STFT} = ˜Wx = ˜WW^{-1}M X, \quad (9.30)
\]

where \(\text{STFT} \) is a column vector containing all STFT vectors \(\text{STFT}_{N_i}(n_i)\), \(i = 0, 1, \ldots, K\), \(X = W_Mx \) is a DFT of the whole signal \(x(n)\), while \(˜W\) is a block matrix \((M \times M)\) formed from the smaller DFT matrices \(W_{N_0}, W_{N_1}, \ldots, W_{N_K}\), as in (9.29). Since the time-varying nonoverlapping STFT corresponds to a decimation-in-time DFT scheme, its calculation is more efficient than the DFT calculation of the whole signal. Illustration of time-varying window STFTs is shown in Fig. 9.19(c), (d). For a signal with \(M\) samples, there is a large number of possible nonoverlapping STFTs with a time-varying window \(N_i \in \{1, 2, 3, \ldots, M\}\). The exact number will be derived later.

Example 9.12. Consider a signal \(x(n)\) with \(M = 16\) samples, whose values are

\[
x = [0.5, 0.5, -0.25, j0.25, 0.25, -j0.25, 0.25, -0.25, 0.25, 0.5, 0.5, -j0.5, j0.5, 0, -1].
\]

Some of its nonoverlapping STFTs are calculated according to (9.29) and shown in Fig. 9.19. Different representations can be compared based on the concentration measures, for example,

\[
\mu[\text{STFT}_{N_i}(n,k)] = \sum_{n_k} |\text{STFT}_{N_i}(n,k)| = \|\text{STFT}\|_1.
\]

The best STFT representation, in this sense, would be the one with the smallest \(\mu[\text{STFT}_{N_i}(n,k)]\). For the considered signal and its four representations shown in Fig. 9.19 the best representation, according to this criterion, is the one shown in Fig. 9.19(b).

Example 9.13. Consider a signal \(x(n)\) with \(M = 8\) samples. Its values are \(x(0) = 0, x(1) = 1, x(2) = 1/2, x(3) = -1/2, x(4) = 1/4, x(5) = -j/4, x(6) = -1/4, \) and \(x(7) = j/4\).

(a) Calculate the STFTs of this signal with rectangular window of the widths \(N = 1, N = 2, N = 4\). Use the following STFT definition

\[
\text{STFT}_N(n,k) = \sum_{m=-N/2}^{N/2-1} x(n+m)e^{-j2\pi nk/N}.
\]

For an odd \(N\), the summation limits are \(\pm (N - 1)/2\). Calculate \(\text{STFT}_1(n,k)\) for \(n = 0, 1, 2, 3, 4, 5, 6, 7\), then \(\text{STFT}_2(n,k)\) for \(n = 1, 3, 5, 7\), then \(\text{STFT}_4(n,k)\) for \(n = 2, 6 \) and \(\text{STFT}_8(n,k)\) for \(n = 4\). For frequency axis use notation \(k = 0, 1, 2, 3, 4, 5, 6, 7\).
(b) Assuming that time-varying approach is used in the nonoverlapping STFT calculation, find the total number of possible representations.

(c) Calculate the concentration measure for each of the cases in (b) and find the representation (nonoverlapping combination of previous STFTs) when the signal is represented with the smallest number of coefficients. Does it correspond to the minimum of $\mu[STFT(n,k)]$?

★(a) The STFT values are:

- for $N = 1$
  \[ STFT_1(n,0) = x(n), \text{ for all } n = 0,1,2,3,4,5,6,7; \]

- for $N = 2$
  \[ STFT_2(n,0) = x(n) + x(n - 1) \]
  \[ STFT_2(1,0) = 1, \]
  \[ STFT_2(3,0) = 0, \]
  \[ STFT_2(5,0) = (1 - j)/4, \]
  \[ STFT_2(7,0) = (-1 + j)/4 \]
  \[ STFT_2(n,1) = x(n) - x(n - 1) \]
  \[ STFT_2(1,1) = 1, \]
  \[ STFT_2(3,1) = -1, \]
  \[ STFT_2(5,1) = (-1 - j)/4, \]
  \[ STFT_2(7,1) = (1 + j)/4 \]

- for $N = 4$ and $n = 2, 6$
  \[ STFT_4(n,0) = x(n - 2) + x(n - 1) + x(n) + x(n + 1) \]
  \[ STFT_4(2,0) = 1, \]
  \[ STFT_4(6,0) = 0 \]
  \[ STFT_4(n,1) = -x(n - 2) + jx(n - 1) + x(n) - jx(n + 1) \]
  \[ STFT_4(2,1) = (1 + 3j)/2, \]
  \[ STFT_4(6,1) = 0 \]
  \[ STFT_4(n,2) = x(n - 2) - x(n - 1) + x(n) - x(n + 1) \]
  \[ STFT_4(2,2) = 0, \]
  \[ STFT_4(6,2) = 0, \]
  \[ STFT_4(n,3) = -x(n - 2) - jx(n - 1) + x(n) + jx(n + 1) \]
  \[ STFT_4(2,3) = (1 - 3j)/2, \]
  \[ STFT_4(6,3) = -1 \]
(b) Now we have to make all possible nonoverlapping combinations of these transforms and to calculate the concentration measure for each of them. Total number of combinations is 25. The absolute STFT values are shown in Fig. 9.20, along with measure

\[ \mu[STFT(n,k)] = \sum_n \sum_k |STFT(n,k)| \]

for each case. (c) By measuring the concentration for all of them, we will get

\[ M \]

The optimal representation, with respect to this measure, is presented with thicker gridlines. Time axis is \( n = 0,1,2,3,4,5,6,7 \) and the frequency axis is \( k = 0,1,2,3,4,5,6,7 \).
that the optimal combination, to cover the time-frequency plane, is

\[
\begin{align*}
STFT_1(0,0) &= x(0) = 0 \\
STFT_1(1,0) &= x(1) = 1 \\
STFT_2(3,1) &= x(3) - x(2) = -1 \\
STFT_2(3,0) &= x(3) + x(2) = 0 \\
STFT_4(6,0) &= x(4) + x(5) + x(6) + x(7) = 0 \\
STFT_4(6,1) &= -x(4) + jx(5) + x(6) - jx(7) = 0 \\
STFT_4(6,2) &= x(4) - x(5) + x(6) - x(7) = 0 \\
STFT_4(6,3) &= -x(4) - jx(5) + x(6) + jx(7) = -1
\end{align*}
\]

with just three nonzero transformation coefficients. It corresponds to the minimum of \( \mu[\text{SPEC}(n,k)] \).

In this case there is an algorithm for efficient optimal lattice determination, based on two regions consideration, starting from lattices 1, 19, and 25 from the Fig. 9.20, corresponding to the constant window widths of \( N = 1 \), \( N = 2 \), and \( N = 4 \) samples.

Example 9.14. Discrete signal \( x(n) \) for \( n = 0,1,2,3,4,5 \) is considered. Time-frequency plane is divided as presented in Fig. 9.21.

(a) Denote each region in the figure by appropriate coefficient \( STFT_{N_i}(n,k) \), where \( N \) is window length, \( n \) is the time index, and \( k \) is the frequency index.

(b) Write relations for coefficients calculation and write transformation matrix \( T \).

(c) By using the transformation matrix, find STFT values if signal samples are \( x(0) = 2, x(1) = -2, x(2) = 4, x(3) = \sqrt{3}, x(4) = -\sqrt{3}, x(5) = 0 \).
(d) If the STFT coefficients for signal \( y(n) \) are

\[
\begin{align*}
\text{STFT}_2(1,0) &= 4, & \text{STFT}_2(1,1) &= 0 \\
\text{STFT}_1(2,0) &= 1, & \text{STFT}_3(4,0) &= 0 \\
\text{STFT}_3(4,1) &= 3, & \text{STFT}_3(4,2) &= 3
\end{align*}
\]

find the signal samples \( y(n) \).

★(a) Denoted areas are presented in Fig. 9.22. (b) The STFT values are obtained using

\[
\begin{align*}
\text{STFT}_N(n,k) &= \sum_{m=-(N-1)/2}^{(N-1)/2-1} x(n+m)e^{-j2\pi mk/N} \quad \text{or} \\
\text{STFT}_N(n,k) &= \sum_{m=-N/2}^{N/2-1} x(n+m)e^{-j2\pi mk/N}
\end{align*}
\]

for and odd and even number of samples \( N \), respectively. It follows

\[
\begin{align*}
\text{STFT}_2(1,0) &= x(0) + x(1) \\
\text{STFT}_2(1,1) &= -x(0) + x(1) \\
\text{STFT}_1(2,0) &= x(2)
\end{align*}
\]
$STFT_3(4,0) = x(3) + x(4) + x(5)$
$STFT_3(4,1) = -\frac{1}{2} + j\sqrt{3} x(3) + x(4) + \frac{-1 + j\sqrt{3}}{2} x(5)$
$STFT_3(4,2) = -\frac{1}{2} - j\sqrt{3} x(3) + x(4) + \frac{-1 - j\sqrt{3}}{2} x(5)$.

The transformation matrix (where the STFT coefficients are arranged into column vector $S$) is

$$
T = \begin{bmatrix}
1 & 1 & 0 & 0 & 0 & 0 \\
-1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & -1 + j\sqrt{3} & 1 & -1 - j\sqrt{3} \\
0 & 0 & 0 & -1 - j\sqrt{3} & 1 & -1 + j\sqrt{3}
\end{bmatrix}.
$$

(c) The STFT coefficients are

$$
S = \begin{bmatrix}
1 & 1 & 0 & 0 & 0 & 0 \\
-1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & -1 + j\sqrt{3} & 1 & -1 - j\sqrt{3} \\
0 & 0 & 0 & -1 - j\sqrt{3} & 1 & -1 + j\sqrt{3}
\end{bmatrix} \begin{bmatrix}
2 \\
-2 \\
4 \\
\sqrt{3} \\
-\sqrt{3} \\
0
\end{bmatrix} = \begin{bmatrix}
0 \\
-4 \\
4 \\
0 \\
-3 + j\sqrt{3} \\
3 - j\sqrt{3}
\end{bmatrix}.
$$

(d) The signal samples $y(n)$ are obtained as $T^{-1}S$ resulting in

$$
\begin{bmatrix}
y(5) \\
y(4) \\
y(3) \\
y(2) \\
y(1) \\
y(0)
\end{bmatrix}^T = \begin{bmatrix}
2 \\
2 \\
1 \\
-1 \\
2 \\
-1
\end{bmatrix}^T.
$$

Example 9.15. A discrete signal $x(n)$ is considered for $0 \leq n < M$. Find the number of the STFTs of this signal with time-varying windows.

(a) Consider arbitrary window widths from 1 to $M$.

(b) Consider dyadic windows, that is, windows whose width is $2^m$, where $m$ is an integer, such that $2^m \leq M$. In this case find the number of time-varying window STFTs for $M = 1, 2, 3, \ldots, 15, 16$.

★(a) Let us analyze the problem recursively. Denote by $F(M)$ the number of STFTs for a signal with $M$ samples. It is obvious that $F(1) = 1$, that is, for one-sample signal there is only one STFT (signal sample itself). If $M > 1$, we can use window with widths $k = 1, 2, \ldots, M$, as the first analysis window. Now let us analyze remaining $(M-k)$ samples in all possible ways, so we can write a recursive relation for the total number of the STFTs. If
the first window is one-sample window, then the number of the STFTs is \( F(M - 1) \). When the first window is a two-sample window, then the total number of the STFTs is \( F(M - 2) \), and so on, until the first window is the \( M \)-sample window, when \( F(M - M) = 1 \). Thus, the total number of the STFTs for all cases is

\[
F(M) = F(M - 1) + F(M - 2) + \ldots + F(1) + 1
\]

We can introduce \( F(0) = 1 \) (meaning that if there are no signal samples we have only one way to calculate time-varying window STFT) and obtain

\[
F(M) = F(M - 1) + F(M - 2) + \ldots + F(1) + F(0) = \sum_{k=1}^{M} F(M - k)
\]

Now, for \( M > 1 \) we can write

\[
F(M - 1) = \sum_{k=1}^{M-1} F(M - 1 - k) = \sum_{k=2}^{M} F(M - k)
\]

and

\[
F(M) - F(M - 1) = \sum_{k=1}^{M} F(M - k) - \sum_{k=2}^{M} F(M - k) = F(M - 1)
\]

\[
F(M) = 2F(M - 1).
\]

resulting in \( F(M) = 2^{M-1} \).

(b) In a similar way, following the previous analysis, we can write

\[
F(M) = F(M - 2^0) + F(M - 2^1) + F(M - 2^2) + \ldots + F(M - 2^m)
\]

\[
= \sum_{m=0}^{\lfloor \log_2 M \rfloor} F(M - 2^m)
\]

where \( \lfloor \log_2 M \rfloor \) is an integer part of \( \log_2 M \). Here we cannot write a simple recurrent relation as in the previous case. It is obvious that \( F(1) = 1 \). We can also assume that \( F(0) = 1 \). By unfolding recurrence we will get

\[
F(2) = F(1) + F(0) = 2
\]

\[
F(3) = F(2) + F(1) = 3
\]

\[
F(4) = F(3) + F(2) + F(0) = 6
\]

\[
\vdots
\]

The results are presented in the table

<table>
<thead>
<tr>
<th>( M )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F(M) )</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>10</td>
<td>18</td>
<td>31</td>
<td>56</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( M )</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F(M) )</td>
<td>98</td>
<td>174</td>
<td>306</td>
<td>542</td>
<td>956</td>
<td>1690</td>
<td>2983</td>
<td>5272</td>
</tr>
</tbody>
</table>
Note that the approximative formula
\[ F(M) \approx \left[ 1.0366 \cdot (1.7664)^{M-1} \right] \]
where \([\cdot]\) is an integer part of the argument, holds, with relative error smaller then 0.4% for \(1 \leq M \leq 1024\). For example, for \(M = 16\) we have 5272 different ways to split time-frequency plane into non-overlapping time-frequency regions.

### 9.2.11 Frequency-Varying Window

The STFT may use frequency-varying window as well. For a given DFT frequency \(p_i\) the window width in time is constant, Fig.9.23

\[
STFT_{N_i}(n,k_i) = \sum_{m=-N_i/2}^{N_i/2-1} w(m)x(n + m)e^{-j \frac{2\pi}{N_i} mk_i}.
\]

For example, value of \(STFT_{4}(2,-1)\) is

\[
STFT_{4}(2,-1) = \sum_{m=-2}^{2-1} x(2 + m)e^{-j2\pi m(-1)/4}.
\]

It position in the time-frequency plane is shown in 9.23(left).

For the signal used to illustrate the frequency-varying STFT in 9.23, the best concentration (out of the presented four) is the one shown in the last subplot. Optimization can be done in the same way as in the case of time-varying windows.

The STFT can be calculated by using the signal’s DFT instead of the signal. There is a direct relation between the time and the frequency domain STFT via coefficients of the form \(\exp(j2\pi nk/M)\). A dual form of the STFT is:

\[
\begin{align*}
STFT(n,k) &= \frac{1}{M} \sum_{i=0}^{M-1} P(i)X(k + i)e^{j2\pi in/M}, \\
STFT_M(k) &= W^{-1}_M P^{-1}_M X(k).
\end{align*}
\]

Frequency domain window \(P(i)\) may be of frequency varying width. This form is dual to the time-varying form. Forms corresponding to frequency varying windows, dual to the ones for the time-varying windows, can be
easily defined, for example, for a rectangular frequency domain window, as

$$\text{STFT} = \begin{bmatrix} W_{N_0}^{-1} & 0 & \cdots & 0 \\ 0 & W_{N_1}^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W_{N_K}^{-1} \end{bmatrix} X, \quad (9.32)$$

where $X = [X(0), X(1), \ldots, X(M - 1)]^T$ is the DFT vector.
9.2.12 Hybrid Time-Frequency-Varying Windows

In general, spectral content of signal changes in time and frequency in an arbitrary manner. Combining time-varying and frequency-varying windows we get hybrid time–frequency-varying windows with \( STFT_{N(i,j)}(n_i, k_l) \),

\[
STFT_{N(i,j)}(n_i, k_l) = \sum_{m=-N_{(i,j)}/2}^{N_{(i,j)}/2-1} w_{(i,j)}(m) x(n_i + m) e^{-j \frac{2\pi}{N_{(i,j)}} mk_l}
\]  

(9.33)

For a graphical representation of the STFT with varying windows, the corresponding STFT value should be assigned to each instant \( n = 0, 1, \ldots, M - 1 \) and each DFT frequency \( p = -M/2, -M/2 + 1, \ldots, M/2 - 1 \) within a block. In the case of a hybrid time–frequency-varying window the matrix form is obtained from the definition for each STFT value. For example, for the STFT calculated as in Fig. 9.24, for each STFT value an expression based on (9.33) should be written. Then the resulting matrix \( STFT \) can be formed.

There are several methods in the literature that adapt windows or basis functions to the signal form for each time instant or even for every considered time and frequency point in the time-frequency plane. Selection of the most appropriate form of the basis functions (windows) for each time–frequency point includes a criterion for selecting the optimal window width (basis function scale) for each point.

9.3 WAVELET TRANSFORM

The first form of functions having the basic property of wavelets was used by Haar at the beginning of the twentieth century. At the beginning of 1980’s, Morlet introduced a form of basis functions for analysis of seismic signals, naming them “wavelets”. Theory of wavelets was linked to the image processing by Mallat in the following years. In late 1980s Daubechies presented a whole new class of wavelets that can be implemented in a simple way, by using digital filtering ideas. The most important applications of the wavelets are found in image processing and compression, pattern recognition and signal denoising. Here, we will only link the basics of the wavelet transform to the time-frequency analysis.

Common STFT is characterized by a constant window and constant time and frequency resolutions for both low and high frequencies. The basic idea behind the wavelet transform, as it was originally introduced by Morlet, was to vary the resolution with scale (being related to frequency)
in such a way that a high frequency resolution is obtained for signal components at low frequencies, whereas a high time resolution is obtained for signal at high frequency components. This kind of resolution change could be relevant for some practical applications, like for example seismic signals. It is achieved by introducing a frequency variable window width. Window width is decreased as frequency increases.

The basis functions in the STFT are

\[
STFT_{II}(t, \Omega_0) = \int_{-\infty}^{\infty} x(\tau) w(\tau - t) e^{-j\Omega_0 \tau} d\tau
\]

\[
= \langle x(\tau), w(\tau - t) e^{-j\Omega_0 \tau} \rangle = \langle x(\tau), h^*(\tau - t) \rangle = \int_{-\infty}^{\infty} x(\tau) h^*(\tau - t) d\tau
\]

where \( h(\tau - t) = w(\tau - t) e^{j\Omega_0 \tau} \) is a band-pass signal. It is obtained when a real-valued window \( w(\tau - t) \) is modulated by \( e^{j\Omega_0 \tau} \).
When the above idea about wavelet transform is translated into the mathematical form and related to the STFT, one gets the definition of a continuous wavelet transform

$$WT(t,a) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(\tau) h^* \left( \frac{\tau - t}{a} \right) d\tau$$ (9.34)

where $h(t)$ is a band-pass signal, and the parameter $a$ is the scale. This transform produces a time-scale, rather than the time-frequency signal representation. For the Morlet wavelet the relation between the scale and the frequency is $a = \Omega_0 / \Omega$. In order to establish a strong formal relationship between the wavelet transform and the STFT, we will choose the basic Morlet wavelet $h(t)$ in the form

$$h(t) = w(t)e^{j\Omega_0 t}$$ (9.35)

where $w(t)$ is a window function and $\Omega_0$ is a constant frequency. For the Morlet wavelet we have a modulated Gaussian function

$$h(t) = \sqrt{\frac{1}{2\pi}} e^{-\alpha t^2} e^{j\Omega_0 t}$$

where the values of $\alpha$ and $\Omega_0$ are chosen such that the ratio of $h(0)$ and the first maximum is $1/2$, $\Omega_0 = 2\pi \sqrt{\alpha / \ln 2}$. From the definition of $h(t)$ it is obvious that small $\Omega$ (i.e., large $a$) corresponds to a wide wavelet, i.e., a wide window, and vice versa. The basic idea of the wavelet transform and its comparison with the STFT is illustrated in Fig. 9.25.

Substitution of (9.35) into (9.34) leads to a continuous wavelet transform form suitable for a direct comparison with the STFT:

$$WT(t,a) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(\tau) w^* \left( \frac{\tau - t}{a} \right) e^{-j\Omega_0 \frac{\tau - t}{a}} d\tau.$$ (9.36)

From the filter theory point of view the wavelet transform, for a given scale $a$, could be considered as the output of system with impulse response $h^*(-t/a)\sqrt{|a|}$, i.e.,

$$WT(t,a) = x(t) *_t h^*(-t/a) \sqrt{|a|},$$

where $*_t$ denotes a convolution in time. Similarly the STFT, for a given $\Omega$, may be considered as $STFTII(t,\Omega) = x(t) *_t [w^*(-t)e^{j\Omega t}]$. If we consider
these two band-pass filters from the bandwidth point of view we can see that, in the case of STFT, the filtering is done by a system whose impulse response $w^*(-t)e^{j\Omega t}$ has a constant bandwidth, being equal to the width of the Fourier transform of $w(t)$.

**Constant Q-Factor Transform:** The quality factor $Q$ for a band-pass filter, as measure of the filter selectivity, is defined as

$$Q = \frac{\text{Central Frequency}}{\text{Bandwidth}}$$

In the STFT the bandwidth is constant, equal to the window Fourier transform width, $B_w$. Thus, factor $Q$ is proportional to the considered frequency,

$$Q = \frac{\Omega}{B_w}.$$ 

In the case of the wavelet transform the bandwidth of impulse response is the width of the Fourier transform of $w(t/a)$. It is equal to $B_0/a$, where $B_0$ is the constant bandwidth corresponding to the mother wavelet (wavelet in
Figure 9.26  Illustration of the wavelet transform (a) of a sum of two delta pulses and two sinusiods compared with STFT (b)

scale \( a = 1 \). It follows

\[
Q = \frac{\Omega}{B_0/a} = \frac{\Omega_0}{B_0} = \text{const.}
\]

Therefore, the continuous wavelet transform corresponds to the passing a signal through a series of band-pass filters centered at \( \Omega \), with constant factor \( Q \). Again we can conclude that the filtering, that produces Wavelet transform, results in a small bandwidth (high frequency resolution and low time resolution) at low frequencies and wide bandwidth (low frequency and high time resolution) at high frequencies.

**Example 9.16.** Find the wavelet transform of signal (9.3)

\[
x(t) = \delta(t - t_1) + \delta(t - t_2) + e^{j\Omega_1 t} + e^{j\Omega_2 t}.
\]

\[ (9.37) \]

\[ \star \]

Its continuous wavelet transform is

\[
WT(t,a) = \frac{1}{\sqrt{|a|}} \left[ w((t_1 - t)/a)e^{-j\Omega_0(t_1 - t)/a} + w((t_2 - t)/a)e^{-j\Omega_0(t_2 - t)/a} \right]
+ \sqrt{|a|} \left[ e^{j\Omega_1 t} W[a(\Omega_0/a - \Omega_1)] + e^{j\Omega_2 t} W[a(\Omega_0/a - \Omega_2)] \right].
\]

\[ (9.38) \]

where \( w(t) \) is a real-valued function. The transform (9.38) has nonzero values in the region depicted in Fig. 9.26(a).

In analogy with spectrogram, the scalogram is defined as the squared magnitude of a wavelet transform:

\[
\text{SCAL}(t,a) = |WT(t,a)|^2.
\]

\[ (9.39) \]
The scalogram obviously loses the linearity property, and fits into the category of quadratic transforms.

9.3.1 Filter Bank and Discrete Wavelet

This analysis will start by splitting the signal’s spectral content into its high frequency and low frequency part. Within the STFT framework, this can be achieved by a two sample rectangular window

\[ w(n) = \delta(n) + \delta(n + 1), \]

with \( N = 2 \). A two-sample window STFT is

\[
STFT(n,0) = \frac{1}{\sqrt{2}} \sum_{m=0}^{1} x(n + m)e^{-j0} \\
= \frac{1}{\sqrt{2}} (x(n) + x(n + 1)) = x_L(n),
\]

for \( k = 0 \), corresponding to low frequency \( \omega = 0 \) and

\[
x_H(n) = \frac{1}{\sqrt{2}} (x(n) - x(n + 1))
\]

for \( k = 1 \) corresponding to high frequency \( \omega = \pi \). A time-shifted (anticausal) version of the STFT

\[
STFT(n,k) = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} x(n + m)e^{-j2\pi km/N}
\]

is used, instead of \( STFT(n,k) = \sum_{m=-N/2}^{N/2-1} x(n + m)e^{-j2\pi km/N} \) in order to remain within the common wavelet literature notation. For the same reason the STFT is scaled by \( \sqrt{N} \) (a form when the DFT and IDFT have the same factor \( 1/\sqrt{N} \)).

This kind of signal analysis leads to the Haar (wavelet) transform. In the Haar wavelet transform the high-frequency part, \( x_H(n) \) is not processed any more. It is kept with this (high) two-samples resolution in time. The resolution in time of \( x_H(n,1) \) is just slightly (two-times) lower than the original signal sampling interval. The lowpass part \( x_L(n) = (x(n) + x(n + 1)) / \sqrt{2} \) will be further processed. After the signal samples \( x(n) \) and \( x(n + 1) \) are processed using (9.40) and (9.41), then next two samples \( x(n + 2) \) and \( x(n + 3) \) are analyzed. The highpass part is again calculated \( x_H(n + 2) = \)
(x(n + 2) – x(n + 3)) / √2 and kept as it is. Lowpass part \( x_L(n + 2) = (x(n + 2) + x(n + 3)) / √2 \) is considered as a new signal, along with its corresponding previous sample \( x_L(n) \).

Spectral content of the lowpass part of signal is divided, in the same way, into its low and high frequency part,

\[
x_{LL}(n) = \frac{1}{\sqrt{2}}(x_L(n) + x_L(n + 2)) = \frac{1}{2}[x(n) + x(n + 1) + x(n + 2) + x(n + 3)]
\]

\[
x_{LH}(n) = \frac{1}{\sqrt{2}}(x_L(n) - x_L(n + 2)) = \frac{1}{2}[x(n) + x(n + 1) - [x(n + 2) + x(n + 3)]].
\]

The highpass part \( x_{LH}(n) \) is left with resolution four in time, while the lowpass part is further processed in the same way, by dividing spectral content of \( x_{LL}(n) \) and \( x_{LL}(n + 4) \) into its low and high frequency part. This process is continued until the full length of signal is achieved. The Haar wavelet transformation matrix in the case of signal with 8 samples is

\[
\begin{bmatrix}
\sqrt{2}W_1(0,H) \\
\sqrt{2}W_1(2,H) \\
\sqrt{2}W_1(4,H) \\
\sqrt{2}W_1(6,H) \\
2W_2(0,H) \\
2W_2(4,H) \\
2\sqrt{2}W_4(0,H) \\
2\sqrt{2}W_4(0,L)
\end{bmatrix}
= \begin{bmatrix}
-1 1 0 0 0 0 0 0 \\
0 0 -1 -1 0 0 0 0 \\
0 0 0 1 -1 0 0 0 \\
0 0 0 0 0 1 -1 0 \\
1 1 -1 -1 0 0 0 0 \\
0 0 0 0 1 1 1 -1 \\
1 1 1 1 -1 -1 -1 -1 \\
1 1 1 1 1 1 1 1
\end{bmatrix} \begin{bmatrix}
x(0) \\
x(1) \\
x(2) \\
x(3) \\
x(4) \\
x(5) \\
x(6) \\
x(7)
\end{bmatrix}.
\]

This kind of signal transformation was introduced by Haar more than a century ago. In this notation scale \( a = 1 \) values of the wavelet coefficients \( W_1(2n,H) \) are equal to the highpass part of signal calculated using two samples, \( W_1(2n,H) = x_H(2n) \). The scale \( a = 2 \) wavelet coefficients are \( W_2(4n,H) = x_{LH}(4n) \). In scale \( a = 4 \) there is only one highpass and one lowpass coefficient at \( n = 0 \), \( W_4(8n,H) = x_{LLH}(8n) \) and \( W_4(8n,L) = x_{LLL}(8n) \). In this way any length of signal \( N = 2^m \) can be decomposed into Haar wavelet coefficients.

The Haar wavelet transform has a property that its highpass coefficients are equal to zero if the analyzed signal is constant within the analyzed time interval, for considered scale. If signal has large number of constant
value samples within the analyzed time intervals, then many Haar wavelet transform coefficients are zero valued. They can be omitted in signal storage or transmission. In recovery their values are assumed as zeros and the original signal is obtained. The same can be done in the case of noisy signals, when all coefficients below an assumed level of noise can be zero-valued and the signal-to-noise ratio in the reconstructed signal improved.

9.3.1.1 Lowpass and Highpass Filtering and Downsampling

Although the presented Haar wavelet analysis is quite simple we will use it as an example to introduce the filter bank framework of the wavelet transform. Obvious results from the Haar wavelet will be used to introduce other wavelet forms. For the Haar wavelet calculation two signals \( x_L(n) \) and \( x_H(n) \) are formed according to (9.40) and (9.41), based on the input signal \( x(n) \). Transfer functions of the discrete-time systems producing these two signals are

\[
H_L(z) = \frac{1}{\sqrt{2}} (1 + z) \tag{9.43}
\]

\[
H_H(z) = \frac{1}{\sqrt{2}} (1 - z).
\]

Frequency responses of these systems assume the form

\[
H_L(e^{j\omega}) = \frac{1}{\sqrt{2}} (1 + e^{j\omega})
\]

\[
H_H(e^{j\omega}) = \frac{1}{\sqrt{2}} (1 - e^{j\omega})
\]

with amplitude characteristics \( |H_L(e^{j\omega})| = \sqrt{2} |\cos(\omega/2)| \), and \( |H_H(e^{j\omega})| = \sqrt{2} |\sin(\omega/2)| \), presented in Fig.9.27. As expected, they represent a quite rough forms of lowpass and highpass filters. In general, this principle is kept for all wavelet transforms. The basic goal for all of them is to split the frequency content of a signal into its lowpass part and highpass part providing, in addition, a possibility of simple and efficient signal reconstruction.

After the values representing lowpass and highpass part of signal are obtained, next values of the signals \( x_L(n) = [x(n) + x(n+1)] / \sqrt{2} \) and \( x_H(n) = [x(n) - x(n+1)] / \sqrt{2} \) are calculated after one time instant is skipped. Therefore the output signal is downsampled by factor of two. The
new downsampled signals will be denoted by

\[ s_L(n) = x_L(2n) \]
\[ s_H(n) = x_H(2n). \]

(9.44)

Downsampling of a signal \( x(n) \) to get the signal \( y(n) = x(2n) \) is described in the z-transform domain by the function

\[ Y(z) = \frac{1}{2} X(z^{1/2}) + \frac{1}{2} X(-z^{1/2}). \]

This relation can easily be verified using the z-transform definition

\[ X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n} \]

\[ X(z^{1/2}) + X(-z^{1/2}) = \sum_{n=-\infty}^{\infty} x(n)(z^{-1/2})^n + (-z^{-1/2})^n = \sum_{n=-\infty}^{\infty} 2x(2n)z^{-n} \]

\[ \mathcal{Z}\{x(2n)\} = Y(z) = \frac{1}{2} X(z^{1/2}) + \frac{1}{2} X(-z^{1/2}). \]  

(9.45)

For the signals \( s_L(n) = x_L(2n) \) and \( s_H(n) = x_H(2n) \) the system implementation is presented in Fig. 9.28.
If the signals $s_L(n)$ and $s_H(n)$ are passed through the lowpass and highpass filters $H_L(z)$ and $H_H(z)$ and then downsampled,

$$S_L(z) = \frac{1}{2} H_L(z^{1/2}) X(z^{1/2}) + \frac{1}{2} H_L(-z^{1/2}) X(-z^{1/2})$$

$$S_H(z) = \frac{1}{2} H_H(z^{1/2}) X(z^{1/2}) + \frac{1}{2} H_H(-z^{1/2}) X(-z^{1/2})$$

hold.

9.3.1.2 Upsampling

Let us assume that we are not going to transform the signals $s_L(n)$ and $s_H(n)$ any more. The only goal is to reconstruct the signal $x(n)$ based on its downsampled lowpass and highpass part signals $s_L(n)$ and $s_H(n)$. The first step in the signal reconstruction is to restore the original sampling interval of the discrete-time signal. It is done by upsampling the signals $s_L(n)$ and $s_H(n)$.

Upsampling of a signal $x(n)$ is described by

$$y(n) = [...x(-2), 0, x(-1), 0, x(0), 0, x(1), 0, x(2), 0, ...]$$

Its $z$-transform domain form is

$$Y(z) = X(z^2),$$
since

\[ X(z^2) = \sum_{n=-\infty}^{\infty} x(n)z^{-2n} = \ldots x(-1)z^2 + 0 \cdot z^1 + x(0) + 0 \cdot z^{-1} + x(1)z^{-2} + \ldots \] (9.46)

Upsampling of a signal \( x(n) \) is defined by

\[ y(n) = \begin{cases} x(n/2) & \text{for even } n \\ 0 & \text{for odd } n \end{cases} = \mathcal{Z}^{-1}\{X(z^2)\}. \]

If a signal \( x(n) \) is downsampled first and then upsampled, the resulting signal transform is

\[
\begin{align*}
Y(z) &= \frac{1}{2}X\left(\left(z^{1/2}\right)^2\right) + \frac{1}{2}X\left(-\left(z^{1/2}\right)^2\right) \\
Y(z) &= \frac{1}{2}X(z) + \frac{1}{2}X(-z).
\end{align*}
\] (9.47)

In the Fourier domain it means \( Y(e^{j\omega}) = (X(e^{j\omega}) + X(e^{j(\omega+\pi)}) \). This form indicates that an aliasing component \( X(e^{j(\omega+\pi)}) \) appeared in this process.

### 9.3.1.3 Reconstruction Condition

In general, when the signal is downsampled and upsampled the aliasing appears since the component \( X(-z) \) exists in addition to the original signal \( X(z) \) in (9.47). The upsampled versions of signals \( s_L(n) \) and \( s_H(n) \) should be appropriately filtered and combined in order to eliminate aliasing. The conditions to avoid the aliasing in the reconstructed signal will be studied next.

In the reconstruction process the signals are upsampled \((S_L(z) \rightarrow S_L(z^2) \) and \( S_H(z) \rightarrow S_H(z^2)\)) and passed through the reconstruction filters \( G_L(z) \) and \( G_L(z) \) before being added up to form the output signal, Fig.9.29.
Figure 9.29  One stage of the filter bank with reconstruction, corresponding to the one stage of the wavelet transform realization.

The output signal transforms are

\[
Y_L(z) = S_L(z^2)G_L(z) = \left[ \frac{1}{2}H_L(z)X(z) + \frac{1}{2}H_L(-z)X(-z) \right]G_L(z)
\]

\[
Y_H(z) = S_H(z^2)G_H(z) = \left[ \frac{1}{2}H_H(z)X(z) + \frac{1}{2}H_H(-z)X(-z) \right]G_H(z)
\]

\[
Y(z) = Y_L(z) + Y_H(z)
\]

\[
= \left[ \frac{1}{2}H_L(z)G_L(z) + \frac{1}{2}H_H(z)G_H(z) \right]X(z)
\]

\[
+ \left[ \frac{1}{2}H_L(-z)G_L(z) + \frac{1}{2}H_H(-z)G_H(z) \right]X(-z).
\]

Condition for alias-free reconstruction is

\[
Y(z) = X(z).
\]

It means that

\[
H_L(z)G_L(z) + H_H(z)G_H(z) = 2 \quad (9.48)
\]

\[
H_L(-z)G_L(z) + H_H(-z)G_H(z) = 0. \quad (9.49)
\]

These are general conditions for a correct (alias-free) signal reconstruction.
Based on the reconstruction conditions we can show that the lowpass filters satisfy

\[ H_L(z)G_L(z) + H_L(-z)G_L(-z) = 2 \]  
\[ P(z) + P(-z) = 2, \]  
where \( P(z) = H_L(z)G_L(z) \).

From (9.49) we may write

\[ G_H(z) = \frac{H_L(-z)G_L(z)}{H_H(-z)} \]  
\[ H_H(z) = \frac{H_L(z)G_L(-z)}{G_H(-z)}. \]

Second expression is obtained from (9.49) with \( z \) being replaced by \( -z \), when \( H_L(z)G_L(-z) + H_H(z)G_H(-z) = 0 \). Substituting these values into (9.48) we get

\[ H_L(z)G_L(z) + \frac{H_L(-z)G_L(z)}{H_H(-z)} \frac{H_L(z)G_L(-z)}{G_H(-z)} = 2 \]

or

\[ \frac{H_L(z)G_L(z)}{H_H(-z)G_H(-z)} [H_H(-z)G_H(-z) + H_L(-z)G_L(-z)] = 2. \]

Since the expression within the brackets is equal to 2 (reconstruction condition (9.48) with \( z \) being replaced by \( -z \)) then

\[ \frac{H_L(z)G_L(z)}{H_H(-z)G_H(-z)} = 1 \]

and (9.50) follows with

\[ H_H(z)G_H(z) = H_L(-z)G_L(-z). \]

In the Fourier transform domain the reconstruction conditions are

\[ H_L(e^{j\omega})G_L(e^{j\omega}) + H_H(e^{j\omega})G_H(e^{j\omega}) = 2 \]  
\[ H_L(-e^{j\omega})G_L(e^{j\omega}) + H_H(-e^{j\omega})G_H(e^{j\omega}) = 0. \]
9.3.1.4 Orthogonality Conditions

The wavelet transform is calculated using downsampling by a factor 2. One of the basic requirements that will be imposed to the filter impulse response for an efficient signal reconstruction is that it is orthogonal to its shifted version with step 2 (and its multiples). In addition the wavelet functions in different scales should be orthogonal. Orthogonality of wavelet function in different scales will be discussed later. The orthogonality condition for the impulse response is

\[ \langle h_L(m), h_L(m - 2n) \rangle = \delta(n) \]

\[ \sum_m h_L(m) h_L(m - 2n) = \delta(n). \]

For the Haar wavelet transform this condition is obviously satisfied. In general, for wavelet transforms when the duration of impulse response \( h_L(n) \) is greater than two, the previous relation can be understood as a downsamped convolution of \( h_L(n) \) and \( h_L(-n) \)

\[ r(n) = h_L(n) * h_L(-n) = \sum_m h_L(m) h_L(m - n), \]

\[ Z\{r(n)\} = H_L(z)H_L(z^{-1}) \]

\[ \text{FT}\{r(n)\} = \left| H_L(e^{j\omega}) \right|^2. \]

The Fourier transform of the downsampled convolution, for real-valued \( h_L(n) \) is, (9.45)

\[ \text{FT}\{r(2n)\} = \frac{1}{2} \left| H_L(e^{j\omega/2}) \right|^2 + \frac{1}{2} \left| H_L(-e^{j\omega/2}) \right|^2. \]

From \( r(2n) = \delta(n) \) follows

\[ \left| H_L(e^{j\omega}) \right|^2 + \left| H_L(-e^{j\omega}) \right|^2 = 2. \]

The impulse response is orthogonal, in the sense of (9.54), if the frequency response satisfies

\[ \left| H_L(e^{j\omega}) \right|^2 + \left| H_L(e^{j(\omega + \pi)}) \right|^2 = 2. \]
Time domain form of relation (9.50) is

\[ h_L(n) \ast g_L(n) + [(-1)^n h_L(n)] \ast [(-1)^n g_L(n)] = 2 \delta(n) \]
\[ \sum_m h_L(m) g_L(n - m) + \sum_m (-1)^n h_L(m) g_L(n - m) = 2 \delta(n) \]
\[ \sum_m h_L(m) g_L(2n - m) = \delta(n). \]

If the impulse response \( h_L(n) \) is orthogonal, as in (9.54), then the last relation is satisfied for

\[ g_L(n) = h_L(-n). \]

In the z-domain it holds

\[ G_L(z) = H_L(z^{-1}) \]

and we may write (9.48) in the form

\[ G_L(z) G_L(z^{-1}) + G_L(-z) G_L(-z^{-1}) = 2 \]  \hspace{1cm} (9.55)

or \( P(z) + P(-z) = 2 \) with \( P(z) = G_L(z) G_L(z^{-1}) \). Relation (9.48) may also written for \( H_L(z) \) as well

\[ H_L(z) H_L(z^{-1}) + H_L(-z) H_L(-z^{-1}) = 2. \]

### 9.3.1.5 FIR Filter and Orthogonality Condition

Consider a lowpass anticausal FIR filter of the form

\[ h_L(n) = \sum_{k=0}^{K-1} h_k \delta(n + k) \]

and the corresponding causal reconstruction filter

\[ g_L(n) = h_L(-n) = \sum_{k=0}^{K-1} h_k \delta(n - k) \]

\[ G_L(e^{j\omega}) = H_L(e^{-j\omega}) \]
If the highpass filters are obtained from corresponding lowpass filters by reversal, in addition to common multiplication by \((-1)^n\), then

\[ g_H(n) = (-1)^n g_L(K - n) \]

\[ G_H(e^{j\omega}) = \sum_{n=0}^{K} g_H(n) e^{-j\omega n} = \sum_{n=0}^{K} (-1)^n g_L(K - n) e^{-j\omega n} \]

\[ = \sum_{m=0}^{K} (-1)^{K-m} g_L(m) e^{-j\omega(K-m)} = (-1)^K e^{-j\omega K} \sum_{m=0}^{K} e^{j\pi m} g_L(m) e^{-j(-\omega)m} \]

\[ = -e^{-j\omega K} G_L(e^{-j(\omega - \pi)}) = -e^{-j\omega K} G_L(-e^{-j\omega}) \]

or

\[ G_H(e^{j\omega}) = -e^{-j\omega K} G_L(-e^{-j\omega}) = -e^{-j\omega K} H_L(-e^{j\omega}) \]

for \( G_L(e^{j\omega}) = H_L(e^{-j\omega}) \). Similar relation holds for the anticausal \( h_H(n) \) impulse response

\[ h_H(n) = (-1)^n h_L(-K - n). \]

\[ H_H(e^{j\omega}) = \sum_{n=-K}^{0} h_H(n) e^{-j\omega n} = \sum_{n=-K}^{0} (-1)^n h_L(-n - K) e^{-j\omega n} \]

\[ = \sum_{m=-K}^{0} (-1)^{-K-m} h_L(m) e^{j\omega(m+K)} = -e^{j\omega K} H_L(-e^{-j\omega}) \]

The reconstruction conditions are satisfied since, according to (9.48) and (9.52), a relation corresponding to

\[ H_H(z)G_H(z) = H_L(-z)G_L(-z) \]

holds in the Fourier domain

\[ H_H(e^{j\omega})G_H(e^{j\omega}) = \left[-e^{j\omega K} H_L(-e^{-j\omega})\right] \left[-e^{-j\omega K} H_L(-e^{j\omega})\right] \]

\[ = H_L(-e^{-j\omega})H_L(-e^{j\omega}) = G_L(-e^{j\omega})H_L(-e^{j\omega}). \]

In this way all filters are expressed in terms of \( G_L(e^{j\omega}) \) or \( H_L(e^{j\omega}) \).

For example, if \( G_L(e^{j\omega}) \) is obtained using (9.55), with appropriate design conditions, then

\[ H_L(e^{j\omega}) = G_L(e^{-j\omega}) \]

\[ G_H(e^{j\omega}) = -e^{-j\omega K} G_L(-e^{-j\omega}) \]

\[ H_H(e^{j\omega}) = -e^{j\omega K} G_L(-e^{j\omega}). \]  

(9.56)
Note that the following symmetry of the frequency response amplitude functions holds

$$\left| H_L(e^{j\omega}) \right| = \left| G_L(e^{-j\omega}) \right| = \left| H_H(e^{j(\omega+\pi)}) \right| = \left| H_H(e^{-j(\omega+\pi)}) \right|.$$  

The highpass and lowpass response orthogonality

$$\sum_m h_L(m) h_H(m-2n) = 0$$
$$\sum_m g_L(m) g_H(m-2n) = 0 \quad (9.57)$$

is also satisfied with these forms of transfer functions for any $n$. Since

$$\mathcal{Z}\{h_L(n) * h_H(-n)\} = H_L(z)H_H(z^{-1})$$

and $\mathcal{Z}\{h_L(2n) * h_H(-2n)\} = 0$, in the Fourier domain this relation assumes the form

$$H_L(e^{j\omega}) H_H(e^{-j\omega}) + H_L(-e^{j\omega}) H_H(-e^{-j\omega}) = 0.$$  

This identity follows from the second relation in (9.53)

$$H_L(-e^{j\omega}) G_L(e^{j\omega}) + H_H(-e^{j\omega}) G_H(e^{j\omega}) = 0$$

with $H_H(-e^{j\omega}) = e^{j\omega K} H_L(e^{-j\omega})$, $G_H(e^{j\omega}) = -e^{-j\omega K} G_L(-e^{j\omega})$, and $H_L(e^{j\omega}) = G_L(e^{-j\omega})$ as

$$G_L(-e^{-j\omega}) G_L(e^{j\omega}) - e^{j\omega K} G_L(e^{j\omega}) e^{-j\omega K} G_L(-e^{-j\omega}) = 0.$$  

9.3.1.6 Haar Wavelet Implementation

The condition that the reconstruction filter $G_L(z)$ has zero value at $z = e^{j\pi} = -1$ means that its form is $G_L(z) = a(1 + z^{-1})$. This form without additional requirements would produce $a = 1/\sqrt{2}$ from the reconstruction relation $G_L(z) G_L(z^{-1}) + G_L(-z) G_L(-z^{-1}) = 2$. The time domain filter form is

$$g_L(n) = \frac{1}{\sqrt{2}} [\delta(n) + \delta(n-1)].$$

It corresponds to the Haar wavelet. All other filter functions can be defined using $g_L(n)$ or $G_L(e^{j\omega}).$
The same result would be obtained starting from the filter transfer functions for the Haar wavelet already introduced as

\[
H_L(z) = \frac{1}{\sqrt{2}} (1 + z) \\
H_H(z) = \frac{1}{\sqrt{2}} (1 - z).
\]

The reconstruction filters are obtained from (9.48)-(9.49)

\[
\frac{1}{\sqrt{2}} (1 + z) G_L(z) + \frac{1}{\sqrt{2}} (1 - z) G_H(z) = 2 \\
\frac{1}{\sqrt{2}} (1 - z) G_L(z) + \frac{1}{\sqrt{2}} (1 + z) G_H(z) = 0
\]
as

\[
G_L(z) = \frac{1}{\sqrt{2}} (1 + z^{-1}) \\
G_H(z) = \frac{1}{\sqrt{2}} (1 - z^{-1})
\]

with

\[
g_L(n) = \frac{1}{\sqrt{2}} \delta(n) + \frac{1}{\sqrt{2}} \delta(n - 1) \\
g_H(n) = \frac{1}{\sqrt{2}} \delta(n) - \frac{1}{\sqrt{2}} \delta(n - 1).
\]

The values impulse responses in the Haar wavelet transform (relations (9.43) and (9.59)) are:

<table>
<thead>
<tr>
<th>n</th>
<th>$\sqrt{2} h_L(n)$</th>
<th>$\sqrt{2} h_H(n)$</th>
<th>n</th>
<th>$\sqrt{2} g_L(n)$</th>
<th>$\sqrt{2} g_H(n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
</tr>
</tbody>
</table>

A detailed time domain filter bank implementation of the reconstruction process in the Haar wavelet case is described. The reconstruction is implemented in two steps:

1) The signals $s_L(n)$ and $s_H(n)$ from (9.44) are upsampled, according to (9.46), as

\[
r_L(n) = [s_L(0) \ 0 \ s_L(1) \ 0 \ s_L(2) \ 0 \ ... \ s_L(N-1) \ 0] \\
r_H(n) = [s_H(0) \ 0 \ s_H(1) \ 0 \ s_H(2) \ 0 \ ... \ s_H(N-1) \ 0]
\]
These signals are then passed through the reconstruction filters. A sum of the outputs from these filters is

\[ y(n) = r_L(n) * g_L(n) + r_H(n) * g_H(n) \]

\[ = \frac{1}{\sqrt{2}} r_L(n) + \frac{1}{\sqrt{2}} r_L(n - 1) + \frac{1}{\sqrt{2}} r_H(n) - \frac{1}{\sqrt{2}} r_H(n - 1) \]

\[ = \frac{1}{\sqrt{2}} \left[ x_L(0) \ 0 \ x_L(2) \ 0 \ x_L(4) \ ... \ 0 \ x_L(2N - 2) \ 0 \right] + \]

\[ + \frac{1}{\sqrt{2}} \left[ 0 \ x_L(0) \ 0 \ x_L(2) \ ... \ 0 \ x_L(2N - 2) \right] \]

\[ + \frac{1}{\sqrt{2}} \left[ x_H(0) \ 0 \ x_H(2) \ 0 \ x_H(4) \ ... \ 0 \ x_H(2N - 2) \ 0 \right] \]

\[ - \frac{1}{\sqrt{2}} \left[ 0 \ x_H(0) \ 0 \ x_H(2) \ ... \ 0 \ x_H(2N - 2) \right]. \]

where \( s_L(n) = x_L(2n) \) and \( s_H(n) = x_H(2n) \). From the previous relation follows

\[ y(0) = \frac{1}{\sqrt{2}} [x_L(0) + x_H(0)] = x(0) \]

\[ y(1) = \frac{1}{\sqrt{2}} [x_L(0) - x_H(0)] = x(1) \]

\[ ... \]

\[ y(2n) = \frac{1}{\sqrt{2}} [x_L(2n) + x_H(2n)] = x(2n) \]

\[ y(2n + 1) = \frac{1}{\sqrt{2}} [x_L(2n) - x_H(2n)] = x(2n + 1). \]

A system for implementation of the Haar wavelet transform of a signal with eight samples is presented in Fig. 9.30. It corresponds to the matrix form realization (9.42).

Example 9.17. For a signal \( x(n) = [1, 1, 2, 0, 2, 2, 0, 0, 2, 2, 0, 0, 0, 0] \) calculate the Haar wavelet transform coefficients, with their appropriate placement in the time-frequency plane corresponding to a signal with \( M = 16 \) samples.

The wavelet transform of a signal with \( M = 16 \) samples after the stage \( a = 1 \) is shown in Fig. 9.31(a). The whole frequency range is divided into two subregions, denoted by \( L \) and \( H \) within the coefficients \( W_1(n, L) = [x(n) + x(n + 1)] / \sqrt{2} \) and \( W_1(n, H) = [x(n) - x(n - 1)] / \sqrt{2} \) calculated at instants \( n = 0, 2, 3, 6, 8, 10, 12, 14 \). In the second stage \( (a = 2) \) the highpass region is not transformed, while the lowpass part \( s_2(n) = W_1(2n, L) \) is divided
The Haar wavelet has the duration of impulse response equal to two. In one stage, it corresponds to a two-sample STFT calculated using a rectangular window. Its Fourier transform presented in Fig.9.27 is quite rough approximation of a lowpass and highpass filter. In order to improve filter performance, an increase of the number of filter coefficients should be done. A fourth order FIR system will be considered. The impulse response of anticausal fourth order FIR filter is \( h_L(n) = [h_L(0), h_L(-1), h_L(-2), h_L(-3)] \) = \([h_0, h_1, h_2, h_3]\).
The highpass and reconstruction filter coefficients are chosen such that

\[
\begin{align*}
  n & \quad h_L(n) \quad h_H(n) \\
  0 & \quad h_0 \quad h_3 \\
  -1 & \quad h_1 \quad -h_2 \\
  -2 & \quad h_2 \quad h_1 \\
  -3 & \quad h_3 \quad -h_0 \\
\end{align*}
\]


then relation (9.60) is satisfied with \( K = 3 \), since \( h_L(n) = g_L(-n), g_H(n) = (-1)^n g_L(3 - n) \), and \( h_H(n) = (-1)^n g_L(n + 3) \).
The reconstruction conditions

\[ H_L(z)G_L(z) + H_H(z)G_H(z) = 2 \]
\[ H_L(-z)G_L(z) + H_H(-z)G_H(z) = 0 \]

are satisfied if

\[ h_0^2 + h_1^2 + h_2^2 + h_3^2 = 1. \]

Using the z-transform of the corresponding filters, it follows

\[
H_L(z)G_L(z) + H_H(z)G_H(z) = \left( h_0 + h_1 z + h_2 z^2 + h_3 z^3 \right) \left( h_0 + h_1 z^{-1} + h_2 z^{-2} + h_3 z^{-3} \right) \\
+ \left( -h_0 z^3 + h_1 z^2 - h_2 z + h_3 \right) \left( -h_0 z^{-3} + h_1 z^{-2} - h_2 z^{-1} + h_3 \right) = 2(h_0^2 + h_1^2 + h_2^2 + h_3^2) = 2
\]

and

\[
H_L(-z)G_L(z) + H_H(-z)G_H(z) = \left( h_0 - h_1 z + h_2 z^2 - h_3 z^3 \right) \left( h_0 + h_1 z^{-1} + h_2 z^{-2} + h_3 z^{-3} \right) \\
+ \left( h_0 z^3 + h_1 z^2 + h_2 z + h_3 \right) \left( -h_0 z^{-3} + h_1 z^{-2} - h_2 z^{-1} + h_3 \right) = 0.
\]

For the calculation of impulse response values \( h_0, h_1, h_2, h_3 \) of a fourth order system (9.60) four independent equations (conditions) are needed. We already have three conditions. The filter has to satisfy zero-frequency condition \( H_L(e^{j0}) = \sqrt{2} \), high-frequency condition \( H_L(e^{j\pi}) = 0 \) and the reconstruction condition \( h_0^2 + h_1^2 + h_2^2 + h_3^2 = 1 \). Therefore one more condition is needed. In the Daubechies D4 wavelet derivation the fourth condition is imposed so that the derivative of the filter transfer function at \( \omega = \pi \) is equal to zero

\[ \frac{dH_L(e^{j\omega})}{d\omega} \bigg|_{\omega=\pi} = 0. \]

This condition, meaning a smooth approach to zero-value at \( \omega = \pi \), also guarantees that the output of high-pass filter \( H_H(-z) \) to the linear input signal, \( x(n) = an + b \), will be zero. This will be illustrated later. Now we
have a system of four equations,

\[ h_0 + h_1 + h_2 + h_3 = \sqrt{2} \text{ from } H_L(e^{i0}) = \sqrt{2} \]
\[ h_0^2 + h_1^2 + h_2^2 + h_3^2 = 1 \text{ reconstruction condition} \]
\[ h_0 - h_1 + h_2 - h_3 = 0 \text{ from } H_L(e^{i\pi}) = 0 \]
\[ -h_1 + 2h_2 - 3h_3 = 0 \text{ from } \frac{dH_L(e^{i\omega})}{d\omega} \bigg|_{\omega=\pi} = 0. \]

Its solution produces the fourth order Daubechies wavelet coefficients (D4)

\[
\begin{array}{cccccccc}
n & h_L(n) & h_H(n) & n & g_L(n) & g_H(n) \\
0 & 1+\sqrt{3} & 1-\sqrt{3} & 0 & 1+\sqrt{3} & 1-\sqrt{3} \\
-1 & \frac{3+\sqrt{3}}{4\sqrt{2}} & -\frac{3+\sqrt{3}}{4\sqrt{2}} & 1 & \frac{3+\sqrt{3}}{4\sqrt{2}} & \frac{3-\sqrt{3}}{4\sqrt{2}} \\
-2 & \frac{3-\sqrt{3}}{4\sqrt{2}} & \frac{3+\sqrt{3}}{4\sqrt{2}} & 2 & \frac{3-\sqrt{3}}{4\sqrt{2}} & \frac{3+\sqrt{3}}{4\sqrt{2}} \\
-3 & \frac{1-\sqrt{3}}{4\sqrt{2}} & -\frac{1+\sqrt{3}}{4\sqrt{2}} & 3 & \frac{1-\sqrt{3}}{4\sqrt{2}} & \frac{1+\sqrt{3}}{4\sqrt{2}} \\
\end{array}
\]

Note that this is just one of possible symmetric solutions of the previous system of equations, Fig.9.32.

The reconstruction conditions for the fourth order FIR filter

\[ H_L(e^{i\omega}) = h_0 + h_1e^{i\omega} + h_2e^{i2\omega} + h_3e^{i3\omega} \]

with Daubechies wavelet coefficients (D4) can also be checked in a graphical way by calculating

\[
\left|H_L(e^{i\omega})\right|^2 + \left|H_L(e^{i(\omega+\pi)})\right|^2 = 2 \]
\[H_L(e^{i(\omega+\pi)}H_L^*(e^{i\omega}) + H_L(e^{i\omega})H_L^*(e^{i(\omega+\pi)}) = 0. \]

From Fig.9.33, we can see that it is much better approximation of low and high pass filters than in the Haar wavelet case, Fig.9.27.

Another way to derive Daubechies wavelet coefficients (D4) is in using relation (9.55)

\[ P(z) + P(-z) = 2 \]

with

\[ P(z) = G_L(z)H_L(z) = G_L(z)G_L(z^{-1}) \]

Condition imposed on the transfer function \( G_L(z) \) in D4 wavelet is that its value and the value of its first derivative at \( z = -1 \) are zero-valued (smooth
Figure 9.32 Impulse responses of the D4 filters.

Figure 9.33 Amplitude of the Fourier transform of basic Daubechies D4 wavelet and scale function.
approach to the highpass zero value)

\[ G_L(e^{i\omega}) \bigg|_{\omega=\pi} = 0 \]
\[ \frac{dG_L(e^{i\omega})}{d\omega} \bigg|_{\omega=\pi} = 0. \]

Then \( G_L(z) \) must contain a factor of the form \( (1+z^{-1})^2 \). Since the filter order must be even (\( K \) must be odd), taking into account that \( (1+z^{-1})^2 \) would produce a FIR system with 3 nonzero coefficients, then we have to add at least one factor of the form \( a(1+z_1z^{-1}) \) to \( G_L(z) \). Thus, the lowest order FIR filter with an even number of (nonzero) impulse response values is

\[ G_L(z) = \left( 1 + z^{-1} \right)^2 a(1+z_1z^{-1}) \]

with

\[ P(z) = \left( 1 + z^{-1} \right)^2 \left( 1 + z^1 \right)^2 R(z) \]

where

\[ R(z) = \left[ a(1+z_1z^{-1}) \right] \left[ a(1+z_1z^1) \right] = z_0 z^{-1} + b + z_0 z. \]

Using

\[ P(z) + P(-z) = 2 \]

only the terms with even exponents of \( z \) will remain in \( P(z) + P(-z) \) producing

\[
(4z_0 + b)z^2 + 8z_0 + 6b + (4z_0 + b)z^{-1} = 1 \\
8z_0 + 6b = 1 \\
4z_0 + b = 0
\]

The solution is \( z_0 = -1/16 \) and \( b = 1/4 \). It produces \( az_1 = z_0 = -1/16 \) and \( a^2 + z_1^2 = b = 1/4 \) with

\[ a = \frac{1}{4\sqrt{2}} \left( 1 + \sqrt{3} \right) \text{ and } z_1 = \frac{1 - \sqrt{3}}{1 + \sqrt{3}} \]

and

\[ R(z) = \left( \frac{1}{4\sqrt{2}} \right)^2 \left( 1 + \sqrt{3} + \left( 1 - \sqrt{3} \right) z^{-1} \right) \left( 1 + \sqrt{3} + \left( 1 - \sqrt{3} \right) z^1 \right). \]
The reconstruction filter transfer function is

\[ G_L(z) = \frac{1}{4\sqrt{2}}(1 + z^{-1})^2 \left( 1 + \sqrt{3} + \left( 1 - \sqrt{3} \right) z^{-1} \right) \]

with

\[ g_L(n) = \frac{1}{4\sqrt{2}} \left[ (1 + \sqrt{3}) \delta(n) + (3 + \sqrt{3}) \delta(n-1) + (3 - \sqrt{3}) \delta(n-2) + (1 - \sqrt{3}) \delta(n-3) \right] . \]

All other impulse responses follow from this one (as in the presented table).

**Example 9.18.** Consider a signal that is a linear function of time

\[ x(n) = an + b. \]

Show that the condition

\[-h_L(-1) + 2h_L(-2) - 3h_L(-3) = 0 \]

following from \( \frac{dH_L(e^{j\omega})}{d\omega} \bigg|_{\omega=\pi} = 0 \)

is equivalent to the condition that highpass coefficients (output from \( H_H(e^{j\omega}) \)) are zero-valued, Fig.9.33. Show that the lowpass coefficients remain a linear function of time.

★The highpass coefficients after the first stage \( W_1(2n, H) \) are obtained by downsampling \( W_1(n, H) \) whose form is

\[ W_1(n, H) = x(n) * h_H(n) \]

\[ = x(n)h_H(0) + x(n + 1)h_H(-1) + x(n + 2)h_H(-2) + x(n + 3)h_H(-3) \]

\[ = x(n)h_3 - x(n + 1)h_2 + x(n + 2)h_1 - x(n + 3)h_0 \]

\[ = (an + b)h_3 - (n + 1)a + b)h_2 + ((n + 2)a + b)h_1 - (n + 3)a + b)h_0 \]

\[ = (a(n + 3) + b)(h_0 + h_1 - h_2 + h_3) - a(h_1 - 2h_2 + 3h_3) = 0 \]

if

\[-h_0 + h_1 - h_2 + h_3 = 0 \text{ and } h_1 - 2h_2 + 3h_3 = 0.\]

The lowpass coefficients are obtained by downsampling

\[ W_1(n, L) = x(n) * h_L(n) \]

\[ = x(n)h_0 + x(n + 1)h_1 + x(n + 2)h_2 + x(n + 3)h_3 \]

\[ = (an + b)h_0 + ((n + 1)a + b)h_1 + ((n + 2)a + b)h_2 + ((n + 3)a + b)h_3 \]

\[ = (an + b)(h_0 + h_1 + h_2 + h_3) + a(h_1 + 2h_2 + 3h_3) \]

\[ = a_4n + b_1, \]

\[ = a_4n + b_1. \]
where $a_1 = \sqrt{2}a$ and $b_1 = \sqrt{2}b + 0.8966a$.

Thus we may consider that the highpass D4 coefficients will indicate the deviation of the signal from a linear function $x(n) = an + b$. In the first stage the coefficients will indicate the deviation from the linear function within four samples. In the next stage the equivalent length of wavelet is doubled. The highpass coefficient in this stage will indicate the deviation of the signal from the linear function within doubled number of signal samples, and so on. This a significant difference from the STFT nature that is derived based on the Fourier transform and the signal decomposition and tracking its frequency content.

**Example 9.19.** Show that with the conditions

\[ h_0 + h_1 + h_2 + h_3 = \sqrt{2} \text{ from } H_L(e^{j0}) = \sqrt{2} \]
\[ -h_0 + h_1 - h_2 + h_3 = 0 \text{ from } H_L(e^{j\pi}) = 0 \]

the reconstruction condition

\[ h_0^2 + h_1^2 + h_2^2 + h_3^2 = 1 \]

is equivalent to the orthogonality property of the impulse response and its shifted version for step 2

\[
\begin{array}{ccccccccc}
0 & 0 & h_0 & h_1 & h_2 & h_3 & 0 & 0 & 0 \\
0 & 0 & 0 & h_0 & h_1 & h_2 & h_3 & 0 & 0
\end{array}
\]

given by

\[ h_2h_0 + h_3h_1 = 0. \]

★ If we write the sum of squares of the first two equations follows

\[ 2(h_0^2 + h_1^2 + h_2^2 + h_3^2) + 4h_0h_2 + 4h_1h_3 = 2. \]

Therefore, the conditions

\[ h_0^2 + h_1^2 + h_2^2 + h_3^2 = 1 \]

and

\[ h_0h_2 + h_1h_3 = 0 \]

follow from each other if $h_0 + h_1 + h_2 + h_3 = \sqrt{2}$ and $-h_0 + h_1 - h_2 + h_3 = 0$ are assumed.
The matrix for the D4 wavelet transform calculation in the first stage is of the form

\[
\begin{bmatrix}
W_1(0, L) \\
W_1(0, H) \\
W_1(2, L) \\
W_1(2, H) \\
W_1(4, L) \\
W_1(4, H) \\
W_1(6, L) \\
W_1(6, H)
\end{bmatrix}
= \begin{bmatrix}
h_0 & h_1 & h_2 & h_3 & 0 & 0 & 0 & 0 \\
h_3 & -h_2 & h_1 & -h_0 & 0 & 0 & 0 & 0 \\
0 & 0 & h_0 & h_1 & h_2 & h_3 & 0 & 0 \\
0 & 0 & h_3 & -h_2 & h_1 & -h_0 & 0 & 0 \\
0 & 0 & 0 & 0 & h_0 & h_1 & h_2 & h_3 \\
0 & 0 & 0 & 0 & h_3 & -h_2 & h_1 & -h_0 \\
h_2 & h_3 & 0 & 0 & 0 & 0 & h_0 & h_1 \\
h_1 & -h_0 & 0 & 0 & 0 & 0 & h_3 & -h_2
\end{bmatrix}
\begin{bmatrix}
x(0) \\
x(1) \\
x(2) \\
x(3) \\
x(4) \\
x(5) \\
x(6) \\
x(7)
\end{bmatrix}.
\]

In the first row of transformation matrix the coefficients corresponds to \(h_L(n)\), while the second row corresponds to \(h_H(n)\). The first row produces D4 scaling function, while the second row produces D4 wavelet function. The coefficients are shifted for 2 in next rows. As it has been described in the Hann(ing) window reconstruction case, the calculation should be performed in a circular manner, assuming signal periodicity. That is why the coefficients are circularly shifted in the last two rows.

Example 9.20. Consider a signal \(x(n) = 64 - |n - 64|\) within \(0 \leq n \leq 128\). How many nonzero coefficients will be in the first stage of the wavelet transform calculation using D4 wavelet functions. Assume that the signal can appropriately be extended so that the boundary effects can be neglected.

\[\star\] In the first stage all highpass coefficients corresponding to linear four-sample intervals will be zero. It means that out of 64 high pass coefficients (calculated with step two in time) only one nonzero coefficient will exist, calculated for \(n = 62\), including nonlinear interval \(62 \leq n \leq 65\). It means that almost a half of the coefficients can be omitted in transmission or storage, corresponding to 50% compression ratio. In the DFT analysis this would correspond to a signal with a half of (the high frequency) spectrum being equal to zero. In the wavelet analysis this process would be continued with additional savings in next stages of the wavelet transform coefficients calculation. It also means that if there is some noise in the signal, we can filter out all zero-valued coefficients using an appropriate threshold. For this kind of signal (piecewise linear function of time) we will be able to improve the signal-to-noise ratio for about 3 dB in just one wavelet stage.

Example 9.21. For the signal \(x(n) = \delta(n - 7)\) defined within \(0 \leq n \leq 15\) calculate the wavelet transform coefficients using the D4 wavelet/scale function. Repeat the same calculation for the signal \(x(n) = 2\cos(16\pi n/N) + 1\) with \(0 \leq n \leq N - 1\) with \(N = 16\).
The wavelet coefficients in the first stage (scale \( a = 1 \), see also Fig.9.30) are

\[
W_1(2n,H) = x(2n)h_H(0) + x(2n+1)h_H(-1) \\
+ x(2n+2)h_H(-2) + x(2n+3)h_H(-3)
\]

\[
x(2n)h_3 - x(2n+1)h_2 + x(2n+2)h_1 - x(2n+3)h_0
\]

with

\[
[h_3,h_2,h_1,h_0] = \left[ \frac{1 - \sqrt{3}}{4\sqrt{2}}, \frac{3 - \sqrt{3}}{4\sqrt{2}}, \frac{3 + \sqrt{3}}{4\sqrt{2}}, \frac{1 + \sqrt{3}}{4\sqrt{2}} \right].
\]

In specific, \( W_1(0,H) = 0 \), \( W_1(2,H) = 0 \), \( W_1(4,H) = -0.4830 \), \( W_1(6,H) = -0.2241 \), \( W_1(8,H) = 0 \), \( W_1(10,H) = 0 \), \( W_1(12,H) = 0 \), and \( W_1(14,H) = 0 \).

The lowpass part of the first stage values

\[ s_2(n) = W_1(2n,L) = x(2n)h_0 + x(2n+1)h_1 + x(2n+2)h_2 + x(2n+3)h_3 \]

are \( W_1(0,L) = 0 \), \( W_1(2,L) = 0 \), \( W_1(4,L) = -0.1294 \), \( W_1(6,L) = 0.8365 \), \( W_1(8,L) = 0 \), \( W_1(10,L) = 0 \), \( W_1(12,L) = 0 \), and \( W_1(14,L) = 0 \). Values of \( s_2(n) \) are defined for \( 0 \leq n \leq 7 \) as

\[ s_2(n) = -0.1294\delta(n - 2) + 0.8365\delta(n - 3). \]

This signal is the input to the next stage (scale \( a = 2 \)). The highpass output of the stage two is

\[ W_2(4n,H) = s_2(n)h_3 - s_2(n+1)h_2 + s_2(n+2)h_1 - s_2(n+3)h_0. \]

The values of \( W_2(4n,H) \) are: \( W_2(0,H) = -0.5123 \), \( W_2(4,H) = -0.1708 \), \( W_2(8,H) = 0 \), and \( W_2(12,H) = 0 \). The lowpass values at this stage at the input to the next stage (\( a = 3 \)) calculation

\[ s_3(n) = W_2(4n,L) = s_2(n)h_0 + s_2(n+1)h_1 + s_2(n+2)h_2 + s_2(n+3)h_3. \]

They are \( W_2(0,L) = -0.1373 \), \( W_2(4,L) = 0.6373 \), \( W_2(8,L) = 0 \), and \( W_2(12,L) = 0 \).

Since there is only 4 samples in \( s_3(n) \) this is the last calculation. The coefficients in this stage are \( W_3(0,H) = -0.1251 \), \( W_3(8,H) = 0.4226 \) and \( W_3(0,L) = 0.4668 \), \( W_3(8,L) = -0.1132 \). The absolute value of the wavelet transform of \( x(n) \) with D4 wavelet function is shown in Fig.9.34.

For the signal \( x(n) = 2\cos(2\pi 8n/N) + 1 \) with \( 0 \leq n \leq N - 1 \) with \( N = 16 \) the same calculation is done. Here it is important to point out that the circular convolutions should be used. The wavelet transform coefficients are \( W_1(2n,L) = 1.4412 \) and \( W_1(2n,H) = 2.8284 \). Values in the next stage are \( W_2(2n,H) = 0 \) and \( W_2(2n,L) = 2 \). The third stage values are \( W_3(2n,H) = 0 \) and \( W_3(2n,L) = 2.8284 \). Compare these results with Fig. 9.26(a). Since the impulse response duration is 4 and the step is 2 this could be considered as a kind of signal analysis with overlapping.
The inverse matrix for the D4 wavelet transform for a signal with $N = 8$ samples would be calculated from the lowest level in this case for $a = 2$ with coefficients $W_2(0, L)$, $W_2(0, H)$, $W_2(4, L)$, and $W_2(4, H)$. The lowpass part of signal at level $a = 1$ would be reconstructed using

$$
\begin{bmatrix}
W_1(0, L) \\
W_1(2, L) \\
W_1(4, L) \\
W_1(6, L)
\end{bmatrix}
\begin{bmatrix}
h_0 & h_3 & h_2 & h_1 \\
h_1 & -h_2 & h_3 & -h_0 \\
h_2 & h_1 & h_0 & h_3 \\
h_3 & -h_0 & h_1 & -h_2
\end{bmatrix}
\begin{bmatrix}
W_2(0, L) \\
W_2(0, H) \\
W_2(4, L) \\
W_2(4, H)
\end{bmatrix}.
$$

After the lowpass part $W_1(0, L)$, $W_1(2, L)$, $W_1(4, L)$, and $W_1(6, L)$ are reconstructed, they are used with wavelet coefficients from this stage $W_1(0, H)$, $W_1(2, H)$, $W_1(4, H)$, and $W_1(6, H)$ to reconstruct the signal as

$$
\begin{bmatrix}
x(0) \\
x(1) \\
x(2) \\
x(3) \\
x(4) \\
x(5) \\
x(6) \\
x(7)
\end{bmatrix}
= \begin{bmatrix}
h_0 & h_3 & 0 & 0 & 0 & 0 & h_2 & h_1 \\
h_1 & -h_2 & 0 & 0 & 0 & 0 & h_3 & -h_0 \\
h_2 & h_1 & h_0 & h_3 & 0 & 0 & 0 & 0 \\
h_3 & -h_0 & h_1 & -h_2 & 0 & 0 & 0 & 0 \\
0 & 0 & h_2 & h_1 & h_0 & h_3 & 0 & 0 \\
0 & 0 & h_3 & -h_0 & h_1 & -h_2 & 0 & 0 \\
0 & 0 & 0 & 0 & h_2 & h_1 & h_0 & h_3 \\
0 & 0 & 0 & 0 & h_3 & -h_0 & h_1 & -h_2
\end{bmatrix}
\begin{bmatrix}
W_1(0, L) \\
W_1(0, H) \\
W_1(2, L) \\
W_1(2, H) \\
W_1(4, L) \\
W_1(4, H) \\
W_1(6, L) \\
W_1(6, H)
\end{bmatrix}.
$$

(9.62)
This procedure can be continued for signal of length \( N = 16 \) with one more stage. Additional stage would be added for \( N = 32 \) and so on.

**Example 9.22.** For the Wavelet transform from the previous example find its inverse (reconstruct the signal).

\[
\begin{bmatrix}
W_2(0, L) \\
W_2(4, L) \\
W_2(8, L) \\
W_2(12, L)
\end{bmatrix} = \begin{bmatrix}
h_0 & h_3 & h_2 & h_1 \\
h_1 & -h_2 & h_3 & -h_0 \\
h_2 & h_1 & h_0 & h_3 \\
h_3 & -h_0 & h_1 & -h_2
\end{bmatrix} \cdot \begin{bmatrix}
W_3(0, L) \\
W_3(0, H) \\
W_3(8, L) \\
W_3(8, H)
\end{bmatrix}
\]

\[
= \begin{bmatrix}
h_0 & h_3 & h_2 & h_1 \\
h_1 & -h_2 & h_3 & -h_0 \\
h_2 & h_1 & h_0 & h_3 \\
h_3 & -h_0 & h_1 & -h_2
\end{bmatrix} \cdot \begin{bmatrix}
0.4668 \\
-0.1251 \\
-0.1132 \\
-0.4226
\end{bmatrix} = \begin{bmatrix}
-0.1373 \\
0.6373 \\
0 \\
0
\end{bmatrix}.
\]

Then \( W_2(4n, L) = s_3(n) \) are used with the wavelet coefficients \( W_2(4n, H) \) to reconstruct \( W_1(2n, L) \) or \( s_2(n) \) using

\[
\begin{bmatrix}
W_1(0, L) \\
W_1(2, L) \\
W_1(4, L) \\
W_1(6, L) \\
W_1(8, L) \\
W_1(10, L) \\
W_1(12, L) \\
W_1(14, L)
\end{bmatrix} = \begin{bmatrix}
h_0 & h_3 & 0 & 0 & 0 & 0 & h_2 & h_1 \\
h_1 & -h_2 & 0 & 0 & 0 & 0 & h_3 & -h_0 \\
h_2 & h_1 & h_0 & h_3 & 0 & 0 & 0 & 0 \\
h_3 & -h_0 & h_1 & -h_2 & 0 & 0 & 0 & 0 \\
0 & 0 & h_2 & h_1 & h_0 & h_3 & 0 & 0 \\
0 & 0 & h_3 & -h_0 & h_1 & -h_2 & 0 & 0 \\
0 & 0 & 0 & h_2 & h_1 & h_0 & h_3 & 0 \\
0 & 0 & 0 & 0 & h_3 & -h_0 & h_1 & -h_2
\end{bmatrix} \cdot \begin{bmatrix}
W_2(0, L) \\
W_2(0, H) \\
W_2(4, L) \\
W_2(4, H) \\
W_2(8, L) \\
W_2(8, H) \\
W_2(12, L) \\
W_2(12, H)
\end{bmatrix}
\]

The obtained values \( W_1(n, L) \) with the wavelet coefficients \( W_1(n, H) \) are used to reconstruct the original signal \( x(n) \). The transformation matrix in this case is of \( 16 \times 16 \) order and it is formed using the same structure as the previous transformation matrix.

\[\Box\]

### 9.3.1.8 Daubechies D4 Wavelet Functions in Different Scales

Although the wavelet realization can be performed using the same basic function presented in the previous section, here we will consider the equivalent wavelet function \( h_H(n) \) and equivalent scale function \( h_L(n) \) in different scales. To this aim we will analyze the reconstruction part of the system. Assume that in the wavelet analysis of a signal only one coefficient is nonzero. Also assume that this nonzero coefficient is at the exit of all lowpass filters structure. It means that the signal is equal to the basic scale function in
the wavelet analysis. The scale function can be found in an inverse way, by reconstructing signal corresponding to this delta pulse like transform. The system of reconstruction filters is shown in Fig. 9.35. Note that this case and coefficient in the Haar transform would correspond to $W_{i}(0, L) = 1$ in (9.42) or in Fig. 9.30. The reconstruction process consists of signal upsampling and passing it through the reconstruction stages. For example, the output of the third reconstruction stage has the $z$-transform

$$\Phi_2(z) = G_L(z)G_L(z^2)G_L(z^4).$$

In the time domain the reconstruction is performed as

$$\Phi_0(n) = \delta(n) * g_L(n) = g_L(n)$$
$$\Phi_1(n) = [\Phi_0(0) 0 \Phi_0(1) 0 \Phi_0(2) 0 \Phi_0(3)] * g_L(n)$$
$$\Phi_2(n) = [\Phi_1(0) 0 \Phi_1(1) 0 ... \Phi_1(8) 0 \Phi_1(9)] * g_L(n)$$
$$...$$

$$\Phi_{a+1}(n) = \sum_p \Phi_a(p)g_L(n - 2p)$$

where $g_L(n)$ is the four sample impulse response (Daubechies D4 coefficients). Duration of the scale function $\Phi_1(n)$ is $(4 + 3) + 4 - 1 = 10$ samples, while the duration of $\Phi_2(n)$ is $19 + 4 - 1 = 22$ samples. The scale function for
different scales $a$ (exists of different reconstruction stages) are is presented in Fig.9.37. Normalized values $\phi_a(n)2^{(a+1)/2}$ are presented. The amplitudes are scaled by $2^{(a+1)/2}$ in order to keep their values within the same range for various $a$.

In a similar way the wavelet function $\psi(n)$ is calculated. The mother wavelet is obtained in the wavelet analysis of a signal when only one nonzero coefficient exists at the highpass of the lowest level of the signal analysis. To reconstruct the mother wavelet the reconstruction system as in Fig.9.36 is used. The values of $\psi(n)$ are calculated: using the values of $g_H(n)$ at the first input, upsampling it and passing trough the reconstruction system with $g_L(n)$, to obtain $\psi_1(n)$ and repeating this procedure for the next steps. The resulting $z$-transform is:

$$\Psi(z) = G_H(z)G_L(z^2)G_L(z^4).$$

In the Haar transform (9.42) and Fig.9.30 this case would correspond to $W_4(0,H) = 1$. 
Calculation in the time of the wavelet function in different scales is
done using

\[ \psi_0(n) = \delta(n) \ast g_H(n) = g_H(n) \]
\[ \psi_1(n) = [\psi_1(0) 0 \psi_1(1) 0 \psi_1(2) 0 \psi_1(3)] \ast g_L(n) \]
\[ \psi_2(n) = [\psi_2(0) 0 \psi_2(1) 0 ... \psi_2(8) 0 \psi_2(9)] \ast g_L(n) \]

\[ \vdots \]
\[ \psi_{a+1}(n) = \sum_p \psi_a(p)g_L(n - 2p) \]

Different scales of the wavelet function, are presented in Fig. 9.37. Values are normalized using \( \psi_a(n)^2 \).

Wavelet function are orthogonal in different scales, with corresponding steps, as well. For example, it is easy to show that

\[ \langle \psi_0(n - 2m), \psi_1(n) \rangle = 0 \]

since

\[ \langle \psi_0(n - 2m), \psi_1(n) \rangle = \sum_p g_H(p) \left( \sum_n g_H(n - 2m)g_L(n - 2p) \right) = 0 \]

for any \( p \) and \( m \) according to (9.57).

Note that the wavelet and scale function in the last row are plotted as the continuous functions. The continuous wavelet transform (CWT) is calculated by using the discretized versions of the continuous functions. However in contrast to the discrete wavelet transform whose step in time and scale change is strictly defined, the continuous wavelet transform can be used with various steps and scale functions.

**Example 9.23.** In order to illustrate the procedure it has been repeated for the Haar wavelet when \( g_L(n) = [1 1] \) and \( g_H(n) = [1 -1] \). The results are presented in Fig. 9.38.

### 9.3.1.9 Daubechies D6 Wavelet Transform

The results derived for Daubechies D4 wavelet transform can be extended to higher order polynomial functions. Consider a sixth order FIR system

\[ h_L(n) = [h_L(0), h_L(-1), h_L(-2), h_L(-3), h_L(-4), h_L(-5)] \]
\[ = [h_0, h_1, h_2, h_3, h_4, h_5]. \]
Figure 9.37 The Daubechies D4 wavelet scale function and wavelet calculated using the filter bank relation in different scales: $a = 0$ (first row), $a = 1$ (second row), $a = 2$ (third row), $a = 3$ (fourth row), $a = 10$ (fourth row—approximation of a continuous domain). The amplitudes are scaled by $2^{(a+1)/2}$ to keep them within the same range. Values $\psi_a(n)2^{(a+1)/2}$ and $\phi_a(n)2^{(a+1)/2}$ are presented.
Chapter 10

Sparse Signal Processing

A discrete-time signal can be transformed into other domains using different signal transformations. Some signals that cover the whole considered interval in one domain could be sparse in a transformation domain, i.e., could be located within a few nonzero coefficients. Compressive sensing is a field dealing with a model for data acquisition including the problem of sparse signal recovery from a reduced set of observations. A reduced set of observations can be a result of a desire to sense a sparse signal with the lowest possible number of measurements/observations (compressive sensing). It can also be a result of a physical or measurement unavailability to take a complete set of observations. Since the signal samples are linear combinations of the signal transformation coefficients they could be considered as the observations of a sparse signal in the transformation domain. In applications it could also happen that some arbitrarily positioned samples of a signal are so heavily corrupted by disturbances that it is better to omit them and consider as unavailable in the analysis and to try to reconstruct the signal with a reduced set of samples. Although the reduced set of observations/samples appears in the first case as a result of user strategy to compress the information, while in the next two cases the reduced set of samples is not a result of user intention, all of them can be considered within the unified framework. Under some conditions, a full reconstruction of a sparse signal can be performed with a reduced set of observations/samples, as in the case if a complete set of samples/observations were available. A priori information about the nature of the analyzed signal, i.e., its sparsity in a known transformation domain, must be used in this analysis. Sparsity
Sparse Signal Processing

is the main requirement that should be satisfied in order to efficiently apply
the compressive sensing methods for sparse signal reconstruction.

The topic of this chapter is to analyze the signals that are sparse in
one of the common transformations domains. The DFT is used as a study
case. The compressive sensing results and algorithms are presented and
used only as a tool to solve engineering problems, involving sparse signals.

10.1 ILLUSTRATIVE EXAMPLES

Before we start the analysis we will describe few widely known examples
that can be interpreted and solved within the context of sparse signal
processing and compressive sensing.

Consider a large set of real numbers $X(0), X(1),...,X(N - 1)$. Assume
that only one of them is nonzero (or different from a common and known
expected value). We do not know either its position or its value. The aim is to
find the position and the value of this number. This case can easily be related
to many real life examples when we have to find one sample which differs
from other $N - 1$ samples. The nonzero value (or the difference from the
expected value) will be denoted by $X(i)$. A direct way to find the position
of nonzero (different) sample would be to perform up to $N$ measurements
and compare each of them with zero (the expected) value. However, if $N$
is very large and there is only one nonzero (different than expected) sample
we can get the result in just a few observations/measurements. A procedure
for the reduced number of observations/measurements is described next.

Take random numbers as weighting coefficients $a_i, i = 0,1,2,...,N - 1,$
for each sample. Measure the total value of all $N$ weighted samples, with
weights $a_i$, from the set. Since only one is different from the common and
known expected value $m$ (or from zero) we will get the total measured value

\[ M = a_1m + a_2m + ... + a_i(m + X(i)) + ... + a_Nm. \]

From this measured value $M$ subtract the expected value $M_T = (a_1 + a_2 +
... + a_N)m$. The obtained value of this observation/measurement, denoted
by $y(0)$, is

\[ y(0) = M - M_T = \sum_{k=0}^{N-1} a_kX(k) = a_iX(i), \]

since nonzero value in the space of $X(0), X(1),...,X(N - 1)$ is at one position
only, $X(k) = X(i)\delta(k - i)$.

As an illustration consider a set of $N$ bags. Assume that only one bag
contains all false coins of a weight $m + X(i)$. It is different from the known
weight \(m\) of true coins. The goal is to find the position and the difference in weight of false coins. From each of \(N\) bags we will take \(a_i\), \(i = 1, 2, ..., N\), coins. Number of coins from the \(i\)th bag is denoted by \(a_i\). The total measured weight of all coins from \(N\) bags is \(M\), Fig. 10.1.

After the expected value is subtracted the observation/measurement \(y(0)\) is obtained

\[
y(0) = \sum_{k=0}^{N-1} X(k) \psi_k(0),
\]

where the weighting coefficients for this measurement are denoted by \(\psi_k(0) = a_k, k = 0, 1, ..., N - 1\). In the space of unknowns (variables) \(X(0), X(1), ..., X(N - 1)\) this equation represents an \(N\)-dimensional hyperplane. We know that only one unknown \(X(k)\) is nonzero at the unknown position
Sparse Signal Processing

k = i. Any cross-section of hyperplane (10.1) with any of coordinate axes could be a solution of our problem. Assuming that a single X(k) is nonzero a solution will exist for any k. Thus, one measurement would produce N possible single nonzero values equal to

\[ X(k) = y(0) / \psi_k(0), \quad \psi_k(0) \neq 0, k = 0, 1, 2, ..., N - 1. \]

As expected, from one measurement we are not able to solve the problem and to find the position and the value of nonzero sample.

If we perform one more measurement \( y(1) \) with another set of weighting coefficients \( \psi_k(1), k = 0, 1, ..., N - 1 \), and get measured value \( y(1) = X(i)\psi_i(1) \) the result will be a hyperplane

\[ y(1) = \sum_{k=0}^{N-1} X(k)\psi_k(1). \]

This measurement will produce a new set of possible solutions for each \( X(k) \) as

\[ X(k) = y(1) / \psi_k(0), \quad k = 0, 1, 2, ..., N - 1. \]

If these two hyperplanes (sets of solutions) produce only one common value

\[ X(i) = y(0) / \psi_i(0) = y(1) / \psi_i(1). \]

then it is the solution of our problem.

As an example consider \( N = 5 \) sets of coins. The common weight of true coins is 2. In the first measurement we use \( \psi_i(0) = a_i = i \) coins from each set. The total weight of coins is \( M = 31 \). It is obtained by measuring \((1 + 2 + 3 + 4 + 5)2 + iX(i) = M\), where \( X(i) \) is the unknown weight difference of false coins. It means that \( iX(i) = 1 \), since all true coins would produce \((1 + 2 + 3 + 4 + 5)2 = 30\). If the false coins were in the first set the weight difference would be \( X(1) = 1/1 = 1 \), if they were in the second set then \( X(2) = 1/2 \), and so on, \( X(3) = 1/3, X(4) = 1/4, X(5) = 1/5 \). False coins can be in any of five sets. Perform one more measurement with \( \psi_i(1) = a_i = i^2 \) coins from each set. Total measured weight is now \( M = 113 \). It is obtained as \( M = 2(1^2 + 2^2 + 3^2 + 4^2 + 5^2) + i^2X(i) = 113 \). Obviously \( i^2X(i) = 3 \). Again if the false coins were in the first set then \( X(1) = 3/1 \), the second would produce \( X(2) = 3/2^2 = 3/4 \), and so on, \( X(3) = 3/3^2 = 1/3, X(4) = 3/4^2 = 3/16, X(5) = 3/25 \). The solution satisfying both equations is \( X(3) = 1/3 \). Thus, false coins are in the third set. Their weight is \( 2 + 1/3 = 7/3 \). Note that we would not be able to solve the problem with two measurements if we got two values \( X(i) \) and \( X(k) \) for \( i \neq k \) satisfying both equations.
In a matrix form these two measurements are
\[
\begin{bmatrix}
y(0) \\
y(1)
\end{bmatrix}
= \begin{bmatrix}
\psi_0(0) & \psi_1(0) & \ldots & \psi_{N-1}(0) \\
\psi_0(1) & \psi_1(1) & \ldots & \psi_{N-1}(1)
\end{bmatrix}
\begin{bmatrix}
X(0) \\
X(1) \\
\vdots \\
X(N-1)
\end{bmatrix}
\]
\[y = AX\]
where \(A\) is the matrix of coefficients (measurement matrix)
\[
A = \begin{bmatrix}
\psi_0(0) & \psi_1(0) & \ldots & \psi_{N-1}(0) \\
\psi_0(1) & \psi_1(1) & \ldots & \psi_{N-1}(1)
\end{bmatrix}
\]
and \(y\) are observations/measurements of sparse variable \(X\).

Common value for two measurements \(X(i) = y(0)/\psi_i(0)\) and \(X(i) = y(1)/\psi_i(1)\) is unique if
\[
\psi_i(0)\psi_k(1) - \psi_i(1)\psi_k(0) \neq 0
\]
for any \(i \neq k\).

In order to prove this statement assume that two different solutions \(X(i)\) and \(X(k)\), for the case of one nonzero coefficient, satisfy the same measurement hyperplane equations
\[
\psi_i(0)X(i) = y(0), \quad \psi_i(1)X(i) = y(1)
\]
and
\[
\psi_k(0)X(k) = y(0), \quad \psi_k(1)X(k) = y(1).
\]
Then
\[
\psi_i(0)X(i) = \psi_k(0)X(k)
\]
and
\[
\psi_i(1)X(i) = \psi_k(1)X(k).
\]
If we divide these two equations we get
\[
\frac{\psi_i(0)}{\psi_i(1)} = \frac{\psi_k(0)}{\psi_k(1)}
\]
or \(\psi_i(0)\psi_k(1) - \psi_i(1)\psi_k(0) = 0\). This is contrary to the assumption that \(\psi_i(0)\psi_k(1) - \psi_i(1)\psi_k(0) \neq 0\).

The same conclusion can be made considering matrix form relations for \(X(i)\) and \(X(k)\). If both of them may satisfy the same two measurements
then
\[
\begin{bmatrix}
y(0) \\
y(1)
\end{bmatrix} =
\begin{bmatrix}
\psi_i(0) & \psi_k(0) \\
\psi_i(1) & \psi_k(1)
\end{bmatrix}
\begin{bmatrix}
X(i) \\
0
\end{bmatrix}
\]
\[
\begin{bmatrix}
y(0) \\
y(1)
\end{bmatrix} =
\begin{bmatrix}
\psi_i(0) & \psi_k(0) \\
\psi_i(1) & \psi_k(1)
\end{bmatrix}
\begin{bmatrix}
0 \\
X(k)
\end{bmatrix}.
\]

Subtraction of the previous matrix equations results in
\[
\begin{bmatrix}
\psi_i(0) & \psi_k(0) \\
\psi_i(1) & \psi_k(1)
\end{bmatrix}
\begin{bmatrix}
X(i) \\
-X(k)
\end{bmatrix} = 0.
\]

For \(\psi_i(0)\psi_k(1) - \psi_i(1)\psi_k(0) \neq 0\) follows \(X(i) = X(k) = 0\). Therefore two different nonzero solutions \(X(i)\) and \(X(k)\) in this case cannot exist. This concludes the proof that the solution is unique if
\[
\psi_i(0)\psi_k(1) - \psi_i(1)\psi_k(0) = \det
\begin{bmatrix}
\psi_i(0) & \psi_k(0) \\
\psi_i(1) & \psi_k(1)
\end{bmatrix} \neq 0
\]
for any \(i \neq k\). It also means that \(\text{rank}(A_2) = 2\) for any \(A_2\) being a \(2 \times 2\) submatrix of the matrix of coefficients (measurement matrix) \(A\). For additional illustration of this simple problem see Section 10.5.2.

In numerical and practical applications we would not be satisfied, if for example \(\psi_i(0)\psi_k(1) - \psi_i(1)\psi_k(0) \neq 0\) but \(\psi_i(0)\psi_k(1) - \psi_i(1)\psi_k(0) = \epsilon\) close to zero. In this case the theoretical condition for a unique solution would be satisfied, however the analysis and possible inversion would be highly sensitive to any kind of noise, including quantization noise. Thus, a practical requirement is that the determinant is not just different from zero, but that it sufficiently differs from zero so that an inversion stability and robustness to a noise is achieved. Inversion stability for a matrix \(B\) is commonly described by the condition number of matrix
\[
\text{cond}\{B\} = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}}
\]
where \(\lambda_{\text{max}}\) and \(\lambda_{\text{min}}\) are the largest and the smallest eigenvalue of matrix \(B\) (when \(B^H B = BB^H\)). The inversion stability worsens as \(\lambda_{\text{min}}\) approaches to zero (when \(\lambda_{\text{min}}\) is small as compared to \(\lambda_{\text{max}}\)). For stable and robust

1 The value of determinant of matrix \(B\) is equal to the product of its eigenvalues, \(\det\{B\} = \lambda_1\lambda_2...\lambda_N\), where \(N\) is the order of square matrix \(B\). Note that the condition number can be interpreted as a ratio of the norms-two (square roots of energies) of noise \(\epsilon\) and signal \(x\) after and before inversion \(y + y_\epsilon = B^{-1}(x + \epsilon)\). This number is always greater or equal to 1. The best value for this ratio is achieved when \(\lambda_{\text{min}}\) is close to \(\lambda_{\text{max}}\).
calculations a requirement

\[
\frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} \leq 1 + \delta
\]

is imposed, with a nonnegative constant \(\delta\) being sufficiently small. In our example this condition should hold for any submatrix \(A_2 = B\).

The previous experiment can be repeated assuming two nonzero values \(X(i)\) and \(X(k)\), Fig.10.1(right). In the case of two nonzero coefficients, two measurements

\[
y(0) = \sum_{l=0}^{N-1} X(l) \psi_l(0) = X(i)\psi_i(0) + X(k)\psi_k(0) \tag{10.3}
\]
\[
y(1) = \sum_{l=0}^{N-1} X(l) \psi_l(1) = X(i)\psi_i(1) + X(k)\psi_k(1)
\]

will result in \(X(i)\) and \(X(k)\) for any \(i\) and \(k\). They are the solution of a system with two equations and two unknowns. Therefore, with two measurements we cannot get a result of the problem and find the positions and the values of nonzero coefficients. If two more measurements are performed then an additional system with two equations

\[
y(2) = X(i)\psi_i(2) + X(k)\psi_k(2) \tag{10.4}
\]
\[
y(3) = X(i)\psi_i(3) + X(k)\psi_k(3)
\]

is formed. Two systems of two equations (10.3) and (10.4) could be solved for \(X(i)\) and \(X(k)\) for each combination of \(i\) and \(k\). If these two systems produce only one common solution pair \(X(i)\) and \(X(k)\) then this pair is the solution of our problem. As in the case of one nonzero coefficient, we may show that the sufficient condition for a unique solution is

\[
\det \begin{bmatrix}
\psi_{k_1}(0) & \psi_{k_2}(0) & \psi_{k_3}(0) & \psi_{k_4}(0) \\
\psi_{k_1}(1) & \psi_{k_2}(1) & \psi_{k_3}(1) & \psi_{k_4}(1) \\
\psi_{k_1}(2) & \psi_{k_2}(2) & \psi_{k_3}(2) & \psi_{k_4}(2) \\
\psi_{k_1}(3) & \psi_{k_2}(3) & \psi_{k_3}(3) & \psi_{k_4}(3)
\end{bmatrix} \neq 0 \tag{10.5}
\]

for any \(k_1, k_2, k_3\) and \(k_4\) or \(\text{rank}(A_4) = 4\) for any \(A_4\), where \(A_4\) is a \(4 \times 4\) submatrix of the matrix of coefficients \(A\). In numeric realizations, the condition is \(\text{cond}\{A_4\} \leq 1 + \delta\) with sufficiently small \(\delta\) for all \(A_4\). Suppose that (10.5) holds and that two pairs of solutions of the problem \(X(k_1), X(k_2)\)
and $X(k_3), X(k_4)$ exist. Then

$$
\begin{bmatrix}
y(0) \\
y(1) \\
y(2) \\
y(3)
\end{bmatrix}
= 
\begin{bmatrix}
\psi_{k_1}(0) & \psi_{k_1}(1) & \psi_{k_2}(0) & \psi_{k_2}(1) & \psi_{k_3}(0) & \psi_{k_3}(1) & \psi_{k_4}(0) & \psi_{k_4}(1) \\
\psi_{k_1}(2) & \psi_{k_1}(2) & \psi_{k_2}(2) & \psi_{k_2}(2) & \psi_{k_3}(2) & \psi_{k_3}(2) & \psi_{k_4}(2) & \psi_{k_4}(2) \\
\psi_{k_1}(3) & \psi_{k_1}(3) & \psi_{k_2}(3) & \psi_{k_2}(3) & \psi_{k_3}(3) & \psi_{k_3}(3) & \psi_{k_4}(3) & \psi_{k_4}(3)
\end{bmatrix}
\begin{bmatrix}
X(k_1) \\
X(k_2) \\
X(k_3) \\
X(k_4)
\end{bmatrix}
$$

and

$$
\begin{bmatrix}
y(0) \\
y(1) \\
y(2) \\
y(3)
\end{bmatrix}
= 
\begin{bmatrix}
\psi_{k_1}(0) & \psi_{k_1}(1) & \psi_{k_2}(0) & \psi_{k_2}(1) & \psi_{k_3}(0) & \psi_{k_3}(1) & \psi_{k_4}(0) & \psi_{k_4}(1) \\
\psi_{k_1}(2) & \psi_{k_1}(2) & \psi_{k_2}(2) & \psi_{k_2}(2) & \psi_{k_3}(2) & \psi_{k_3}(2) & \psi_{k_4}(2) & \psi_{k_4}(2) \\
\psi_{k_1}(3) & \psi_{k_1}(3) & \psi_{k_2}(3) & \psi_{k_2}(3) & \psi_{k_3}(3) & \psi_{k_3}(3) & \psi_{k_4}(3) & \psi_{k_4}(3)
\end{bmatrix}
\begin{bmatrix}0 \\
0 \\
X(k_3) \\
X(k_4)
\end{bmatrix}.
$$

By subtracting of these two systems we get

$$0 = 
\begin{bmatrix}
\psi_{k_1}(0) & \psi_{k_2}(0) & \psi_{k_3}(0) & \psi_{k_4}(0) \\
\psi_{k_1}(1) & \psi_{k_2}(1) & \psi_{k_3}(1) & \psi_{k_4}(1) \\
\psi_{k_1}(2) & \psi_{k_2}(2) & \psi_{k_3}(2) & \psi_{k_4}(2) \\
\psi_{k_1}(3) & \psi_{k_2}(3) & \psi_{k_3}(3) & \psi_{k_4}(3)
\end{bmatrix}
\begin{bmatrix}
X(k_1) \\
X(k_2) \\
-X(k_3) \\
-X(k_4)
\end{bmatrix}.
$$

Since (10.5) holds then $X(k_1) = X(k_2) = X(k_3) = X(k_4) = 0$, meaning that the assumption about two independent pairs of solutions with two nonzero coefficients is not possible.

This approach to solve a problem (and to check the solution uniqueness) is illustrative, however not computationally feasible. For example, for a simple case with $N = 1024$ and just two nonzero coefficients, in order to find a solution we have to solve two times systems of equations (10.3) and (10.4) for each possible combination of $i$ and $k$ and to compare their solutions. Total number of combinations of two indices out of the total number of $N$ indices is

$$\binom{N}{2} \sim 10^6.$$

In order to check the solution uniqueness we should calculate a determinant value for all combinations of four indices $k_1, k_2, k_3$ and $k_4$ out the set of $N$ values. The number of determinants is $\left(\begin{array}{c}N \\ 4\end{array}\right) \sim 10^{12}$. If one determinant of the forth order is calculated in $10^{-5}$ [sec], then more than 5 days are needed to calculate all determinants for this quite simple case of two nonzero coefficients.

As a next example consider a signal described by a weighted sum of $K$ harmonics from a set of possible oscillatory functions $e^{j2\pi kn/N}$, $k = 0, 1, 2,$
where \( C \) are (different) constants.

More conservative bounds would be obtained if we used the spark definition. Then two minimization formulations produce the same solution if

\[
K < \frac{1}{2} \text{spark}(A) = \frac{1}{2} \left(1 + \frac{1}{\mu}\right)
\]

where \( \mu \) is the coherence index of the measurement matrix \( A \). It is easy to show that this condition is the same as the requirements that the measurement matrix \( A \) (its \( A_{2K} \) submatrices) with coherence \( \mu \) satisfies the restricted isometry property with \( \delta_{2K} = (2K - 1)\mu \)

\[
\left| \frac{1}{E_A} \left\| AX \right\|_2^2 - \left\| X \right\|_2^2 \right| \leq (2K - 1)\mu \left\| X \right\|_2^2.
\]

**Figure 10.19** Minimization using the \( \ell_1 \)-norm and the solution illustration for the case when the measurements line corresponds to noisy data.
The condition that
\[ \delta_{2K} = (2K - 1)\mu < 1 \]
is the same as \( K < \frac{1}{2}(1 + 1/\mu) \). Note that \( \delta_{2K} = (2K - 1)\mu \) is just a bound of \( \delta_{2K} \). For a matrix \( A \) there could be a lower and less restrictive constant satisfying the restricted isometry property.

In an ideal case the matrix \( (A_K^T A_K)^{-1} \) should be identity matrix for any combination of \( K \) columns. It means that the lines are with vector coordinate 1 in each direction. Reconstruction condition would be always satisfied. The transformation \( y = AX \) would correspond to a rotation on a sphere with all axis 1. Each \( X(0), X(1), X(2) \) would be transformed as \( y = AX \) keeping its amplitude. Since this is not the case then the transform \( y = AX \) will change amplitudes in addition to the rotation. For matrix \( A \) (not square matrix) the maximal gain of vector \( X \) is obtained in a direction defined by the maximal eigenvector. In reality

\[
X_K = (A_K^T A_K)^{-1} A_K^T y = (A_K^T A_K)^{-1} X_0
\]

\[
\|X_K\|^2 \leq \frac{1}{d_{\min}^2} \|X_0\|^2
\]

with \( d_{\min}^2 = (1 - \delta_k)^2 \). The condition \( \delta_2 < \sqrt{2} - 1 \) would here mean that \( 1/d_{\min}^2 > 0.343 \). It has been assumed that \( E_A = 1 \).

10.6 MEDIAN BASED FORMULATION

Illustrative explanation of the \( \ell_1 \)-norm based minimization can be presented on a slightly more general case with \( N \)-dimensional signal. Consider the case with \( M = N - 1 \) measurements. All coefficients \( X(k) \) can be expressed as a function of one coefficient, for example, \( X(0) \). For common signal transforms, when a complete set of measurements exists, there is only one degree of freedom in the minimization and the measurements are on a line in the \( N \) dimensional space. First assume ideal directions of measurement line

\[
p_{X(0)} = p_{X(1)} = \ldots = p_{X(N-1)} = 1.
\]
Then

\[
X(1) = X(0) - b_1 \\
X(2) = X(0) - b_2 \\
\vdots \\
X(N-1) = X(0) - b_{N-1}
\]

(10.71)

where \(b_i\) are unknown coefficients. Cost function for minimization is

\[
z = \|X\|_1 = |X(0)| + |X(0) - b_1| + \ldots + |X(0) - b_{N-1}|
\]

(10.72)

The solution of this minimization problem is the median

\[
|X(0)| = \arg\{\min\{z\}\} = \text{median}\{0, b_1, \ldots, b_{N-1}\}.
\]

Then for any value of coefficients \(b_i\) at least one \(X(k)\) will be equal to zero, since at least one of the elements in \(z\) is zero. It means that the solution will be of sparsity \(K = N - 1\) at least.

In order to prove that the median produces position of (10.72) minimum assume that the total number of terms \(N\) is an odd number. Function \(z\) in (10.72) is a sum of the functions of form \(|x - a|\). The rate (derivative) of these functions is +1 for \(x > a\) and −1 for \(x < a\). If there are \(N\) terms, as in (10.72), then the rate of function \(z\) will be \(+N\) for \(x \to \infty\). Going now back from \(x \to \infty\) toward the term with largest shift, the rate will remain \(+N\). At the position of the largest shift, the rate of this term will change from +1 to −1 meaning that the overall rate of \(z\) will be reduced to \(+,(N-2)\).

By passing each term, the rate will be reduced for additional factor of 2. It means that after the \(k\)th term the rate will be \((N-2k)\). The rate of \(z\) will change its sign when \((N-2k) = -1\). This will be the position of function \(z\) minimum. It is \(k = (N+1)/2\) and it corresponds to the middle coefficient positions, i.e., to the median of coefficients (shifts).

**Example 10.21.** As an example consider the case with \(N = 7\) and \(M = 6\) measurements \(AX = y\) producing an ideal line in a seven-dimensional space of the form (10.71), with \(b_1 = 0.7, b_2 = 0.2, b_3 = -0.5, b_4 = 1, b_5 = 0.8,\) and \(b_6 = -0.9\).

For the data presented in Fig.10.20 the solution is \(|X(0)| = \arg\{\min\{z\}\} = \text{median}\{0, 0.7, 0.2, -0.5, 1, 0.8, -0.9\} = 0.2\) with the coefficient corresponding to \(X(2) = X(0) - 0.2 = 0\) being equal to zero.

If the signal sparsity is \(K < N/2\) then there will exist more than \(N/2\) values \(b_i = b\) such that \(|X(0) - b_i| = 0\). The solution of minimization problem then will not depend on other \(b_k \neq b_i = b\) and will be unique.
Figure 10.20  Median as the solution of minimization problem.
arg\{\min\{z\}\} = \text{median}\{0, b_1, ..., b_i, b_{i+1}, ..., b_{N-1}\} = b_i = b. Therefore for one missing sample \(M = N - 1\) the solution is unique for signals whose sparsity is \(K < N/2\).

If the directions are not ideal but \(p_{X(0)} = a_0, p_{X(1)} = a_1, ..., p_{X(N-1)} = a_{N-1}\) then a form corresponding to the weighted median appears. For \(N = 2P + 1\) a weighted median produces the same result as the unweighted median if a sum of the smallest \(P + 1\) coefficient values is greater than a sum of its \(P\) largest values

\[
\sum_{i=0}^{P} |a_i| > \sum_{i=P+1}^{N-1} |a_i|.\
\]

For \(P = 1\) we get

\[
|a_0| + |a_1| + |a_2| - \max\{|a_0|, |a_1|, |a_2|\} > 1.
\]

This relation corresponds to the thick line for the \(\ell_1\)-norm in Fig. 10.16 (top) with \(|a_0| = ||p_{X(0)}|| = \max\{|a_0|, |a_1|, |a_2|\}\) and

\[
\frac{|p_{X(1)}| + |p_{X(2)}|}{|p_{X(0)}|} > 1.
\]

Consider next the case when two degrees of freedom exist (with \(M = N - 2\) measurements). All coefficients \(X(k)\) can be expressed as a function of, for example, \(X(0)\) and \(X(1)\) as

\[
X(2) = a_{2,0}X(0) + a_{2,1}X(1) - b_2,
\]

... 

\[
X(N-1) = a_{N-1,0}X(0) + a_{N-1,1}X(1) - b_{N-1}.
\]

Then

\[
z = ||X(k)||_1 = |X(0)| + |X(1)| + |a_{20}X(0) + a_{21}X(1) - b_2| + ... + |a_{N-1,0}X(0) + a_{N-1,1}X(1) - b_{N-1}|
\]

The solution of the minimization problem is a two-dimensional median. It is a point in \(X(0), X(1)\) plane such that a sum of absolute distances from the
lines

\[
\begin{align*}
X(0) &= 0 \\
X(1) &= 0 \\
a_{2,0}X(0) + a_{2,1}X(1) - b_2 &= 0 \\
&\vdots \\
a_{N-1,0}X(0) + a_{N-1,1}X(1) - b_{N-1} &= 0
\end{align*}
\]

is minimal\(^3\). Median here is not so simple as in the one-dimensional case. Various algorithms have been proposed for multidimensional (multivariate or spatial) median form. Note that by crossing a line \(a_{i,0}X(0) + a_{i,1}X(1) - x_i = 0\) we will always either increase or reduce the rate of the function \(z\), as in one dimensional case.

An illustration of signal with \(N = 6\) is presented in Fig.10.21. Value of \(z\) is presented, along with measurements lines, for the case of two degrees of freedom (two dimensional variable space). From this figure we can see that the number of measurements is \(M = 4\) and the sparsity of signal is \(K = 2\) since the distance of the function \(z\) minimum point from four planes is 0. There are two nonzero distances (to the thick black lines) meaning that there are two nonzero coefficients \(X(k)\). It is interesting that in this case the marginal median (minimization along axes \(X(0)\) and \(X(1)\) independently would produce the same result, since one of the zero values is on the axis).

For any value of variables at least two \(X(k)\) will be equal to zero, since at least two of the elements in \(z\) are zero. It means that, in general, the solution will be of sparsity \(K = N - 2\) at least.

In the case of \(M\) measurements the system

\[AX = y\]

contains \(M\) equations with \(N\) unknowns. It means that there are \(N - M\) free variables, while \(M\) can be calculated based on the free variables. Let us

\(^3\) Distance of a plane \(ax + by + c = 0\) from a point \(x_0, y_0\) is

\[
d_0 = \frac{|ax_0 + by_0 + c|}{\sqrt{a^2 + b^2}} = |ax_0 + by_0 + c|
\]

if \(\sqrt{a^2 + b^2} = 1\).
denote \( M \) unknowns \( X(k) \) by vector \( X_M \). Then it can be written

\[
\begin{bmatrix}
    x(n_1) \\
x(n_2) \\
    \vdots \\
x(n_M)
\end{bmatrix} = \begin{bmatrix}
    \psi_0(n_1) & \psi_1(n_1) & \psi_{N-M-1}(n_1) \\
    \psi_0(n_2) & \psi_1(n_2) & \psi_{N-M-1}(n_2) \\
    \psi_0(n_M) & \psi_1(n_M) & \psi_{N-M-1}(n_M)
\end{bmatrix} + \begin{bmatrix}
    X(0) \\
    X(1) \\
    \vdots \\
    X(N-M) \\
    X(N-M+1) \\
    \vdots \\
    X(N-1)
\end{bmatrix}
\]
\[ y = B_{N-M} X_{0,N-M-1} + C_M X_{N-M,N-1} \]
\[ X_{N-M,N-1} = C_M^{-1} y - C_M^{-1} B_{N-M} X_{0,N-M-1} \]

where \( X_{0,N-M-1} \) is the vector of free variables \( X(0), X(1), \ldots, X(N - M - 1) \), with corresponding measurement coefficients \( C_{N-M}^{-1} B_{N-M} \). Then the minimization problem can be written as

\[
z = \| X(k) \|_1 = |X(0)| + |X(1)| + \ldots + |X(N - M - 1)| + \left\| C_M^{-1} y - C_M^{-1} B_{N-M} X_{0,N-M-1} \right\|_1.
\]

It is reduced to \((N - M)\)-dimensional median formulation over variables \( X(0), X(1), \ldots, X(N - M - 1) \). Note that the multidimensional median calculation is not simple as in the case of one-dimensional problem.

### 10.7 NORM-ONE BASED RECONSTRUCTION ALGORITHMS

In the \( \ell_1 \)-norm based reconstructions the problem is formulated as

\[
\min \| X \|_1 \quad \text{subject to} \quad y = AX
\]

where \( \| X \|_1 = \sum_{k=0}^{N-1} |X(k)| \). The problem can be formulated in Lagrangian form

\[
F(X) = \| y - AX \|_2^2 + \lambda \| X \|_1
\]

where \( F(X) \) is the function to be minimized.

Reformulation of the problem in a constrained form reads

\[
\min \| y - AX \|_2^2 \quad \text{subject to} \quad \| X \|_1 < \varepsilon
\]

or

\[
\min \| X \|_1 \quad \text{subject to} \quad \| y - AX \|_2^2 < \varepsilon,
\]

where \( \varepsilon \) is sufficiently small parameter.

There are many ways to solve the stated problem, based on the constrained or Lagrangian form. Many of them are developed within the regression theory. Here we will present just one of them, based on the absolute selection and shrinkage operator (LASSO) formulation and Lagrangian minimization form.
10.7.1 LASSO- Minimization

The $\ell_1$-norm based minimization can be formulated as the minimization of $y - AX$ with a condition imposed on $X$.

The standard ridge formulation within the regression framework would minimize the error

$$\|y - AX\|_2^2 = (y - AX)^T (y - AX)$$

$$= \|y\|_2^2 - X^TA^Ty - y^TAX + X^TA^TAX$$

subject to the minimal energy values of $X$, i.e. subject to $\|X\|_2^2$. The minimization of the ridge constraint problem can be reformulated in Lagrangian form using a parameter $\lambda$ as

$$X = \arg\min_X \left\{ \|y - AX\|_2^2 + \lambda \|X\|_2^2 \right\}.$$

Minimization of

$$F(X) = \|y - AX\|_2^2 + \lambda \|X\|_2^2$$

$$= \|y\|_2^2 - X^TA^Ty - y^TAX + X^TA^TAX + \lambda X^TX$$

can be obtained in a closed form using the symbolic derivative operator

$$\frac{\partial F(X)}{\partial X^T} = -2A^Ty + 2A^TAx + 2\lambda X = 0$$

as

$$X_{\text{ridge}} = \left( A^TA + I\lambda \right)^{-1} A^Ty.$$

Parameter $\lambda$ balances the error and constraint. Its inclusion makes that the inversion is nonsingular even if $A^TA$ is singular. Real valued matrix $A$ is assumed, otherwise Hermitian conjugate and transpose $A^H$ would be used.

The standard ridge regression minimizes the energy of solution $X(k)$ and not its sparsity, Fig.10.22. That is the reason while the $\ell_1$-norm constraint is introduced in the cost function

$$F(X) = \|y - AX\|_2^2 + \lambda \|X\|_1$$

$$= \|y\|_2^2 - X^TA^Ty - y^TAX + X^TA^TAX + \lambda X^T\text{sign}(X)$$

with the LASSO minimization problem formulation

$$X = \arg\min_X \left\{ \|y - AX\|_2^2 + \lambda \|X\|_1 \right\}.$$
Figure 10.22 Minimization with constraint: in ridge regression (left), LASSO regression (middle), and the $\ell_{1/4}$-norm being a function closer to the $\ell_0$-norm.

Function $\|X\|_1$ promotes sparsity. It produces the same results (under certain conditions) as if $\|X\|_p$, with $p$ close to 0, is used, Fig. 10.22.

10.7.1.1 Iterative Calculation

The minimization problem with the $\ell_1$-norm constraint does not have a close form solution. It is solved in iterative ways. In order to define an iterative procedure we will add a nonnegative term, having zero value at the solution $X_s$ of the problem,

$$G(X) = (X - X_s)^T(\alpha I - A^T A)(X - X_s),$$

to the function $F(X)$. This term will not change the minimization solution. New function is

$$H(X) = F(X) + (X - X_s)^T(\alpha I - A^T A)(X - X_s).$$

where $\alpha$ is such that the added term is always nonnegative. It means $\alpha > \lambda_{\text{max}}$, where $\lambda_{\text{max}}$ is the largest eigenvector of $A^T A$. Gradient of $H(X)$ is

$$\nabla H(X) = \frac{\partial H(X)}{\partial X} = -2A^Ty + 2A^TA X + \lambda \text{sign}(X) + 2(\alpha I - A^T A)(X - X_s).$$

Solution of $\nabla H(X) = 0$ is

$$-A^Ty + \frac{\lambda}{2} \text{sign}(X) - (\alpha I - A^T A)X_s + \alpha X = 0$$

$$X + \frac{\lambda}{2\alpha} \text{sign}(X) = \frac{1}{\alpha} A^T(y - AX_s) + X_s.$$
Corresponding iterative relation is of the form

\[ X_{s+1} + \frac{\lambda}{2\alpha} \text{sign}(X_{s+1}) = \frac{1}{2\alpha} A^T(y - AX_s) + X_s. \]

Note that the solution of scalar equation

\[ x + \lambda \text{sign}(x) = y \]

is obtained using soft-thresholding rule defined by a function \( \text{soft}(y, \lambda) \) as

\[
x = \text{soft}(y, \lambda) = \begin{cases} 
  y + \lambda & \text{for } y < -\lambda \\
  0 & \text{for } |y| \leq \lambda \\
  y - \lambda & \text{for } y > \lambda 
\end{cases}
\]

or

\[
\text{soft}(y, \lambda) = \text{sign}(y) \max\{0, |y| - \lambda\}.
\]

The same rule can be applied to each coordinate of vector \( X_{s+1} \),

\[
X_{s+1} = \text{soft}\left( \frac{1}{\alpha} A^T(y - AX_s) + X_s, \frac{\lambda}{2\alpha} \right)
\]

or

\[
X(k)_{s+1} = \text{soft}\left( \frac{1}{\alpha} (a(k) - b(k)) + X(k)_s, \frac{\lambda}{2\alpha} \right)
\]

where \( a(k) \) and \( b(k) \) are coordinates of vectors \( a \) and \( b \) defined by \( a = A^T y \) and \( b = A^T AX_s \).

This is the iterative soft-thresholding algorithm (ISTA) for LASSO minimization. It can be easily modified to improve convergence to fast ISTA (FISTA). Note that this is just one of possible solutions of the minimization problem with the \( \ell_1 \)-norm.

The Lagrangian constant \( \lambda \) is a balance between the error and the \( \ell_1 \)-norm value, while \( \alpha = 2 \max \{\text{eig}(A^TA)\} \) is commonly used. The algorithms that solve this kind of problem are implemented as functions \( X = \text{lasso}(A, y) \).

**Example 10.22.** Measurement matrix \( A \) is formed as a Gaussian random matrix of the size \( 40 \times 60 \). Since there are 40 measurements the random variable \( N(0, \sigma^2) \) with \( \sigma^2 = 1/40 \) is used. The original sparse signal of the total length \( N = 60 \) is \( X(k) = \delta(k - 5) + 0.5\delta(k - 12) + 0.9\delta(k - 31) - 0.75\delta(k - 45) \) in the transformation domain. It is measured with a matrix \( A \) with 40 measurements stored in vector \( y \). All 60 signal values are reconstructed using these 40 measurements \( y \) and the matrix \( A \), in 1000 iterations. In
the initial iteration $X_0 = 0$ is used. Then for each next $s$ the new values of $X$ are calculated using (10.73), given data $y$ and matrix $A$. Value of $\alpha = 2\max\{\text{eig}\{A^TA\}\}$ is used. The results for $\lambda = 0.01$ and $\lambda = 0.0001$ are presented in Fig. 10.23. For very small $\lambda = 0.0001$ the result is not sparse, since the constraint is too weak.

10.7.2 Signal Domain Reconstruction with a Gradient Algorithm

It is shown that the sparse signal reconstruction can be formulated as a constrained minimization problem. The sparsity measure is minimized having in mind constraints defined by available samples/measurements). If a complete set of samples/measurements can be defined then the signal reconstruction can be formulated as a minimization problem where the missing samples/measurements $y_c$ are considered as minimization variables, while available samples/measurements $y$ remain unchanged. The simplest way to solve this problem is in changing all missing samples within the range of their possible values and then to select the combination of their values which produced the minimal sparsity measure. This method has been illustrated in Example 10.15 on a signal with two missing samples. However, when the number of missing samples is large, then a direct search over all missing sample values cannot be used due to its high calculation complexity.

Minimization of the sparsity measure $M$ can be implemented with gradient descent (or steepest descent) method instead of using a direct
search over missing sample values. Minimum sparsity measure position is
determined through an iterative procedure

\[ y^{(m+1)}_c = y^{(m)}_c - \alpha \left. \frac{\partial M}{\partial y_c} \right|_{y_c = y^{(m)}_c} \]

where \( y^{(m)}_c \) is the vector of missing samples in the \( m \)th iteration and \( M \) is
the sparsity measure. Gradient of sparsity measure calculated at
\( y_c = y^{(m)}_c \) is denoted by \( \left. \frac{\partial M}{\partial y_c} \right|_{y_c = y^{(m)}_c} \), while \( \alpha \) is the iteration step. For the algorithm
convergence a convex measure function is required.

A signal \( x(n) \) that is sparse in a transformation domain \( X(k) = \mathbb{T}\{x(n)\} \) is used for illustration. As in Example 10.15 it has been assumed
that two samples \( x(n_{N-1}) \) and \( x(n_N) \) are not available, \( y_c = (x(n_{N-1}), x(n_N)) \). Signal \( x_d(n) \) is formed. Its values at the available sample positions
\( y = (x(n_1), x(n_2), ..., x(n_M)) \), \( M = N - 2 \), are considered as constants. Samples \( x(n_{N-1}) \) and \( x(n_N) \) at the positions \( q_1 = n_{N-1} \) and \( q_2 = n_N \) are considered as variables. For various values of \( x(n_{N-1}) \) and \( x(n_N) \) the sparsity
measure of \( x_a(n) \) is calculated as \( M = \|\mathbb{T}[x_a(n)]\|_1 = \|X_a\|_1 \) and presented
in Fig. 10.24, along with illustration of the gradient \( \left. \frac{\partial M}{\partial y_c} \right|_{y_c = 0} \)
coordinates at \( x(n_{N-1}) = 0, x(n_N) = 0 \).

Consider a signal \( x(n) \) with available samples at \( n \in M \). Signal is
sparse in a transformation domain \( X(k) = \mathbb{T}\{x(n)\} \). The DFT will be used
as a study case, \( X(k) = \text{DFT}[x(n)] \).

As the initial estimate of reconstructed signal \( x_a^{(0)} \) we will use values
that would follow as a result of the \( \ell_2 \)-norm based minimization of the signal
transform. Values of \( x_a^{(0)} \) are

\[ x_a^{(0)}(n) = \begin{cases} 0 & \text{for missing samples, } n \in N_Q \\ x(n) & \text{for available samples, } n \in M \end{cases} \]

where \( N_Q \) is the set of missing sample positions. The available samples are
considered as constants, while the missing samples are changed through
iterations. Denote by \( x_a^{(m)} \) the values of the signal reconstructed after \( m \)
iterations. The minimization process can be described as

\[ \min \|X_a\|_1 \quad \text{subject to } x_a^{(m)}(n) = x(n) \text{ for } n \in M \]

where \( X_a(k) = \text{DFT}[x_a^{(m)}(n)] \). Since the task is to find the position of function \( z = \|X_a\|_1 \) minimum, through an iterative procedure, the relation for
missing samples calculation can be defined by using the gradient of sparsity measure

\[
y_c^{(m+1)} = y_c^{(m)} - \frac{1}{N} \frac{\partial \|X_a\|_1}{\partial y_c} y_c = y_c^{(m)}
\]

(10.74)

where \(y_c^{(m)}\) is the vector of variables (missing signal sample values) in the \(m\)th iteration, Fig. 10.13. Factor 1/N is introduced for the DFT analysis so that coefficients \(X(k)\) are equal to the signal amplitudes in time. The coordinates of gradient vector \(g(n_i) = \partial \|X_a\|_1 / (\partial y_c N)\) in the \(m\)th iteration can be estimated using finite differences of the sparsity measure calculated for each variable (missing sample) \(n_i \in \mathbb{N}_Q\)

\[
g(n_i) = \frac{\|X_{a+}\|_1 - \|X_{a-}\|_1}{2\Delta N}
\]
where

$$X_a^+(k) = \mathbb{T}\{x_a^+(n)\}$$
$$X_a^-(k) = \mathbb{T}\{x_a^-(n)\}$$

and

$$x_a^+(n) = x_a^m(n) + \Delta \delta(n - n_i)$$
$$x_a^-(n) = x_a^m(n) - \Delta \delta(n - n_i).$$

For \(n_i \in M\) there are no changes of the signal values, \(g(n_i) = 0\). A parameter for finite difference calculation is denoted by \(\Delta\). All \(g(n)\) values form vector denoted by \(G_m\) with elements \(G_m(n)\). The minimum of sparsity measure is obtained when all unavailable samples are equal to the values of the original signal values, i.e., when the signal is reconstructed (assuming that the recovery conditions are satisfied).

### 10.7.2.1 Finite Difference Step

Before presenting the algorithm, the basic idea and parameters in (10.74) will be discussed. Assume first a simple case when a single signal sample at \(n_0 \in \mathbb{N}_Q\) is not available, with \(\text{card}\{M\} = N - 1\). This sample is considered as variable. It may assume an arbitrary signal value \(x_a(n_0) = x(n_0) + z(n_0)\), where \(z(n_0)\) is a variable representing shift from the true signal value at \(n_0\).

In order to estimate the finite difference of the sparsity measure

$$\|X_a\|_1 = \sum_{k=0}^{N-1} |X_a(k)|,$$

due to the change of variable \(z(n_0)\), form the signals

$$x_a^+(n) = x(n) + (z(n) + \Delta) \delta(n - n_0)$$
$$x_a^-(n) = x(n) + (z(n) - \Delta) \delta(n - n_0),$$

where \(\Delta\) is a parameter. The finite difference of the sparsity measure is

$$g(n_0) = \frac{\|X_a^+\|_1 - \|X_a^-\|_1}{2N\Delta}.$$
The pulses $\delta(n - n_0)$ are uniformly spread over all frequencies in the DFT domain. Then

$$X_a^+(k) = X(k) + (z(n_0) + \Delta)e^{j2\pi n_0 k / N}$$

$$X_a^-(k) = X(k) + (z(n_0) - \Delta)e^{j2\pi n_0 k / N}$$

holds. Since the signal is sparse ($K \ll N$) in a rough analysis we may neglect changes in a few nonzero values of $X(k)$. We may approximately write

$$\|X_a^+\|_1 = \sum_{k=0}^{N-1} |X_a^+(k)| \approx \mu + |z(n_0) + \Delta| N$$

$$\|X_a^-\|_1 = \sum_{k=0}^{N-1} |X_a^-(k)| \approx \mu + |z(n_0) - \Delta| N,$$  

where $\mu = \|X\|_1$ is the sparsity measure of the original signal $x(n)$. Therefore the gradient approximation of the sparsity measure $\|X_a\|_1$ along the direction of variable $z(n_0)$ is

$$g(n_0) = \frac{\|X_a^+\|_1 - \|X_a^-\|_1}{2N\Delta} \approx \frac{|z(n_0) + \Delta| - |z(n_0) - \Delta|}{2\Delta}.$$}

For deviations from the true signal value smaller than the step $|z(n_0)| < \Delta$ we get

$$g(n_0) \approx \frac{z(n_0)}{\Delta} \sim z(n_0). \quad (10.75)$$

It means that the gradient value can be used as an indicator of the signal value deviation from the correct value (this property will be later used for detection of impulsive noise in signal samples as well). For a large $|z(n_0)| > \Delta$

$$g(n_1) \approx \frac{1}{2} \text{sign}(z(n_0)). \quad (10.76)$$

In that case the gradient assumes correct direction toward minimum positions, with a deviation independent intensity.

In order to analyze the influence of $\Delta$ to the solution precision, when $z(n_0)$ is very small, assume that we have obtained the exact solution and that the change of sparsity measure is tested on the change of sample $x(n_0)$ for $\pm\Delta$. Then for a signal $x(n) = \sum_{i=1}^{K} A_i e^{j2\pi n_0 k_i / N}$ of sparsity $K$ the DFTs of
\( x^+(n) = x(n) + \Delta \delta(n - n_0) \) and \( x^-(n) = x(n) - \Delta \delta(n - n_0) \) are

\[
\|X^+_a\|_1 = \sum_{i=1}^{K} |A_i| + \Delta e^{-j2\pi n_0 k_i / N} + (N - K)\Delta
\]

\[
\|X^-_a\|_1 = \sum_{i=1}^{K} |A_i| - \Delta e^{-j2\pi n_0 k_i / N} + (N - K)\Delta.
\]

For the worst case analysis, assume that \( A_i \) are in phase with \( e^{-j2\pi n_0 k_i / N} \) and \( \Delta \leq |A_i| \) when

\[
\|X^+_a\|_1 = \sum_{i=1}^{K} |A_i| + K\Delta + (N - K)\Delta = \mu + N\Delta
\]

\[
\|X^-_a\|_1 = \sum_{i=1}^{K} |A_i| - K\Delta + (N - K)\Delta = \mu + (N - 2K)\Delta.
\]

where \( \mu = \|X\|_1 \). Therefore \( g(n_0) = (\|X^+_a\|_1 - \|X^-_a\|_1) / (2N\Delta) \neq 0 \). The correct signal value will not be a stationary state. The algorithm will move the solution from \( x(n_0) \) to \( x(n_0) + b \) in order to produce \( g(n_0) = 0 \) in the stationary point. Then \( \|X^+_a\|_1 = \mu + N(\Delta - b) \) is equal to \( \|X^-_a\|_1 = \mu_0 + (N - 2K)(\Delta + b) \). It means that the stationary point will be biased. The worst case bias \( b \) follows from

\[
N(\Delta - b) = (N - 2K)(\Delta + b)
\]

\[
b = \frac{K}{N - K}\Delta \simeq \frac{K}{N}\Delta \text{ for } K \ll N.
\]

The bias upper limit can be reduced by using very small \( \Delta \). However, calculation with a small \( \Delta \) would be time consuming (with many iterations). Efficient implementation can be done by using \( \Delta \) of an order of signal amplitude in the initial iteration. When the algorithm reaches a stationary point, with a given \( \Delta \), the value of mean squared error will assume its almost constant value. The error will be changing the gradient direction around correct point only, for almost \( \pi \). This fact may be used as an indicator to reduce the step \( \Delta \), in order to approach the signal true value with a given precision. For example, if the signal amplitudes are of order of 1 and \( K/N = 0.1 \) taking \( \Delta = 1 \) in the first iteration will produce the solution with a precision better than 20 [dB]. Then, the step \( \Delta \) should be reduced, for example to \( \Delta = 0.1 \). A precision better than 40 [dB] would be obtained, and so on.

Through simulation study it has been concluded that appropriate step parameter value in (10.74) is related to the finite difference step as \( \alpha = 2\Delta \).
10.7.2.2 Algorithm

The presented analysis is used as a basic idea for the algorithm summarized as follows:

**Step 0:** Set \( m = 0 \) and form the initial signal estimate \( x_a^{(0)}(n) \) defined for \( n \in \mathbb{N} \) as

\[
x_a^{(0)}(n) = \begin{cases} 
0 & \text{for missing samples, } n \in \mathbb{N}_Q \\
x(n) & \text{for available samples, } n \in \mathbb{M}
\end{cases}
\]

(10.79)

where \( \mathbb{N} = \{0,1,\ldots,N-1\} \) and \( \mathbb{N}_Q = \mathbb{N} \setminus \mathbb{M} \) is the complement of \( \mathbb{M} \) with respect to \( \mathbb{N} \). The initial value for an algorithm parameter \( \Delta \) is estimated as

\[
\Delta = \max_{n \in \mathbb{M}} |x(n)|.
\]

(10.80)

**Step 1:** Set \( x_p(n) = x_a^{(m)}(n) \). This signal is used in Step 3 in order to estimate reconstruction precision.

**Step 2.1:** Set \( m = m + 1 \). For each missing sample at \( n_i \in \mathbb{N}_Q \) form signals \( x_a^+(n) \) and \( x_a^-(n) \):

\[
x_a^+(n) = x_a^{(m)}(n) + \Delta \delta(n - n_i)
\]

\[
x_a^-(n) = x_a^{(m)}(n) - \Delta \delta(n - n_i).
\]

(10.81)

**Step 2.2:** Estimate differential of the signal transform measure

\[
g(n_i) = \frac{\sum_{k=0}^{N-1} |X_a^+(k)| - \sum_{k=0}^{N-1} |X_a^-(k)|}{2N\Delta}
\]

(10.82)

where \( X_a^+(k) = \mathcal{T}\{x_a^+(n)\} \) and \( X_a^-(k) = \mathcal{T}\{x_a^-(n)\} \) are transforms of \( x_a^+(n) \) and \( x_a^-(n) \).

**Step 2.3:** Form a gradient vector \( G_m \) with the same length as the signal. At the positions of available samples \( n \in \mathbb{M} \), this vector has value \( G_m(n) = 0 \). At the positions of missing samples \( n \in \mathbb{N}_Q \) its values are \( G_m(n) = g(n) \), calculated by (10.82).

**Step 2.4:** Correct the values of estimated signal \( y_a(n) \) iteratively by

\[
x_a^{(m)}(n) = x_a^{(m-1)}(n) - \alpha G_m(n),
\]

(10.83)

where the step parameter \( \alpha = 2\Delta \) is commonly used.

**Step 2.5:** Calculate angle \( \beta_m \) between successive gradients as

\[
\beta_m = \arccos \frac{\sum_{k=0}^{N-1} G_{m-1}(k) G_m(k)}{\sqrt{\sum_{k=0}^{N-1} G_{m-1}^2(k)} \sqrt{\sum_{k=0}^{N-1} G_m^2(k)}}
\]
If angle $\beta_m$ is lower than $170^\circ$ and the maximal allowed number of iterations is not reached ($m < m_{\text{max}}$) go to **Step 2.1**.

**Step 3:** If the maximal allowed number of iterations is reached stop the algorithm. Otherwise calculate

$$T_r = 10 \log_{10} \frac{\sum_{n \in N_Q} |x_p(n) - x_a^{(m)}(n)|^2}{\sum_{n \in N_Q} |x_a^{(m)}(n)|^2}.$$  

Value of $T_r$ is an estimate of the reconstruction error to signal ratio, calculated for missing samples only. If $T_r$ is above the required precision threshold (for example, if $T_r > -100dB$), the calculation procedure should be repeated with smaller $\Delta$. For example, set new $\Delta$ value as $\Delta/\sqrt{10}$ or $\Delta/10$ and go to **Step 1**.

**Step 4:** Reconstruction with the required precision is obtained in $m$ iterations or the maximal allowed number of iterations is reached. The reconstructed signal is $x_R(n) = x_a^{(m)}(n)$.

By performing presented iterative procedure, the missing values will converge to the true signal values, producing the minimal concentration measure in the transformation domain.

- The inputs to the algorithm are the signal length $N$, the set of available samples $M$, the available signal values $x(n_i), n_i \in M$, the required precision $T_{\text{max}}$, and maximal number of iterations.
- Instead of calculating signals (10.81) and their transforms for each $n_i \in N_Q$ we can calculate

$$|X_a^+(k)| = |X_a^{(m)}(k) + \Delta D_{n_i}(k)|$$

$$|X_a^-(k)| = |X_a^{(m)}(k) - \Delta D_{n_i}(k)|$$

with $X_a^{(m)}(k) = \mathcal{T}\{x_a^{(m)}(n)\}$ and $D_{n_i}(k) = \mathcal{T}\{\delta(n - n_i)\} = \exp(-j2\pi n_i k / N)$, for the DFT and each $n_i \in M$. Since $D_{n_i}(k)$ are independent of the iteration number $m$ they can be calculated independently from the DFT of the signal.

**Example 10.23.** Consider a signal

$$x(n) = 3\sin(20\pi \frac{n}{N})$$

with $N = 8$. Missing samples are $n \in N_Q = \{1, 6\}$. The signal is reconstructed using a simplified gradient based algorithm using **Step 0** to **Step 2.4**, from (10.79) to (10.83), in 60 iterations. The initial algorithm parameter $\Delta = 1$ and
the initial value of missing samples $x(1) = 0$ and $x(6) = 0$ are used. The values of missing samples in the first 20 iterations are presented by dots (connected by a line) in Fig.10.25. After about 6 iterations the algorithm with $\Delta = 1$ does not significantly change the missing sample values (zoomed changes are shown in lower subplot within the figure). Close to the stationary point obtained for $\Delta = 1$ the gradient coordinates are almost zero-valued (with direction changes for almost $\pi$), since the measures are on the contour with almost the same measure (circles). After the step is reduced to $\Delta = 0.1$ in the 20th iteration, the algorithm resumes its fast approach toward the exact value, until a new stationary state. With a new change of $\Delta$ to $\Delta = 0.01$ the approach is again continued.

The stationary state bias for $\Delta = 1$ is lower than $\frac{\pi}{\sqrt{2}}\Delta = 1/4$ (it corresponds to the bias caused MSE lower than 15.5 [dB]). By each reduction of $\Delta$ to $\Delta/10$ the bias caused MSE will be lower for 20 [dB]. The reconstruction result and the MSE for the estimated missing values $x(1)$ and $x(6)$ is presented in Fig.10.26.

The calculation is repeated with the signal

$$x(n) = 3\sin(20\pi \frac{n}{N}) + 2\cos(60\pi \frac{n}{N}) + 0.5\sin(46\pi \frac{n}{N})$$

and $N = 32$. Missing samples are $n \in \mathbb{N}_Q = \{2, 4, 5, 7, 9, 13, 17, 19, 24, 26, 28, 31\}$. The result for this case is shown in Fig.10.27.
10.7.2.3 Comments on the Algorithm

- In a gradient-based algorithm, a possible divergence is related to the algorithm behavior for large steps $\Delta$. Small steps influence the rate of the algorithm approach to the solution only (assuming that it exists). Here, we will examine the algorithm behavior for a large value of step $\Delta$. We can write

$$
|X_+^a(k)| - |X_-^a(k)| = |X_+^{(m)}(k) + \Delta D_{n_1}(k)| - |X_-^{(m)}(k) - \Delta D_{n_1}(k)|
$$

$$
= \Delta |D_{n_1}(k)| \left( \left| 1 + \frac{X_+^{(m)}(k)}{\Delta D_{n_1}(k)} \right| - \left| 1 - \frac{X_-^{(m)}(k)}{\Delta D_{n_1}(k)} \right| \right).
$$

Considering the complex number $a = X_+^{(m)}(k) / (\Delta D_{n_1}(k))$, with $|a| \ll 1$ for a large $\Delta$, from the problem geometry it is easy to show that the following bounds hold $0 \leq ||1 + a| - |1 - a|| \leq 2|a|$. Exact value of this expression depends on the phase of $a$. Therefore,

$$
0 \leq ||X_+^a(k)| - |X_-^a(k)|| \leq 2 \left| X_+^{(m)}(k) \right|.
$$

Lower limit 0 is obtained if $a$ is imaginary-valued, while the upper limit $2 \left| X_+^{(m)}(k) \right|$ follows if $a$ is real-valued.

It means that the value of the finite difference $|X_+^a(k)| - |X_-^a(k)|$, that is used to correct the missing signal samples, does not depend on the value of the step $\Delta$ if $\Delta$ is large. The missing signal values will be adapted for
a value independent on $\Delta$ in that case. The values of missing samples will oscillate within the range of the original signal values of order $|X_{a}(k)|/N$, until $\Delta$ is reduced in the iterations below the signal magnitude. Then the missing samples will start approaching to the position of the sparsity measure minimum. The initial values will be arbitrary changed within the signal amplitude order as far as $\Delta$ is too large. It will not influence further convergence of the algorithm, when the step $\Delta$ assumes appropriate values.

- Since two successive gradient vectors are required to calculate the gradient angle $\beta_m$, it is calculated starting from the second iteration for each $\Delta$.

- Algorithm output is the reconstructed signal $x_R(n), n = 0, 1, ..., N - 1$.

- Other signal transforms can be used instead of the DFT. The only requirement is that signal is sparse in that transform domain (two-dimensional DCT will be presented later).

**Figure 10.27** Gradient-based reconstruction of a sparse signal.
Example 10.24. Consider a signal

\[ x(t) = \sum_{i=1}^{K/2} A_i \cos(2\pi t k_i / T + \phi_i), \]  

(10.84)

with \( t = n\Delta t \), \( \Delta t = 1 \), and the total number of samples \( N = T/\Delta t \). The sparsity parameter \( K \) is changed from \( K = 2 \) to \( K = N/2 \). The amplitudes \( A_i \), frequencies \( k_i \), and phases \( \phi_i \) are taken randomly. Amplitude values are modeled as Gaussian random variables with variance 1, the frequency indices assume random numbers within \( 1 \leq k_i \leq N - 1 \), and the phases assume uniform random values within \( 0 \leq \phi_i \leq 2\pi \), in each realization. The reconstruction is performed by using 100 realizations for each \( K \) with random sets of missing \( Q = N - M \) samples in each realization. The reconstructed signals \( x_R(n) \) are obtained. The results are presented in Fig. 10.28 in a form of the signal-to-reconstruction-error ratio (SRR) in [dB]

\[ SRR = 10\log \frac{\sum_{n=0}^{N-1} |x(n)|^2}{\sum_{n=0}^{N-1} |x(n) - x_R(n)|^2}. \]  

(10.85)

Bright colors indicate the region where the algorithm had fully recovered missing samples in all realizations, while dark colors indicate the region where the algorithm could not recover missing samples in any realization. In the transition region for \( M \) slightly greater than \( 2K \) we have cases when the signal recovery is not achieved and the cases of full signal recovery. The simulations are done for \( N = 128 \) and for \( N = 64 \), Fig. 10.28(a), (b). A stopping criterion for the accuracy of 120 [dB] is used. It corresponds to a precision in the recovered signal of an input samples precision if they are acquired by a 20-bit A/D converter. The case with \( N = 64 \) is repeated with an additive input Gaussian noise such that the input signal-to-noise ratio is 20 [dB] in each realization Fig. 10.28(c). The reconstruction error in this case is limited by the input signal-to-noise value. The number of iterations to achieve the required precision is presented in Fig. 10.28(d). We can see that the number of iterations is well bellow 100 for the most important region where the reconstruction was achieved in all realizations (high values of \( M \) and small value of \( K \), \( M \gg K \)). The number of iterations is quite small in the region where the reconstruction can be achieved.

An illustration of the algorithm performance regarding to the SRR and the gradient angle \( \beta_m \) in one realization, with \( K = 6 \), is presented in Fig. 10.29. The algorithm reached 120 [dB] accuracy in 47 iterations. From the gradient angle graph we see that the algorithm step is reduced to \( \Delta/\sqrt{10} \rightarrow \Delta \) in about each 4 iterations. According to (10.77) the expected MSE improvement by each reduction of \( \Delta \) is \( 20\log(\sqrt{10}) = 10 \) [dB].

\[ \square \]
Figure 10.28 Signal-to-reconstruction-error (SRR) averaged over 100 realizations for various sparsity $K$ and number of available samples $M$: (a) The total number of samples is $N = 128$. (b) The total number of samples is $N = 64$. (c) With a Gaussian noise in the input signal, $SNR = 20$ [dB] and $N = 64$. (d) Number of iterations to reach the solution with the defined precision.

10.8 ON THE UNIQUENESS OF THE DFT OF SPARSE SIGNALS

In general, the reconstructed signal uniqueness is guaranteed if the restricted isometry property is used and checked with appropriate isometry constants. However, two problems exist in the implementation of this approach. For a specific measurement matrix it produces quite conservative bounds. In practice it would produce a large number of false alarms for nonuniqueness. In addition, uniqueness check with the restricted isometry property requires a combinatorial approach, which is an NP hard problem.
In the adaptive gradient-based method the missing samples (measurements) are considered as the minimization variables. The available samples values are known and fixed. The number of variables in the minimization process is equal to the number of missing samples/measurements in the observation domain. This approach is possible when the common signal transforms are the domains of signal sparsity. Then the missing and available samples/measurements form a complete set of samples/measurements.

The DFT will be considered here as the signal sparsity domain. The solution uniqueness is defined in the sense that the variation of the missing sample values cannot produce another signal of the same sparsity. In the case when the signal is already reconstructed then the uniqueness is checked in the sense that there is no other signal of the same or lower sparsity with the same set of available samples.

Consider a signal $x(n)$ with $n \in \mathbb{N} = \{0, 1, 2, ..., N - 1\}$. Assume that $Q$ of its samples at the positions $q_m \in \mathbb{N}_Q = \{q_1, q_2, ..., q_Q\}$ are missing/omitted. The signal is sparse in the DFT domain, with sparsity $K$. The reconstruction goal is to get $x(n)$, for all $n \in \mathbb{N}$ using available samples at

![Figure 10.29](image-url)
\( n \in \mathbf{M} = \mathbf{N} \setminus \mathbf{N}_Q \). A new signal of the form

\[
x_a(n) = x(n) + z(n)
\]

will be analyzed here. For the available signal positions \( n \in \mathbf{M} \) the value of \( z(n) \) is fixed \( z(n) = 0 \), while \( z(n) \) may take arbitrary value at the positions of missing samples \( n = q_m \in \mathbf{N}_Q = \{q_1, q_2, \ldots, q_Q\} \). If \( x(n) \) is a \( K \) sparse signal then the DFT of \( x_a(n) \) is

\[
X_a(k) = X(k) + Z(k)
\]

\[
= N \sum_{i=1}^{K} A_i \delta(k - k_{0i}) + \sum_{m=1}^{Q} z(q_m) e^{-j2\pi q_m k / N}.
\]

Positions of nonzero values in \( X(k) \) are \( k_{0i} \in \mathbf{K} = \{k_{01}, k_{02}, \ldots, k_{0K}\} \) with amplitudes \( X(k_{0i}) = NA_i \). The values of missing samples of \( x_a(n) = x(n) + z(n) \) for \( n \in \mathbf{N}_Q \) are considered as variables. The goal of reconstruction process is to get \( x_a(n) = x(n) \), or \( z(n) = 0 \) for all \( n \in \mathbf{N} \). This goal should be achieved by minimizing a sparsity measure of the signal transform \( X_a(k) \). Existence of the unique solution of this problem depends on the number of missing samples, their positions, and the signal form.

If a signal with the transform \( X(k) \) of sparsity \( K \) is obtained using a reconstruction method, with a set of missing samples, then the reconstruction \( X(k) \) is unique if there is no other signal of the same or lower sparsity that satisfies the same set of available samples (using the same set of missing samples as variables).

**Example 10.25.** Consider the simplest case of one missing sample at position \( n = q \). The signal sparsity is \( K \). Signal reconstruction is based on \( x_a(n) = x(n) + z \delta(n - q) \) where \( z \) indicates an arbitrary deviation from the true signal value, since the missing sample \( x(q) \) is considered as variable. The DFT of \( x_a(n) \) is

\[
X_a(k) = N \sum_{i=1}^{K} A_i \delta(k - k_{0i}) + ze^{-j2\pi qk/N}.
\]

The number of nonzero DFT coefficients is

\[
\text{card}\{X_a\} = \|X_a\|_0 = \sum_{i=1}^{K} |NA_i + ze^{-j2\pi qk_{0i}/N}|^0 + \sum_{i=K+1}^{N} |z|^0
\]
Possible sparsity of $X_a(k)$ is

$$
\|X_a\|_0 = \begin{cases} 
N \text{ for } |z| \neq 0 \text{ and } z \neq -NA_i e^{2\pi k_0 q/N} \text{ for any } i \\
N - 1 \text{ for } |z| \neq 0 \text{ and } z = -NA_i e^{2\pi k_0 q/N} \text{ for one } i \text{ only} \\
\cdots \\
N - K \text{ for } |z| \neq 0 \text{ and } z = -NA_i e^{2\pi k_0 q/N} \text{ for } i = 1, \ldots, K \\
K \text{ for } |z| = 0.
\end{cases}
$$

(10.86)

With just one missing value and arbitrary signal, the minimum of $\|X_a\|_0$ is achieved at $|z| = 0$ only if the signal sparsity is lower than the lowest possible sparsity with $|z| \neq 0$, $K < N - K$.

It means $K < N/2$. For $K = N/2$ the last two rows of (10.86) will produce the same result $N - K = N/2$ and $K = N/2$. In that case the minimum of $\|X_a\|_0$ is not unique. Note that this is true only if the considered signal $x(n)$ has a very specific form

$$
A_1 e^{2\pi k_0 q/N} = A_2 e^{2\pi k_0 q/N} = A_3 e^{2\pi k_0 q/N} = \cdots = A_K e^{2\pi k_0 q/N} = C. \quad (10.87)
$$

In reality the case that all components have equal amplitudes $|A_1| = |A_2| = |A_3| = \cdots = |A_K|$ and that the missing sample position $q$ is such that

$$
\arg \{A_1\} + 2\pi k_0 q/N = \arg \{A_2\} + 2\pi k_0 q/N = \cdots = \arg \{A_K\} + 2\pi k_0 q/N \quad (10.88)
$$

is a zero probability event.

It is interesting to note that if the last two conditions are satisfied by a signal $x(n)$ then the DFT coefficients from (10.87) are the frequency domain samples of a harmonic signal $B \exp(e^{2\pi k q/N})$, at $k \in \{k_{01}, k_{02}, \ldots, k_{06}\}$. Its IDFT is a delta pulse with the group delay at the position of missing sample

$$
\text{IDFT}\{B \exp(e^{2\pi k q/N})\} = B \delta(n - q).
$$

Example 10.26. Consider a signal $x(n)$ with $N = 32$ and two missing samples at $q_m \in \mathbb{N}_Q = \{3, 19\}$.

Signal sparsity is $K$. In order to simplify the notation assume that one DFT value of the reconstructed signal is $X(5) = 2$.

(a) Show that the limit for sparsity $K$ (when we can claim that the reconstructed sparse signal is unique, assuming that all signal amplitudes may assume arbitrary values) is $K < 8$.

(b) What are the properties that a signal must satisfy in the limit case $K = 8$ so that the solution is not unique?

(c) What is the sparsity limit if the missing samples are at $q_m \in \mathbb{N}_Q = \{5, 9\}$?
Figure 10.38  Sparsity limit probability distribution for the worst possible case of signal with $Q = 72$ out of $N = 128$ samples in 100,000 random realizations.

case signal with sparsity $K = 10$ is unique for $Q = 72$ is

**Probability** $[K = 10$ is unique, with $Q = 72] = 0.8723$.

### 10.11 IMAGE RECONSTRUCTION

The gradient based algorithm is applied on the image $x(n,m)$. As the transformation domain the two-dimensional DCT (in symmetric form) will be used

$$C(k,l) = v_k v_l \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} x(m,n) \cos \left( \frac{2\pi(2m + 1)k}{4N} \right) \cos \left( \frac{2\pi(2n + 1)l}{4N} \right),$$

where $v_0 = \sqrt{1/N}$ and $v_k = \sqrt{1/N}$ for $k \neq 0$. Assume that random set of pixels is available (not corrupted) at $(n,m) \in \mathbf{M}$. The goal is to reconstruct

---

This section is written by *Isidora Stanković.*
unavailable pixels. In order to apply a CS reconstruction algorithm, the image sparsity is assumed in the DCT domain. The DCT of an image is usually calculated by using 8x8 blocks. Most of the common images could be considered as sparse in the DCT domain without any additional processing. If we want to be sure that the original image, which will be processed, is sparse we can pre-process it by calculating the DCT of its 8x8 blocks and set the lowest amplitude coefficients to zero. By making the image sparse in the DCT domain we will not make a notable visual difference with respect to the original image.

Using the available pixels (measurements), an initial image is formed. It assumes the original image values at the positions of available pixels, while the missing pixels are set to zero (or arbitrary) value. This new image is defined as

\[ x^{(0)}_a(m,n) = \begin{cases} x(m,n) & \text{for } (n,m) \in \mathbf{M} \\ 0 & \text{for } (n,m) \in \mathbf{N}_Q \end{cases} \]

Note that for the missing pixels any value within the possible image values range can be assumed in the initial step. The algorithm will reconstruct the true image values at these positions. For graphical representation of missing pixels the value 255 corresponding to a white pixel or 0 will be used. Then the corrupted pixels are black or white pixels, Fig. 10.39.

For each missing sample signals \( x^+_a(m,n) \) and \( x^-_a(m,n) \) are formed:

\[
\begin{align*}
x^+_a(m,n) &= x^{(p)}(m,n) + \Delta \delta(m - m_i, n - n_i) \\
x^-_a(m,n) &= x^{(p)}(m,n) - \Delta \delta(m - m_i, n - n_i).
\end{align*}
\]

The finite difference of the signal transform measure is calculated

\[ g(m_i, n_i) = \frac{||C^+_a(k,l)||_1 - ||C^-_a(k,l)||_1}{2\Delta} \tag{10.107} \]

where \( C^+_a(k,l) = \text{DCT}[x^+_a(m,n)] \) and \( C^-_a(k,l) = \text{DCT}[x^-_a(m,n)] \).

A gradient matrix \( G_{m,n} \) is of the same size as the image. At the positions of available samples \((n,m) \in \mathbf{M}\), this matrix has zero value, \( G_{m,n} = 0 \). At the missing sample positions \( n \in \mathbf{N}_Q \) its values are \( G_{m,n} = g(m,n) \), calculated using (10.107).

The image values are corrected iteratively as

\[ x^{(p)}_a(m,n) = x^{(p-1)}_a(m,n) - 2\Delta G_{m,n}. \tag{10.108} \]

The change of step \( \Delta \) and the stopping criterion are the same as in one-dimensional case. The results in 50 iterations are shown in Fig. 10.39. Reconstructed image after 1, 3, and 50 iterations are presented.
Figure 10.39  Reconstruction of image using the gradient-based algorithm.
Index

Adaptive reconstruction, 765
Adaptive systems, 417
Allpass system, 239
Ambiguity function, 623
Analog signals, 10
Analytic part, 34
Antenna array, 467
Anticausal systems, 180
Attenuation, 50
Auto-regressive (AR), 179
Autocorrelation function, 330
Autocovariance function, 330
Backward difference, 220
Bandpass filter, 238
Bilinear transform, 224
Binomial random variable, 336
Blackman window, 528
Block LMS algorithm, 477
Born-Jordan distribution, 651
Butterworth filter, 47, 234
discrete-time, 231
Capon’s method, 608
local polynomial Fourier transform (LPFT), 614
short-time Fourier transform (STFT), 612
Cascade realization, 270
Causal system, 28, 61
Causal systems, 175
Central limit theorem, 338
Characteristic polynomial, 180
Choi-Willimas distribution, 657
Cohen class of distributions, 649
discrete form, 654
kernel decomposition, 656
Coherence, 678, 681
Complex sinusoidal signal, 16
Compressive sensing, 660
Continuous signals, 15
Convolution
circular, 109
continuous, 27
discrete-time, 59
in frequency domain, 33, 67
Cosine series, 23
Derivative
complex function, 20
Difference equation, 177, 180
Differential equation, 45
Differentiator, 67
Digital signals, 10
Direct realization I, 262
Direct realization II, 262
Dirichlet conditions, 18
Discrete Cosine transform (DCT), 134
Discrete Fourier transform (DFT), 101, 118
Discrete Hartley transform (DCT), 159
Discrete pseudo Wigner distribution, 633
Discrete Sine transform (DST), 137
Discrete system, 58
Discrete-time signals (discrete signals), 53
Displacement, 131
Downsampling, 570
Duality property, 30
Eigenvalues, 432
Eigenvectors, 432
Energy, 18, 57
Equiangular Tight Frame, 680
Ergodicity, 331
Error function, 339
Error signal
  adaptive system, 423
Fast Fourier transform, 120
  decimation-in-frequency, 120
  decimation-in-time, 122
Finite impulse response (FIR), 179
  frequency domain design, 285
  realization, 278
First-order statistics, 313
Fixed point arithmetic, 376
Floating point arithmetic, 384
  IEEE standard, 386
  mu-law and A-law, 386
Fourier series, 21, 35, 116
Fourier transform, 29, 35, 116
  matrix, 106
  of discrete-time signals, 61, 118
  properties, 31
  signum function, 31
Fractional Fourier transform, 607
  relation to the LPFT, 608
  windowed, 608
Frequency estimation, 130
Goertzel algorithm, 182
Gradient, 762
Gramm matrix, 686
Group delay, 40, 244, 279, 627
Haar transform, 146, 569
Hadamard transform, 145
Hamming window, 527
Hann(ing) window, 64, 525
Hartley series, 23
Highpass filter, 236
Hilbert transform, 34
Homogeneous equation, 181
Image reconstruction, 803
Impulse invariance method, 212
Impulse signal
  continuous (delta function), 16
  discrete-time, 54
Indirect measurements, 786
Infinite impulse response (IIR), 179
Initial condition
  continuous, 44
Instantaneous frequency, 36, 627
Interpolation, 114
Inverse system, 241
Isometry, 677
ISTA algorithm, 763
Kalman filter, 489
Kronecker delta function, 54
L-statistics, 325
Lagrangian, 760
Laplace transform, 42
LASSO minimization, 761
Leakage effect, 129
Linear adaptive adder, 421
Linear phase systems, 279
Linear system, 27
LMS algorithm, 451
  antenna systems, 467
  block, 477
  complex, 483
  convergence, 453
  echo cancellation, 473
  identification, 454
  noise cancelation, 458
  prediction, 464
  sign, 475
  sinusoidal disturbance, 462
  variable step, 479
Local polynomial Fourier transform, 604
  moments, 605
  relation to fractional Fourier transform, 607
Lowpass filter, 230
Magnitude, 18
Marginal properties, 647
Matched filter, 363
Matched z-transform method, 217
Measurement Matrix, 673
  Bernoulli Random, 676
  Gaussian Random, 675
  Indirect, 674
  Partial DFT, 675
  Structured Random, 676, 787
Median, 323, 350
Minimum phase system, 241
Moments
  LPFT, 606
  Morlet wavelet, 565
  Moving average (MA), 179
Moyal property, 629
MUSIC, 612

Narrowband signals
spectral estimation, 360

Neural networks, 495
activation function, 498
acyclic, 500
continuous output, 506
cyclic, 500
error backpropagation, 510
layer, 500
multilayer, 510
network function, 497
perceptron, 502
supervised, 502
unsupervised, 513
voting machines, 514

Neuron, 496
Noise, 334
binary, 335
complex Gaussian, 343
Gaussian, 338
Laplacian, 344
missing samples, 712
reconstruction, 723
uniform, 334

Noisy signal
Fourier transform, 346

Norm zero, 704
Norm-one, 729
ball, 742
Norm-zero, 671
Notch filter, 176, 206

Optimal filter, 366, 428
Orthogonality principle, 439
Overflow, 378

Parallel realization, 274
Parseval’s theorem, 68, 108, 676
Perceptron, 502
Period of a discrete signal, 56
Pisarenko method, 613
Power, 18, 57
Power spectral density, 332, 356
Probability, 320
density function, 320
Probability density function, 322
Quantization, 370

Random signals, 313
Rank of matrix, 698
Rayleigh distribution
uniform, 343
Reconstruction uniqueness, 777
Reconstruction uniqueness, 781
Rectangular window, 64
Recursive systems
adaptive, 488
Reduced interference distributions
discrete form, 651
Region of convergence, 164
Resolution, 527
Restricted isometry, 677, 681
canonical, 677, 682
eigenvalues, 685
uniqueness, 695
Ridge regression, 761
RLS algorithm
variable step, 483
S-method, 641
S-transform (the Stockwell transform), 601
Sampling
nonuniform, 786
Sampling theorem, 95
for periodic signals, 124
in the frequency domain, 35
in the time domain, 71
Schwartz’s inequality, 680
Second-order statistics, 330
Sensitivity of system, 265
Short-time Fourier transform (STFT), 516
discrete, 532
discrete-time, 529
filter bank, 536
frequency-varying, 561
hybrid, 563
inversion, 523, 540
optimal window, 519
optimisation, 554
overlapping, 538
recursive, 535
time-varying, 550
Sign LMS algorithm, 475
Sinc distribution
discrete form, 651
Soft-thresholding, 763
Spark of matrix, 698
coherence, 754
uniqueness, 700
Sparse signals, 660
Sparsity, 670
Stable system, 28, 61
Stable systems, 175
Standard deviation, 325
Stationary phase method, 36
Stationary signals, 331
Steepest descend method, 440
Taylor series, 36
Trace of matrix, 443

Unit step signal
continuous (Heaviside function), 15
discrete-time, 54
Unitary matrix, 676
Upsampling, 572

Variance, 325
Voting machines, 514

Walsh-Hadamard transform, 145
Wavelet transform, 563
Coifflet, 600
Daubechies D4, 582
Daubechies D6, 596
filter bank, 568
Haar, 579
orthogonality, 576
reconstruction condition, 573
scale function, 594
wavelet function, 594
Welch bound, 678
Welch periodogram, 362
Wide sense stationary signals, 331
Wiener filter, 366
Wigner distribution, 617, 650
auto-terms, 621
cross-terms, 621
discrete form, 633
properties, 625
pseudo, 630
S-method, 641
smoothed, 630
Window, 523
Bartlett (triangular), 283, 524
Blackman, 528
Hamming, 285, 527
Hann(ing), 285
Hann(ing) (Hann), 525
Kaiser, 529
rectangular, 523
Windows, 281
Yule-Walk equation, 361
z-transform, 163, 185
inverse, 168
Zero-padding, 114
Zhao-Atlas-Marks distribution
discrete form, 651
Bibliography


Ljubiša Stanković was born in Montenegro on June 1, 1960. He received a BSc degree in electrical engineering from the University of Montenegro in 1982 with the award as the best student at the University. As a student he won several competitions in mathematics in Montenegro and former Yugoslavia. He received an MSc degree in communications from the University of Belgrade, and a PhD in theory of electromagnetic waves from the University of Montenegro in 1988. As a Fulbright grantee, he spent 1984-1985 academic year at the Worcester Polytechnic Institute, Worcester, MA. Since 1982, he has been on the faculty at the University of Montenegro, where he has been a full professor since 1995. In 1997-1999, he was on leave at the Ruhr University Bochum, Germany, supported by the Alexander von Humboldt Foundation. At the beginning of 2001, he was at the Technische Universität Eindhoven, The Netherlands, as a visiting professor. During the period of 2003-2008, he was the rector of the University of Montenegro. He was ambassador of Montenegro to the United Kingdom, Iceland, and Ireland 2011-2015. During his stay in United Kingdom he was a visiting academic at the Imperial College London, 2013-2014. His current interests are in signal processing. He published about 400 technical papers, more than 140 of them in the leading journals. Stanković received the highest state award of Montenegro in 1997 for scientific achievements.

Stanković was an associate editor of the IEEE Transactions on Image Processing, an associate editor of the IEEE Signal Processing Letters, and an associate editor of the IEEE Transactions on Signal Processing. Stanković is a member of the Editorial Board of Signal Processing. He is a member of the National Academy of Sciences and Arts of Montenegro (CANU) since 1996 and a member of the European Academy of Sciences and Arts. Stanković is a Fellow of the IEEE for contributions to time-frequency signal analysis.