Ljubisa Stankovic is a professor at the University of Montenegro, IEEE Fellow for contributions to Time-Frequency Signal Analysis, a member of the Montenegrin Academy of Sciences and Arts (CANU), and a member of the Academia Europaea. He has been an Associate Editor, Senior Area Editor, and Deputy Editor of several world-leading journals in Signal Processing.

The book is a result of the 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 through 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.

Printed by CreateSpece, An Amazon.com Company

L. Stanković

Digital Signal Processing

Ljubiša Stanković

DIGITAL SIGNAL PROCESSING

with selected topics

ADAPTIVE SYSTEMS AND NEURAL NETWORKS
TIME-FREQUENCY ANALYSIS
SPARSE SIGNAL PROCESSING - COMPRESSIVE SENSING

2015

DIGITAL SIGNAL PROCESSING

Basic Theory and Applications

Adaptive Systems and Neural Networks
Time-Frequency Signal Analysis
Sparse Signal Processing – Compressive Sensing

Ljubiša Stanković

Revised edition 2020

Library of Congress Cataloging-in-Publication Data
Library of Congress Control Number: 2015912465

ISBN-13: 978-1514179987
ISBN-10: 1514179989

© 2015 Ljubiša Stanković, All Rights Reserved

Printed by CreateSpace Independent Publishing Platform,
An Amazon.com Company
North Charleston, South Carolina, USA.

Available from Amazon.com and other on-line and bookstores

All right reserved. Printed and bounded in the United States of America.

© 2020 Ljubiša Stanković, All Rights Reserved

No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the copyright holder.

To
my parents
Božo and *Cana* ,

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

Contents

I Review of Continuous-Time Signals and Systems **16**

Chapter 1 Continuous-Time Signals and Systems 17
1.1 Continuous-Time Signals 17
1.2 Linear Systems 20
1.3 Periodic Signals and Fourier Series 23
1.3.1 Fourier Series of Real-Valued Signals 25
1.4 Fourier Transform 30
1.4.1 Fourier Transform and Linear Time-Invariant Systems 33
1.4.2 Properties of the Fourier Transform 33
1.4.3 Relationship Between the Fourier Series and the Fourier Transform 36
1.5 Fourier Transform and the Stationary Phase Method 37
1.6 Laplace Transform 42
1.6.1 Properties of the Laplace Transform 44
1.6.2 Table of the Laplace Transform 46
1.6.3 Linear Systems Described by Differential Equations 47
1.7 Butterworth Filter 52

II Deterministic Discrete-Time Signals and Systems **57**

Chapter 2 Discrete-Time Signals and Transforms 58
2.1 Discrete-Time Signals 58
2.2 Discrete-Time Systems 62
2.3 Fourier Transform of Discrete-Time Signals 66
2.3.1 Properties 68
2.3.2 Spectral Energy and Power Density 72
2.4 Sampling Theorem in the Time Domain 73
2.5 Problems 78
2.6 Exercise 82
2.7 Solutions 84

Chapter 3 Discrete Fourier Transform 102
3.1 DFT Definition 102

3.2 DFT Properties 107
3.3 Zero-Padding and Interpolation 112
3.4 Relation among the Fourier Representations 117
3.5 Fast Fourier Transform 119
3.6 Sampling of Periodic Signals 123
3.7 Analysis of a Sinusoid Using the DFT 126
3.7.1 Leakage Effect 127
3.7.2 Displacement 129
3.8 Discrete Cosine and Sine Transforms 132
3.9 Discrete Walsh-Hadamard and Haar Transforms 137
3.9.1 Discrete Walsh-Hadamard Transform 140
3.9.2 Discrete Haar Wavelet Transform 143
3.10 Problems 146
3.11 Exercise 148
3.12 Solutions 150

Chapter 4 z-Transform 158
4.1 Definition of the z-transform 158
4.2 Properties of the z-transform 160
4.2.1 Linearity 160
4.2.2 Time-Shift 160
4.2.3 Multiplication by an exponential signal: Modulation 161
4.2.4 Differentiation 161
4.2.5 Convolution in time 162
4.2.6 Initial and Stationary State Signal Value 162
4.2.7 Table of the z-transform 163
4.3 Inverse z-transform 163
4.3.1 Direct Power Series Expansion 163
4.3.2 Theorem of Residues Based Inversion 167
4.4 Discrete systems and the z-transform 169
4.5 Difference equations 171
4.5.1 Solution Based on the z-transform 172
4.5.2 Solution to Difference Equations in the Time Domain 174
4.6 Relation of the z-transform to other Transforms 179
4.7 Problems 181
4.8 Exercise 184
4.9 Solutions 186

Chapter 5 From Continuous to Discrete Systems 200
5.1 Impulse Invariance Method 200
5.2 Matched z-transform method 205
5.3 Differentiation and Integration 207
5.4 Bilinear Transform 211
5.5 Discrete Filters Design 216
5.5.1 Lowpass filters 216
5.5.2 Highpass Filters 222
5.5.3 Bandpass Filters 223

5.5.4 Allpass Systems - System Stabilization 224
5.5.5 Inverse and Minimum Phase Systems 226
5.6 Problems 230
5.7 Exercise 232
5.8 Solutions 234

Chapter 6 Realization of Discrete Systems 242
6.1 Realization of IIR systems 242
6.1.1 Direct realization I 242
6.1.2 Direct realization II 243
6.1.3 Sensitivity of the System Poles/Zeros 246
6.1.4 Cascade Realization 249
6.1.5 Parallel realization 253
6.1.6 Inverse realization 256
6.2 FIR Systems and their Realizations 257
6.2.1 Linear Phase Systems and Group Delay 257
6.2.2 Windows 258
6.2.3 Design of a FIR System in the Frequency Domain 262
6.2.4 Realizations of the FIR systems 262
6.3 Problems 269
6.4 Exercise 271
6.5 Solutions 274

III Random Discrete-Time Signals and Systems 286

Chapter 7 Discrete-Time Random Signals 287
7.1 Basic Statistical Definitions 287
7.1.1 Mean Value – Sample Average 287
7.1.2 Median 293
7.1.3 Variance and Standard Deviation 297
7.1.4 Linear Regression Analysis 299
7.1.5 Random Sample Consensus (RANSAC) 301
7.1.6 Ridge Regression 304
7.2 Basic Probability Definitions 306
7.2.1 Probability 306
7.2.2 Expected Value and Variance 311
7.2.3 Probability Density Function 312
7.3 Second-Order Statistics 318
7.3.1 Correlation and Covariance 318
7.3.2 Stationarity and Ergodicity 320
7.3.3 Characteristic Function and Moments 321
7.3.4 Power Spectral Density 322
7.4 Noise and Random Signal Examples 327
7.4.1 Uniform Noise 327
7.4.2 Binary, Bernoulli, and Binomial Random Signal 328
7.4.3 Bayesian Inference for Binary and Binomial Random Signal 330

7.4.4 Gaussian Noise 334
7.4.5 Estimation of the Gaussian Distribution Parameters 337
7.4.6 Cramer-Rao Bound 341
7.4.7 Confidence Intervals 347
7.4.8 Bootstrap Method 354
7.4.9 Hypothesis Testing 356
7.4.10 Complex Gaussian Noise and Rayleigh Distribution 359
7.4.11 Impulsive Noises 361
7.4.12 Poisson Noise 363
7.4.13 Exponential Random Signal 364
7.4.14 Noisy Signals 366
7.5 Discrete Fourier Transform of Noisy Signals 367
7.5.1 Expected Value and Variance of the DFT 367
7.5.2 Spectral Estimation 369
7.5.3 Bias in the Fourier Transform of the Windowed Signals 370
7.5.4 Variance in the Fourier transform of Noisy Signals 371
7.5.5 Bias-to-Variance Trade-Off: Optimum Window Width 372
7.5.6 Periodogram 373
7.5.7 Detection of a Sinusoidal Signal Frequency 380
7.6 Linear Systems and Random Signals 383
7.6.1 Spectral Estimation of Narrowband Signals 388
7.6.2 Detection and Matched Filter 392
7.6.2.1 Matched Filter 392
7.6.2.2 Two Hypothesis Decision 395
7.6.3 Optimal Wiener Filter 397
7.7 Quantization effects 400
7.7.1 Input signal quantization 401
7.7.2 Quantization of the results 404
7.7.2.1 Fixed point arithmetic 404
7.7.2.2 Discrete rounding error 408
7.7.2.3 Floating point arithmetic 410
7.8 Problems 415
7.9 Exercise 420
7.10 Solutions 424

IV Adaptive Systems and Neural Networks 445

Chapter 8 Adaptive Systems 446
8.1 Introduction 446
8.2 Linear Adaptive Adder 449
8.2.1 Error Signal 451
8.2.2 Autocorrelation Matrix Eigenvalues and Eigenvectors 458
8.2.3 Error Signal Analysis 462
8.2.4 Orthogonality Principle 464
8.3 Steepest Descend Method 465
8.4 LMS Algorithm 474

8.4.1 Convergence of the LMS algorithm 475
8.5 LMS Application Examples 476
8.5.1 Identification of Unknown System 476
8.5.2 Noise Cancellation 479
8.5.3 Sinusoidal Disturbance Cancellation 482
8.5.4 Signal Prediction 483
8.5.5 Adaptive Antenna Arrays 487
8.5.6 Acoustic Echo Cancellation 492
8.6 Variations on the LMS Algorithm 493
8.6.1 Sign LMS 493
8.6.2 Block LMS 495
8.6.3 Normalized LMS Algorithm 496
8.6.4 LMS with Variable Step Size 497
8.6.5 Complex LMS 498
8.7 RLS Algorithm 499
8.8 Adaptive Recursive Systems 503
8.9 From the LMS algorithm to the Kalman filters 505

Chapter 9 Neural Networks 510
9.1 Neural Network Structure 510
9.1.1 Neuron 511
9.1.2 Network Function 511
9.1.3 Activation Function 512
9.1.4 Neural Network Topology 513
9.1.5 Network with Supervised Learning 515
9.1.6 One-Layer Network with Binary Output - Perceptron 515
9.1.7 One-Layer Neural Network with Continuous Output 519
9.1.8 Multilayer Neural Networks 521
9.1.9 Neural Networks with Unsupervised Learning 524
9.1.10 Voting Machines 524
9.2 Convolutional Neural Network – CNN 525
9.2.1 Forward Propagation 526
9.2.2 Updating Convolution Weights: Back-propagation 536
9.2.2.1 Initial Weights 536
9.2.2.2 Back-propagation in a two-layer CNN 538
9.2.2.3 Softmax Output Layer 540
9.2.2.4 Back-Propagation in a Multi-Layer CNN 541
9.2.2.5 Delta error back-propagation 542
9.2.2.6 Additional operations 546

V Time-Frequency Analysis 549

Chapter 10 Linear Time-Frequency Representations 550
10.1 Short-Time Fourier Transform 550
10.2 STFT Inversion 555
10.3 Windows 558

10.3.1 Rectangular Window 558
10.3.2 Triangular (Bartlett) Window 558
10.3.3 Hann(ing) Window 559
10.3.4 Hamming Window 561
10.3.5 Blackman and Kaiser Windows 562
10.4 Realizations of the STFT 564
10.4.1 Discrete Form and Realizations of the STFT 564
10.4.2 Recursive STFT Realization 568
10.4.3 Filter Bank STFT Implementation 569
10.4.3.1 Overlapping windows 569
10.5 Signal Reconstruction from the Discrete STFT 572
10.6 Varying Windows in the STFT 581
10.6.1 Time-Varying Windows 581
10.6.1.1 Time-Varying Hann(ing) Windows 590
10.6.1.2 Time-Varying Sine Windows 591
10.6.2 Frequency-Varying Window 592
10.6.3 Hybrid Time-Frequency-Varying Windows 594
10.7 Local Polynomial Fourier Transform 594
10.7.1 Fractional Fourier Transform with Relation to the LPFT 598
10.8 High-Resolution STFT 599
10.8.1 Capon's STFT 599
10.8.2 MUSIC STFT 602
10.8.3 Capon's LPFT 604

Chapter 11 Quadratic Time-Frequency Representations 606
11.1 Wigner Distribution 606
11.1.1 Auto-Terms and Cross-Terms in the Wigner Distribution 609
11.1.2 Wigner Distribution Properties 613
11.1.3 Pseudo and Smoothed Wigner Distribution 616
11.1.4 Discrete Pseudo Wigner Distribution 618
11.2 From the STFT to the Wigner Distribution via S-Method 624
11.3 General Quadratic Time-Frequency Distributions 628
11.3.1 Reduced Interference Distributions 632
11.3.2 Kernel Decomposition Method 636

Chapter 12 Wavelet Transform 639
12.1 Continuous Morlet Wavelet Transform 639
12.1.1 S-Transform 642
12.1.2 Spectral Meyer Wavelet Transform 643
12.2 Filter Bank and Discrete Wavelet 645
12.2.1 Lowpass and Highpass Filtering and Downsampling 646
12.2.2 Upsampling 648
12.2.3 Reconstruction Condition 649
12.2.4 Orthogonality Conditions 651
12.2.5 FIR Filter and Orthogonality Condition 652
12.2.6 Haar Wavelet Implementation 653
12.2.7 Daubechies D4 Wavelet Transform 655

12.2.8 Daubechies D4 Wavelet Functions in Different Scales 664
12.2.9 Daubechies D6 Wavelet Transform 668
12.2.10 Coifflet Transform 669
12.2.11 Discrete Wavelet Transform - STFT 670

VI Sparse Signal Processing and Compressive Sensing 673

Chapter 13 Sensing of Sparse Signals 674
13.1 Illustrative Examples 674
13.2 Basic Definitions 680
13.2.1 Sparsity 680
13.2.2 Measurements 681
13.2.2.1 Sparsity Aware Form of the Measurements 682
13.2.2.2 Signal Samples as Measurements 682
13.2.2.3 Indirect Measurements 684
13.3 Measurement Matrix Parameters 684
13.3.1 Measurement Matrix Examples 684
13.3.2 Isometry and Restricted Isometry Property 686
13.3.3 Coherence Index 688
13.3.4 Restricted Isometry Property and Coherence 691
13.3.5 Restricted Isometry Property and Eigenvalues 697
13.3.6 Spark and Rank of a Matrix 701
13.4 Reconstruction Uniqueness 702
13.4.1 Unique Reconstruction Condition and the RIP 702
13.4.2 Spark and the Solution Uniqueness 705
13.4.3 Coherence Index Uniqueness Relations 706
13.4.4 Spark and Coherence Index Relation 707

Chapter 14 Sparse Signals Reconstruction 709
14.1 Norm-Zero Based Reconstruction 709
14.1.1 Problem Formulation 709
14.1.2 Known Coefficient Positions 710
14.1.3 Direct Combinatorial Search 712
14.1.4 Pseudoinverse matrix 712
14.1.5 Estimation of Unknown Positions 713
14.1.5.1 Initial Estimate 714
14.1.5.2 Sparsity Aware Initial Estimate 715
14.1.5.3 Reconstruction Algorithm 715
14.1.6 Iterative Reconstruction 718
14.1.7 CoSaMP Reconstruction Algorithm 722
14.1.8 Initial Estimate and the Coherence Based Uniqueness Relation 722
14.1.9 Solution Exactness 724
14.1.10 The Initial Estimate Noise 726
14.1.10.1 Expected Value of the Initial Estimate 726
14.1.10.2 Variance of the Initial Estimate 726

14.1.10.3 Distribution of the Initial Estimate in the Partial DFT Matrix 730
14.1.11 Influence of Additive Input Noise 732
14.1.11.1 Signal-To-Noise Ratio Bounds 733
14.1.11.2 Exact Signal-To-Noise Ratio 733
14.1.11.3 Exact Signal-To-Noise Ratio in the Partial DFT 734
14.1.12 Nonsparse Signal Reconstruction 736
14.1.13 Iterative Hard Thresholding Reconstruction Algorithm 739
14.2 Norm-One Based Reconstruction 739
14.2.1 Uniqueness Illustration in the Sparsity Domain 741
14.2.2 Illustration of the ℓ_0-norm and ℓ_1-norm Solution Equivalence 743
14.2.3 Conditions for the ℓ_0-norm and ℓ_1-norm Solution Equivalence 752
14.3 Norm-one Based Reconstruction Algorithms 754
14.3.1 LASSO- Minimization 754
14.3.1.1 Iterative Calculation 755
14.3.1.2 Norm-Two Based Minimization Solution 758
14.3.2 Reconstruction in the Signal Domain 758
14.3.2.1 Direct Search in the Signal Domain 759
14.3.2.2 Reconstruction with a Gradient Algorithm 760
14.3.2.3 Algorithm 762
14.3.2.4 Finite Difference Step 765
14.3.2.5 Comments on the Algorithm 767
14.4 On the Uniqueness of the DFT of Sparse Signals 770
14.5 Indirect Measurements 778
14.6 Processing of Sparse Signals with Impulsive Noise 785
14.6.1 Direct Search Procedure 786
14.6.2 Criteria for Selecting Samples 787
14.6.2.1 Iterative procedure 788
14.6.3 Uniqueness of the Obtained Solution 789
14.7 RANSAC-Based Signal Denoising Using Compressive Sensing 791
14.7.1 RANSAC-based CS Signal Denoising 791
14.7.2 Expected Signal-to-Noise Ratio 794
14.7.3 Numerical and Statistical Examples 794
14.8 Image Reconstruction 799
14.9 Median Based Formulation and Reconstruction 800
14.10 Bayesian-Based Reconstruction 806

About the Author 819

Preface

THIS book is a result of the 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 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 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 an 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 topics 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. The theory is explained in a simple way with a necessary mathematical rigor. The book provides simple examples and explanations for every presented transform, method, algorithm or approach. Sophisticated results in signal processing theory are illustrated by simple numerical examples.

Part three of the book contains a 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 would like to thank colleagues: prof. Zdravko Uskoković, prof. Srdjan Stanković, prof. Igor Djurović, prof. Veselin Ivanović, prof. Miloš Daković, prof. Božo

Krstajić, prof. Vesna Popović-Bugarin, prof. Slobodan Djukanović, prof. Irena Orović, dr. Nikola Žarić, dr Marko Simeunović, M.Sc. Predrag Raković, M.Sc. Andjela Draganić and M.Sc. Isidora Stanković for careful reading of the initial version of this book and for the comments that helped to improve the presentation.

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,
July 2013 - July 2015.

Author

Preface to the Revised Edition

The book has been slightly edited, keeping the main structure unchanged. *The chapter dealing with random signals is updated to provide the basis for machine learning, compressive sensing, graph signal processing*, and other modern signal processing areas.

Podgorica, Montenegro
March - June 2020.

Author

Last revision on March 9, 2021.

Introduction

A signal is a physical process, mathematical function, or any other physical or symbolic representation of information. Signal theory and processing are the areas dealing with the efficient generation, description, transformation, transmission, reception, and interpretation of signals. 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 tools are strongly related to 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 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 the 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 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 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 the sampling frequency of up to megasample (some even up to a 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,

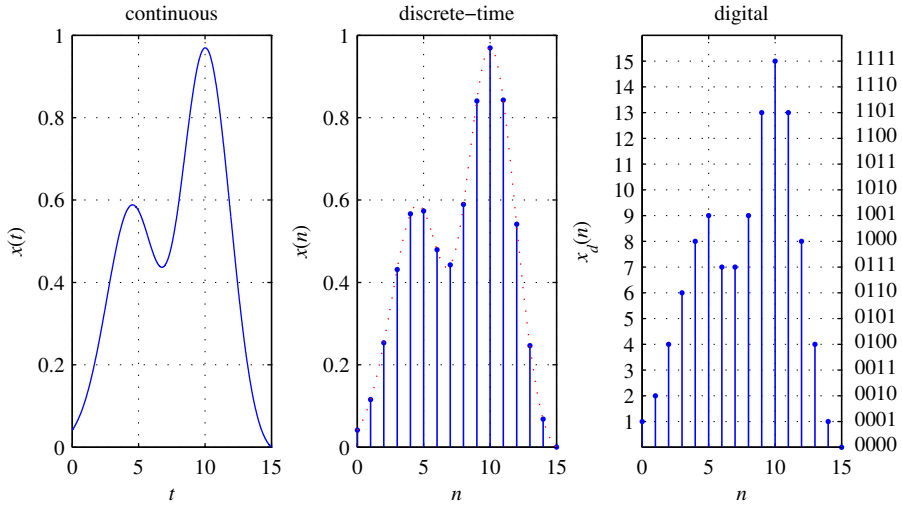


Figure 1 A continuous-time analog signal (left) and its discrete-time (middle) and digital version (right).

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.

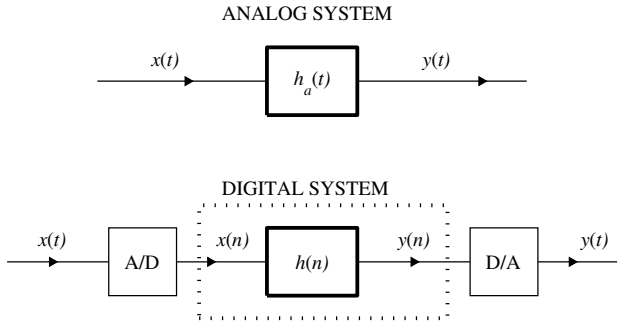


Figure 2 Illustration of an analog and a digital system used to process an analog signal.

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.

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 of Continuous-Time Signals and Systems

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

$$u(t) = \begin{cases} 1, \text{ for } t \geq 0 \\ 0, \text{ for } t < 0. \end{cases} \tag{1.1}$$

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. The 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

$$\delta(t) = 0, \text{ for } t \neq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} \delta(t)dt = 1. \tag{1.2}$$

The impulse signal is equal to 0 everywhere, except at $t = 0$, where it takes 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 using the impulse signal, as

$$x(t) = \int_{-\infty}^{\infty} x(t-\tau)\delta(\tau)d\tau = \int_{-\infty}^{\infty} x(\tau)\delta(t-\tau)d\tau. \tag{1.3}$$

The unit-step signal can be related to the impulse signal using the previous relation as

$$u(t) = \int_{-\infty}^{\infty} \delta(\tau)u(t-\tau)d\tau = \int_{-\infty}^{t} \delta(\tau)d\tau$$

or

$$\frac{du(t)}{dt} = \delta(t). \tag{1.4}$$

The sinusoidal signal, with amplitude A, frequency Ω_0, and initial phase φ, is a signal of the form

$$x(t) = A\sin(\Omega_0 t + \varphi). \tag{1.5}$$

This signal is periodic in time, since it satisfies the periodicity condition

$$x(t+T) = x(t). \tag{1.6}$$

In this case, the period is $T = 2\pi/\Omega_0$.

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

The complex-valued sinusoidal signal whose definition is given by

$$x(t) = Ae^{j(\Omega_0 t+\varphi)} = A\cos(\Omega_0 t + \varphi) + jA\sin(\Omega_0 t + \varphi) \tag{1.7}$$

is also periodic with period $T = 2\pi/\Omega_0$. Fig. 1.1 depicts basic continuous-time signals.

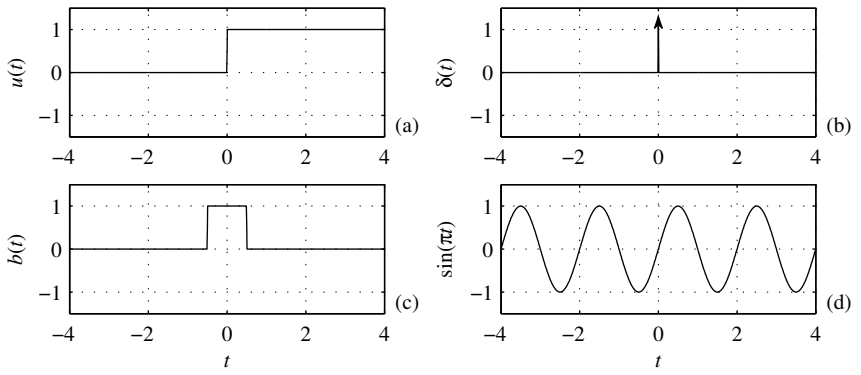


Figure 1.1 Continuous-time signals: (a) unit-step signal, (b) impulse signal, (c) boxcar signal, and (d) sinusoidal signal.

Example 1.1. Find the periods of the 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 the 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 in it. Thus, the signal $x_5(t)$ is not periodic.

Example 1.2. Find the period of the signal defined by

$$x(t) = \sum_{n=0}^{N} A_n e^{jn\Omega_0 t}.$$

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

Some parameters that can be used to describe a signal are:

- Maximum absolute value (magnitude) of the signal $x(t)$ is defined by

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

- Signal energy of the same signal is

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

- Signal $x(t)$ instantaneous power

$$P_x(t) = |x(t)|^2. \tag{1.10}$$

The average signal $x(t)$ power is defined by

$$P_{AV} = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} |x(t)|^2 dt. \tag{1.11}$$

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.

Example 1.3. Find the magnitude, energy, instantaneous power, and average power of the signal $x(t)$ given by

$$x(t) = te^{-t}u(t). \tag{1.12}$$

★ The signal $x(t)$ is a nonnegative continuous function with the initial and the final value equal to $x(0) = 0$ and $\lim_{t\to\infty} x(t) = 0$, respectively. The magnitude of this signal is obtained as its maximum, from

$$\frac{dx(t)}{dt} = e^{-t} - te^{-t} = (1-t)e^{-t} = 0, \quad \text{for} \quad t > 0. \tag{1.13}$$

The maximum $M_x = 1/e$ is achieved at $t = 1$. The energy of this signal is equal to

$$E_x = \int_0^\infty t^2 e^{-2t} dt = -\frac{e^{-2t}}{2}t^2\Big|_0^\infty + \int_0^\infty te^{-2t}dt = -\frac{e^{-2t}}{2}t\Big|_0^\infty + \frac{1}{2}\int_0^\infty e^{-2t}dt = \frac{1}{4}, \tag{1.14}$$

where the integration in parts is used twice, with $\lim_{t\to\infty} e^{-2t}t^2 = 0$. The instantaneous power of the signal $x(t)$ is $P_x(t) = t^2e^{-2t}u(t)$. The average power of this signal is $P_{AV} = 0$.

1.2 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, $\mathbb{T}\{\circ\}$. The output signal can be written as

$$y(t) = \mathbb{T}\{x(t)\}. \tag{1.15}$$

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

$$y(t) = \mathbb{T}\{a_1x_1(t) + a_2x_2(t)\} = a_1\mathbb{T}\{x_1(t)\} + a_2\mathbb{T}\{x_2(t)\}. \tag{1.16}$$

A system is *time-invariant* if its properties and parameters do not change over time. For a time-invariant system, the following relation:

$$\text{if } y(t) = \mathbb{T}\{x(t)\} \quad \text{then} \quad \mathbb{T}\{x(t-t_0)\} = y(t-t_0), \tag{1.17}$$

holds for any t_0.

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

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

then for any signal $x(t)$ at the system input, the output can be obtained using (1.3), as

$$y(t) = \mathbb{T}\{x(t)\} = \mathbb{T}\left\{\int_{-\infty}^{\infty} x(\tau)\delta(t-\tau)d\tau\right\}$$
$$\stackrel{Linearity}{=} \int_{-\infty}^{\infty} x(\tau)\mathbb{T}\{\delta(t-\tau)\}d\tau \stackrel{Time-invariance}{=} \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau.$$

The last integral is of particular importance in signals and systems. It is called the *convolution* in time of $x(t)$ and $h(t)$, and has a specific notation

$$y(t) = x(t) *_t h(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau. \tag{1.18}$$

The convolution is a commutative operation, since

$$x(t) *_t h(t) = h(t) *_t x(t) \tag{1.19}$$

holds.

Example 1.4. Find the convolution of the two boxcar signals $x(t) = u(t) - u(t-5)$ and $h(t) = u(t) - u(t-2)$.

★The signals $x(\tau)$ and $h(t-\tau)$ are shown in Fig. 1.2 for $t = 0$ and $t = 1.25$. For example, the convolution value at $t = 0$ is obtained using the integral of the product of $x(\tau)$ and $h(-\tau)$, that is

$$y(0) = \int_{-\infty}^{\infty} x(\tau)h(-\tau)d\tau = 0. \tag{1.20}$$

For $t < 0$, the nonzero values of $x(\tau)$ and $h(t-\tau)$ do not overlap, resulting in $y(t) = 0$. For $0 \leq t < 2$, the output signal is $y(t) = \int_0^t d\tau = t$, while for $2 \leq t < 5$, $y(t) = 2$. For $5 \leq t < 7$, the value of $y(t)$ is $y(t) = 7 - t$. Finally, for $t \geq 7$ the convolution value is equal to zero, $y(t) = 0$, as shown in Fig. 1.2.

Duration of the convolution, $y(t) = x(t) *_t h(t)$, is equal to the sum of durations of $x(t)$ and $h(t)$, that is $T_y = T_x + T_h$, where T_x, T_h, and T_y, are the respective durations of $x(t)$, $h(t)$, and $y(t)$.

Example 1.5. Find the convolution of the two signals $x(t) = u(t+1) - u(t-1)$ and $h(t) = e^{-t}u(t)$.

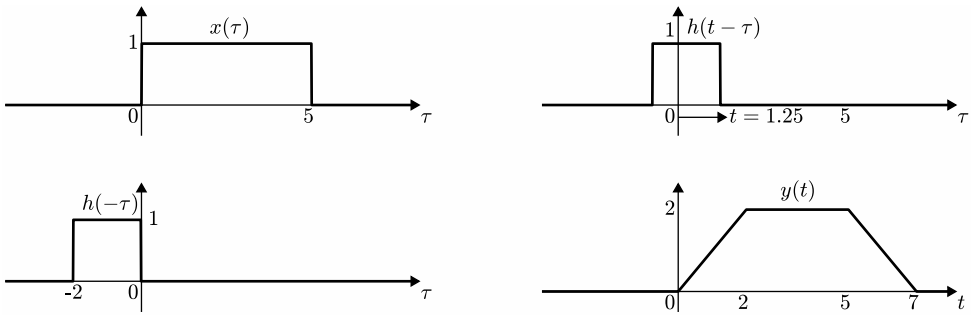


Figure 1.2 Calculation of the convolution, $y(t) = x(t) *_t h(t)$, of signals $x(t) = u(t) - u(t-5)$ and $h(t) = u(t) - u(t-2)$.

★By using the convolution definition, we get

$$y(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau = \int_{-1}^{1} 1 \cdot e^{-(t-\tau)}u(t-\tau)d\tau$$

$$= -\int_{t+1}^{t-1} e^{-\lambda}u(\lambda)d\lambda = \begin{cases} \int_{t-1}^{t+1} e^{-\lambda}d\lambda = e^{-t}(e-1/e), & \text{for } t \geq 1 \\ \int_{0}^{t+1} e^{-\lambda}d\lambda = 1 - e^{-(t+1)}, & \text{for } -1 \leq t < 1 \\ 0 & \text{for } t < -1. \end{cases}$$

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

$$\int_{-\infty}^{\infty} |h(\tau)|d\tau < \infty \tag{1.21}$$

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 (1.21) 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 PERIODIC SIGNALS AND FOURIER SERIES

Consider a periodic signal $x(t)$ with a period T. It can be expanded onto periodic complex sinusoidal functions $\phi_n(t) = e^{j2\pi nt/T}$, for $-\infty < n < \infty$,

$$\begin{aligned} x(t) &= \cdots + X_{-1}e^{-j2\pi t/T} + X_0 e^{-j0} + X_1 e^{j2\pi t/T} + \cdots \\ &= \sum_{n=-\infty}^{\infty} X_n e^{j2\pi nt/T} \end{aligned} \tag{1.22}$$

if the following (Dirichlet) conditions are met:

(1) The signal $x(t)$ has a finite number of finite discontinuities within the period T;

(2) It is absolutely integrable over the period T, that is $\int_{-T/2}^{T/2} |x(t)|dt \leq c < \infty$; and

(3) The signal $x(t)$ has a finite number of maxima and minima.

Since signal analysis deals with real-world physical signals, rather than mathematical generalizations, these conditions are almost always met.

The set of basis functions $\phi_n(t) = e^{j2\pi nt/T}$, $-\infty < n < \infty$, is an orthogonal set of functions since their inner product is

$$\left\langle e^{j2\pi mt/T}, e^{j2\pi nt/T} \right\rangle = \frac{1}{T}\int_{-T/2}^{T/2} e^{j2\pi mt/T} e^{-j2\pi nt/T} dt = \begin{cases} 1 & \text{for } m = n \\ \frac{\sin(\pi(m-n))}{\pi(m-n)} = 0 & \text{for } m \neq n. \end{cases}$$

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

$$X_n = \left\langle x(t), e^{j2\pi nt/T} \right\rangle = \frac{1}{T}\int_{-T/2}^{T/2} x(t) e^{-j2\pi nt/T} dt. \tag{1.23}$$

This relation follows after a simple multiplication of the right and left side of (1.22) by $e^{-j2\pi mt/T}$ and a normalized integration within the period, that is $\frac{1}{T}\int_{-T/2}^{T/2}(\cdot)\,dt$.

Normalization is achieved using the factor of $1/T$ in the scalar product definition. If this factor were not used, then the orthonormal set of basis functions would be defined by $\phi_n(t) = e^{j2\pi t/T}/\sqrt{T}$, $-\infty < n < \infty$, with the same conclusions and relations.

Since the signal and the basis functions are periodic with period T, we can use

$$\frac{1}{T}\int_{-T/2}^{T/2} x(t) e^{-j2\pi nt/T} dt = \frac{1}{T}\int_{-T/2+\Lambda}^{T/2+\Lambda} x(t) e^{-j2\pi nt/T} dt \tag{1.24}$$

in all previous integrals, where Λ is an arbitrary constant.

The signal expansion in (1.22) is known as the *Fourier series*, and the coefficients X_n are the Fourier series coefficients.

Example 1.6. What are the Fourier series coefficients of the periodic signal $x(t) = \cos^2(\pi t/4)$. What will be the coefficient values if the 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 (1.22) form, we get

$$x(t) = \frac{1}{4}e^{-j2\pi t/4} + \frac{1}{2} + \frac{1}{4}e^{j2\pi t/4}.$$

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 (1.23)). 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 of the form $e^{j2\pi nt/8}$ (see relation (1.22)). The signal can be written as

$$x(t) = \frac{1}{4}e^{-j2\pi 2t/8} + \frac{1}{2} + \frac{1}{4}e^{j2\pi 2t/8}. \tag{1.25}$$

Thus, by 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 $X_2 = 1/4$. The remaining coefficients X_n are equal to zero.

Example 1.7. Calculate the Fourier series coefficients of a periodic signal $x(t)$ defined as

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

with

$$x_0(t) = u(t + 1/4) - u(t - 1/4). \tag{1.26}$$

★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. The Fourier series coefficients are obtained from

$$X_n = \frac{1}{2}\int_{-1/4}^{1/4} 1e^{-j2\pi nt/2}dt = \frac{\sin(\pi n/4)}{\pi n}, \tag{1.27}$$

with $X_0 = 1/4$. Values of X_n are shown in Fig. 1.3 (right).

The signal $x(t)$ can be reconstructed using the Fourier series (1.22). In numeric calculations, a finite number of M terms is used,

$$x_M(t) = \sum_{n=-M}^{M} X_n e^{j\pi nt} = \frac{1}{4} + \sum_{n=1}^{M} \frac{\sin(\pi\frac{n}{4})}{\pi n}\left(e^{-j\pi nt} + e^{j\pi nt}\right) = \frac{1}{4} + \sum_{n=1}^{M} \frac{\sin(\pi\frac{n}{4})}{\pi\frac{n}{2}}\cos(\pi nt).$$

The reconstructed signal, with $M = 1,2,6$, and 30, is shown in Fig. 1.4.

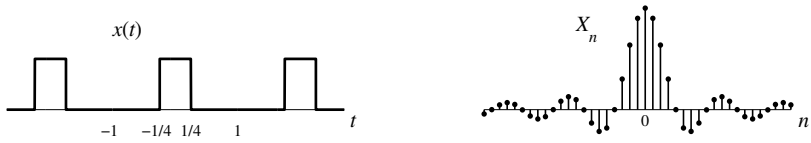


Figure 1.3 Periodic signal, $x(t)$, (left) and its Fourier series coefficients, X_n, (right).

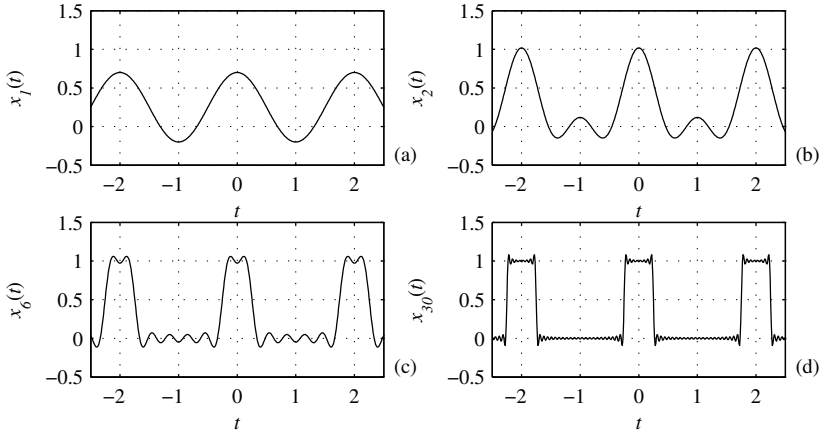


Figure 1.4 Reconstruction of the signal $x(t)$ using a finite Fourier series with: (a) the coefficients X_n within $-1 \leq n \leq 1$, (b) the coefficients X_n within $-2 \leq n \leq 2$, (c) the coefficients X_n within $-6 \leq n \leq 6$, and (d) the coefficients X_n within $-30 \leq n \leq 30$.

1.3.1 Fourier Series of Real-Valued Signals

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

$$X_n = \frac{1}{T}\int_{-T/2}^{T/2} x(t)\cos(\frac{2\pi nt}{T})dt - j\frac{1}{T}\int_{-T/2}^{T/2} x(t)\sin(\frac{2\pi nt}{T})dt = \frac{A_n - jB_n}{2}, \qquad (1.28)$$

where $A_n/2$ and $-B_n/2$ are the 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 equal to

$$A_n = X_n + X_{-n} = \frac{2}{T}\int_{-T/2}^{T/2} x(t)\cos(\frac{2\pi nt}{T})dt,$$
$$B_n = \frac{X_n - X_{-n}}{-j} = \frac{2}{T}\int_{-T/2}^{T/2} x(t)\sin(\frac{2\pi nt}{T})dt. \tag{1.29}$$

The Fourier series form of real-valued signals is

$$\begin{aligned} x(t) &= \sum_{n=-\infty}^{-1} X_n e^{j2\pi nt/T} + X_0 + \sum_{n=1}^{\infty} X_n e^{j2\pi nt/T} \\ &= X_0 + \sum_{n=1}^{\infty}(X_n e^{j2\pi nt/T} + X_{-n}e^{-j2\pi nt/T}) \\ &= X_0 + \sum_{n=1}^{\infty}(X_n + X_{-n})\cos(\frac{2\pi nt}{T}) + j(X_n - X_{-n})\sin(\frac{2\pi nt}{T}) \\ &= \frac{A_0}{2} + \sum_{n=1}^{\infty} A_n\cos(\frac{2\pi nt}{T}) + \sum_{n=1}^{\infty} B_n\sin(\frac{2\pi nt}{T}), \end{aligned} \tag{1.30}$$

with $|X_n| = \sqrt{A_n^2 + B_n^2}/2$. For real-valued signals the integrals in (1.29), corresponding to A_n and B_n, are respectively even and odd function of n. Therefore, it is possible to calculate

$$H_n = \frac{1}{T}\int_{-T/2}^{T/2} x(t)\left[\cos(\frac{2\pi nt}{T}) + \sin(\frac{2\pi nt}{T})\right]dt = \frac{1}{T}\int_{-T/2}^{T/2} x(t)\mathrm{cas}(\frac{2\pi nt}{T})dt \tag{1.31}$$

and to get

$$A_n = H_n + H_{-n}$$
$$B_n = H_n - H_{-n}.$$

The coefficients calculated by (1.31) are the Hartley series coefficients. For a real-valued and even signal, $x(t) = x(-t)$, the Hartley series reduces to

$$C_n = X_n = \frac{A_n}{2} = \frac{1}{T}\int_{-T/2}^{T/2} x(t)\cos(\frac{2\pi nt}{T})dt = \frac{2}{T}\int_{0}^{T/2} x(t)\cos(\frac{2\pi nt}{T})dt,$$

corresponding to the Fourier *cosine series* coefficients,

$$x(t) = C_0 + \sum_{n=1}^{\infty} 2C_n\cos(\frac{2\pi nt}{T}). \tag{1.32}$$

A similar expression is obtained for an odd and real-valued signal $x(t)$, when the Fourier series reduces to the Fourier *sine series* coefficients,

Example 1.8. Consider the Fourier series based reconstruction of the the signal

$$x(t) = t[u(t) - u(t-1/2)],$$

whose duration (nonzero values) is limited to $0 \leq t < 1/2$. For the Fourier series expansion, a periodic extension of the signal must be formed. The rate of the Fourier series coefficients convergence depends on the way how the periodic extension of this signal is formed.
(a) Calculate the Fourier series of the original signal extended periodically 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 extended first with its reversed version,

$$x_c(t) = x(t) + x(1-t),$$

and then extended periodically with the 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

$$X_n = \frac{1}{1/2} \int_0^{1/2} te^{-j2\pi n/(1/2)t} dt = \frac{1}{-j4\pi n}$$

with $X_0 = 1/4$. The signal reconstructed 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 several values of M is shown in Fig. 1.5.
(b) The Fourier series coefficients of the signal $x(t)$ extended with period 1 are

$$X_n = \frac{1}{1} \int_0^{1/2} te^{-j2\pi nt} dt = \frac{1}{-j2\pi n} te^{-j2\pi nt} \Big|_0^{1/2} + \frac{1}{(2\pi n)^2} e^{-j2\pi nt} \Big|_0^{1/2} = \frac{(-1)^n}{-j4\pi n} + \frac{(-1)^n - 1}{(2\pi n)^2},$$

with $X_0 = 1/8$. Note that the relation between the Fourier coefficients in (a) and (b) is $2X_{2n}^{(b)} = X_n^{(a)}$. The reconstruction is given in Fig. 1.6.

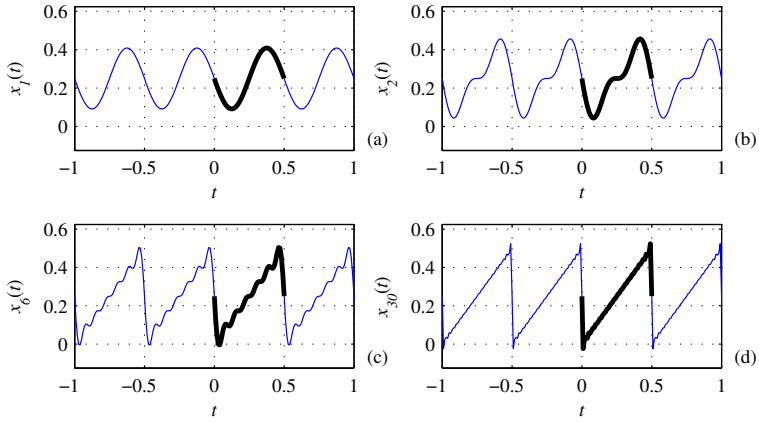


Figure 1.5 Reconstruction of the signal $x(t)$ using the Fourier series. Reconstructed signal is denoted by $x_M(t)$, where M indicates the number of coefficients used in reconstruction.

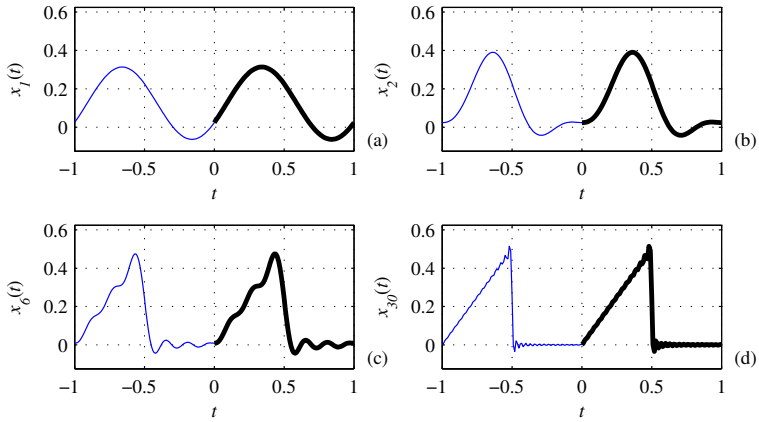


Figure 1.6 Reconstruction of the periodic signal $x(t)$, with a zero interval extension before the Fourier series is used.

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

$$X_n = C_n = \frac{1}{1}\Big(\int_0^{1/2} te^{-j2\pi nt}dt + \int_{1/2}^{1}(1-t)e^{-j2\pi nt}dt\Big) = 2\int_0^{1/2} t\cos(2\pi nt)dt = \frac{(-1)^n-1}{2\pi^2n^2}$$

with $C_0 = 1/4$. The reconstruction formula is

$$x_M(t) = \frac{1}{4} + 2\sum_{n=1}^{M}\frac{(-1)^n-1}{2\pi^2n^2}\cos(2\pi nt) = \frac{1}{4} - 2\sum_{n=1}^{M}\frac{1}{\pi^2(2n-1)^2}\cos\Big(2\pi(2n-1)t\Big).$$

The reconstructed signal is shown in Fig. 1.7.

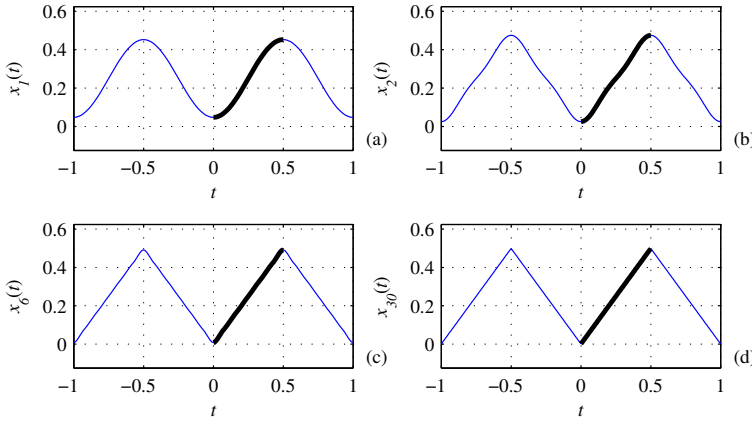


Figure 1.7 Reconstruction of a periodic signal after an even extension before using the Fourier series (cosine Fourier series).

(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 a given number of coefficients, will be achieved in case (c). Also, for a given reconstruction error the smallest number of reconstruction terms M would be required in case (c). This kind of signal extension (even signal extension) will be later used as a basis for the definition of the so called cosine signal transforms. From these periodic extensions, we can also conclude that an extension that avoids signal discontinuities at the interval ending instants improves the series convergence.

Example 1.9. Show that the Fourier series coefficients X_n of a periodic signal $x(t)$ can be obtained by minimizing the mean squared error between the signal and $\sum_{n=-N}^{N} X_n e^{j2\pi nt/T}$ within the period T.

★The mean squared value of the error, defined by

$$e(t) = x(t) - \sum_{n=-N}^{N} X_n e^{j2\pi nt/T},$$

within the period, T, is given by

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

From $\partial I/\partial X_m^* = 0$ follows

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

$$X_m = \frac{1}{T}\int_{-T/2}^{T/2} x(t)e^{-j2\pi mt/T}dt. \tag{1.33}$$

Note: The derivative of a complex function $F(z) = u(x,y) + jv(x,y)$, with $z = x + jy$, are $u(x,y)$, $v(x,y)$ are real-valued functions, is defined by

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

Commonly, a half of these values is used in the definition.

In order to justify the complex derivation $\partial I/\partial X_m^*$ in (1.33) let us denote: (i) the complex-valued variable X_m by $z = x + jy$, (ii) all terms in $x(t) - \sum_{n=-N}^{N} X_n e^{j2\pi nt/T} = f(z)$ which do not depend on $z = X_m = x + jy$ by $a + jb$, and (iii) the value of $-e^{j2\pi mt/T}$ by $e^{j\alpha}$. Now 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

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

For the minimization of the real-valued function $|f(z)|^2$ of two variables x and y we need partial derivatives

$$\frac{\partial |f(z)|^2}{\partial x} = 2\cos\alpha(a + x\cos\alpha - y\sin\alpha) + 2\sin\alpha(b + x\sin\alpha + y\cos\alpha) \tag{1.34}$$
$$= 2\,\mathrm{Re}\{e^{-j\alpha}f(z)\}$$

and

$$\frac{\partial |f(z)|^2}{\partial y} = 2\,\mathrm{Im}\{e^{-j\alpha}f(z)\}. \tag{1.35}$$

Therefore, all calculations with two real-valued equations (1.34) and (1.35) are the same as one complex-valued relation

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

1.4 FOURIER TRANSFORM

The Fourier series has been introduced and presented for periodic signals, with a period T. Assume now that the signal is of limited duration and that the period for its expansion is extended toward 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

$$\lim_{T\to\infty} X_n T = \lim_{T\to\infty} \int_{-T/2}^{T/2} x(t)e^{-j2\pi nt/T}dt = \int_{-\infty}^{\infty} x(t)e^{-j\Omega t}dt \tag{1.36}$$

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

$$X(\Omega) = \int_{-\infty}^{\infty} x(t)e^{-j\Omega t}dt, \tag{1.37}$$

is called the *Fourier transform* (FT) of a signal $x(t)$. For the Fourier transform existence it is sufficient that the signal is absolutely integrable, that is

$$\int_{-\infty}^{\infty} |x(t)|dt < \infty. \tag{1.38}$$

There are some signals that do not satisfy this condition, such as the unit-step signal, whose Fourier transform exists in the form of generalized functions.

The inverse Fourier transform (IFT) can be obtained by multiplying both sides of (1.37) by $e^{j\Omega\tau}$ and integrating over Ω,

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

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

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\Omega)e^{j\Omega t}d\Omega. \tag{1.39}$$

Example 1.10. Calculate the Fourier transform of the signal

$$x(t) = Ae^{-at}u(t).$$

★According to the Fourier transform definition we can write

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

The Fourier transform of this signal exists for $a > 0$, when

$$\int_0^\infty |x(t)|dt = A\int_0^\infty e^{-at}dt = A\frac{e^{-at}}{-a}\Big|_0^\infty = \frac{A}{a} < \infty.$$

Example 1.11. Find the Fourier transform of the signal

$$x(t) = \text{sign}(t) = \begin{cases} 1 & \text{for } t > 0 \\ 0 & \text{for } t = 0 \\ -1 & \text{for } t < 0 \end{cases}. \tag{1.40}$$

★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 the signal $x(t)$ can be obtained as the following limit

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

The Fourier transform of $x(t)$ can be calculated from

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

where

$$X_a(\Omega) = \int_{-\infty}^0 -e^{at}e^{-j\Omega t}dt + \int_0^\infty e^{-at}e^{-j\Omega t}dt = \frac{2\Omega}{ja^2 + j\Omega^2}. \tag{1.41}$$

It results in

$$X(\Omega) = \frac{2}{j\Omega}. \tag{1.42}$$

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 $2\pi x(-\Omega)$

$$\begin{aligned} X(\Omega) &= \text{FT}\{x(t)\} \\ 2\pi x(-\Omega) &= \text{FT}\{X(t)\}, \end{aligned} \tag{1.43}$$

where $\text{FT}\{\circ\}$ denotes the Fourier transform operator.

Example 1.12. Find the Fourier transform of the signals $\delta(t)$, $x(t)=1$, and $u(t)$.

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

$$\mathrm{FT}\{\delta(t)\}=\int_{-\infty}^{\infty}\delta(t)e^{-j\Omega t}dt=1. \tag{1.44}$$

According to the duality property, the Fourier transform of $x(t)=1$ is

$$\mathrm{FT}\{1\}=2\pi\delta(\Omega). \tag{1.45}$$

Finally, for the unit-step signal we get

$$\mathrm{FT}\{u(t)\}=\mathrm{FT}\left\{\frac{\mathrm{sign}(t)+1}{2}\right\}=\frac{1}{j\Omega}+\pi\delta(\Omega). \tag{1.46}$$

1.4.1 Fourier Transform and Linear Time-Invariant Systems

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

$$\begin{aligned}y(t)=x(t)*_t h(t)&=\int_{-\infty}^{\infty}Ae^{j(\Omega_0(t-\tau)+\varphi)}h(\tau)d\tau\\&=Ae^{j(\Omega_0 t+\varphi)}\int_{-\infty}^{\infty}h(\tau)e^{-j\Omega_0\tau}d\tau=H(\Omega_0)x(t),\end{aligned} \tag{1.47}$$

where

$$H(\Omega)=\int_{-\infty}^{\infty}h(t)e^{-j\Omega t}dt \tag{1.48}$$

is the Fourier transform of $h(t)$. The linear time-invariant system does not change the form of an input complex harmonic signal $x(t)=Ae^{j(\Omega_0 t+\varphi)}$. It remains complex harmonic signal after passing through the linear time-invariant system, with the same frequency Ω_0. The amplitude of the input signal $x(t)$ is changed for $|H(\Omega_0)|$ and the phase is changed for $\arg\{H(\Omega_0)\}$.

1.4.2 Properties of the Fourier Transform

The Fourier transform satisfies the following properties:

1. *Linearity*: The Fourier transform of a linear combination of two signals, $x_1(t)$ and $x_2(t)$ is

$$\mathrm{FT}\{a_1x_1(t)+a_2x_2(t)\}=a_1X_1(\Omega)+a_2X_2(\Omega), \tag{1.49}$$

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-valued (that is, $X^*(\Omega) = X(\Omega)$), if

$$x^*(-t) = x(t),$$

since

$$X^*(\Omega) = \int_{-\infty}^{\infty} x^*(t)e^{j\Omega t}dt \overset{t\to -t}{=} \int_{-\infty}^{\infty} x^*(-t)e^{-j\Omega t}dt = X(\Omega), \tag{1.50}$$

if $x^*(-t) = x(t)$.
3. *Modulation*: If the signal $x(t)$ is modulated by $e^{j\Omega_0 t}$ the Fourier transform of the modulated signal is shifted in frequency, that is

$$\mathrm{FT}\{x(t)e^{j\Omega_0 t}\} = \int_{-\infty}^{\infty} x(t)e^{j\Omega_0 t}e^{-j\Omega t}dt = X(\Omega - \Omega_0) \tag{1.51}$$

$$\mathrm{FT}\{2x(t)\cos(\Omega_0 t)\} = X(\Omega - \Omega_0) + X(\Omega + \Omega_0).$$

4. *Shift in time*: The Fourier transform of the signal $x(t)$ shifted in time for t_0 is modulated in the frequency domain,

$$\mathrm{FT}\{x(t - t_0)\} = \int_{-\infty}^{\infty} x(t - t_0)e^{-j\Omega t}dt = X(\Omega)e^{-jt_0\Omega}. \tag{1.52}$$

5. *Time-scaling*: For a signal scaled in time by factor a the Fourier transform is given by

$$\mathrm{FT}\{x(at)\} = \int_{-\infty}^{\infty} x(at)e^{-j\Omega t}dt = \frac{1}{|a|}X(\frac{\Omega}{a}). \tag{1.53}$$

6. *Convolution*: The Fourier transform of the convolution of signals $x(t)$ and $h(t)$ is equal to the product of their corresponding Fourier transforms, that is

$$\mathrm{FT}\{x(t) *_t h(t)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x(\tau)h(t - \tau)e^{-j\Omega t}d\tau dt \tag{1.54}$$

$$\overset{t-\tau\to u}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x(\tau)h(u)e^{-j\Omega(\tau+u)}d\tau du = X(\Omega)H(\Omega).$$

7. *Multiplication*: For the product of two signals holds

$$\mathrm{FT}\{x(t)h(t)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} x(t)\int_{-\infty}^{\infty} H(\theta)e^{j\theta t}d\theta e^{-j\Omega t}dt \tag{1.55}$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} H(\theta)X(\Omega - \theta)d\theta = X(\Omega) *_\Omega H(\Omega) = H(\Omega) *_\Omega X(\Omega).$$

Convolution in frequency domain is denoted by $*_\Omega$ with a factor of $1/2\pi$ being included.

8. *Parseval's theorem*: The inner products of the signals $x(t)$ and $y(t)$ in the time domain and the frequency domain satisfy the following relations

$$\int_{-\infty}^{\infty} x(t)y^*(t)dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\Omega)Y^*(\Omega)d\Omega, \tag{1.56}$$
$$\int_{-\infty}^{\infty} |x(t)|^2 dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} |X(\Omega)|^2 d\Omega.$$

9. *Differentiation* of a signal in the time domain corresponds to the Fourier transform multiplication by $j\Omega$,

$$\mathrm{FT}\left\{\frac{dx(t)}{dt}\right\} = \mathrm{FT}\left\{\frac{d}{dt}\left(\frac{1}{2\pi}\int_{-\infty}^{\infty} X(\Omega)e^{j\Omega t}d\Omega\right)\right\} = j\Omega X(\Omega). \tag{1.57}$$

10. *Integration*: The Fourier transform of an integral of the signal $x(t)$,

$$\int_{-\infty}^{t} x(\tau)d\tau,$$

can be calculated as the Fourier transform of the convolution of the signals $x(t)$ and $u(t)$,

$$x(t) *_t u(t) = \int_{-\infty}^{\infty} x(\tau)u(t-\tau)d\tau = \int_{-\infty}^{t} x(\tau)d\tau.$$

Then, the Fourier transform of the signal $x(t)$ integral is obtained from

$$\mathrm{FT}\left\{\int_{-\infty}^{t} x(\tau)d\tau\right\} = \mathrm{FT}\{x(t)\}\mathrm{FT}\{u(t)\} = \tag{1.58}$$
$$X(\Omega)\left(\frac{1}{j\Omega} + \pi\delta(\Omega)\right) = \frac{1}{j\Omega}X(\Omega) + \pi X(0)\delta(\Omega).$$

If the mean value of the signal $x(t)$ is zero, when $X(0) = 0$, a multiplication by $1/(j\Omega)$ in the Fourier transform domain corresponds to the signal integration in the time domain.

11. *An analytic part* of a signal $x(t)$, whose Fourier transform is $X(\Omega)$, is a signal with the Fourier transform defined by

$$X_a(\Omega) = \begin{cases} 2X(\Omega) & \text{for } \Omega > 0 \\ X(0) & \text{for } \Omega = 0 \\ 0 & \text{for } \Omega < 0 \end{cases}. \tag{1.59}$$

It can be written as

$$X_a(\Omega) = X(\Omega) + X(\Omega)\text{sign}(\Omega) = X(\Omega) + jX_h(\Omega), \tag{1.60}$$

where $X_h(\Omega)$ is the Fourier transform of the *Hilbert transform* of the signal $x(t)$. From Example 1.11, 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 the signal $x(t)$, in the time domain, takes the form

$$x_a(t) = x(t) + jx_h(t) = x(t) + x(t) *_t \frac{j}{\pi t} = x(t) + j\frac{1}{\pi} \underset{p.v.}{\int_{-\infty}^{\infty}} \frac{x(\tau)}{t-\tau} d\tau, \tag{1.61}$$

where $p.v.$ stands for Cauchy principal value of the considered integral.

1.4.3 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 limited duration (that is, $x(t) = 0$ for $|t| > T_0/2$). Then,

$$X(\Omega) = \int_{-T_0/2}^{T_0/2} x(t)e^{-j\Omega t} dt. \tag{1.62}$$

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

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

This periodic signal $x_p(t)$ can be expanded into Fourier series with the coefficients

$$X_n = \frac{1}{T} \int_{-T/2}^{T/2} x_p(t)e^{-j2\pi nt/T} dt. \tag{1.63}$$

If $T > T_0$ it is easy to conclude that

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

or

$$X_n = \frac{1}{T} X(\Omega)_{|\Omega=2\pi n/T}. \tag{1.64}$$

It means that the Fourier series coefficients are equal to 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 its periodic extension (that is, $T > T_0$). The sampling interval in frequency is

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

This sampling interval 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*. It states that: *the values of* $X(\Omega)$ *can be recovered for any* Ω*, from its samples* $X(2\pi n/T) = X_n T$ *if* $T > T_0$. 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 the periodic signal $x_p(t)$, formed by a periodic extension of $x(t)$ with period T, can be written as

$$x_p(t) = \sum_{n=-\infty}^{\infty} x(t+nT) = x(t) *_t \sum_{n=-\infty}^{\infty} \delta(t+nT). \tag{1.65}$$

The Fourier transform of this periodic signal is

$$X_p(\Omega) = \mathrm{FT}\left\{ x(t) *_t \sum_{n=-\infty}^{\infty} \delta(t+nT) \right\} \tag{1.66}$$

$$= X(\Omega) \cdot \frac{2\pi}{T} \sum_{n=-\infty}^{\infty} \delta\left(\Omega - \frac{2\pi}{T}n\right) = \frac{2\pi}{T} \sum_{n=-\infty}^{\infty} X\left(\frac{2\pi}{T}n\right) \delta\left(\Omega - \frac{2\pi}{T}n\right)$$

since

$$\mathrm{FT}\left\{ \sum_{n=-\infty}^{\infty} \delta(t+nT) \right\} = \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(t+nT) e^{-j\Omega t} dt$$

$$= \sum_{n=-\infty}^{\infty} e^{j\Omega nT} = \frac{2\pi}{T} \sum_{n=-\infty}^{\infty} \delta\left(\Omega - \frac{2\pi}{T}n\right). \tag{1.67}$$

The Fourier transform of a periodic signal is a series of generalized impulse signals at $\Omega = 2\pi n/T$, with weighting factors $X(\frac{2\pi}{T}n)/T$ being equal to the Fourier series coefficients X_n. The relation between the periodic generalized impulse signals in the time and frequency domain will be explained (derived) later, (see Example 2.8).

1.5 FOURIER TRANSFORM AND THE STATIONARY PHASE METHOD

When a signal

$$x(t) = A(t)e^{j\phi(t)} \tag{1.68}$$

is not of a simple analytic form, it may be possible, in some cases, to obtain an approximative expression for its Fourier transform 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

$$\int_{-\infty}^{\infty} A(t)e^{j\phi(t)}e^{-j\Omega t}dt \simeq A(t_0)e^{j\phi(t_0)}e^{-j\Omega t_0}\sqrt{\frac{2\pi j}{|\phi''(t_0)|}}, \tag{1.69}$$

where t_0 is the solution to

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

The most significant contribution to the integral on the left side of (1.69) 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 the instantaneous frequency of a signal. Around the stationary phase instant t_0 the following relation holds

$$\frac{d\,(\phi(t) - \Omega t)}{dt}_{|t=t_0} = 0$$
$$\phi'(t_0) - \Omega = 0.$$

In the vicinity of the stationary phase instant, t_0, the phase can be expanded into a Taylor series,

$$\phi(t) - \Omega t = [\phi(t_0) - \Omega t_0] + [\phi'(t_0) - \Omega] + \frac{1}{2}\phi''(t_0)t^2 + \dots$$

Since $\phi'(t_0) - \Omega = 0$ the integral in (1.69) can be written in the form

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

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

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

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 (1.69) 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 1.13. Consider signal

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

Find its Fourier transform approximation using the stationary phase method.

★According to the stationary phase method, the instant of stationary phase follows from $\phi'(t_0) - \Omega = 0$, that is

$$8\pi t_0 + 10\pi = \Omega$$
$$t_0 = \frac{\Omega - 10\pi}{8\pi}$$

and

$$\phi''(t_0) = 8\pi. \tag{1.70}$$

The amplitude of $X(\Omega)$ is

$$|X(\Omega)| \simeq A(t_0)\left|\sqrt{\frac{2\pi}{\phi''(t_0)}}\right| = \exp(-(t_0^2-1)t_0^2)\sqrt{\frac{2\pi}{8\pi}}$$
$$= \frac{1}{2}\exp\left(-\left[\left(\frac{\Omega-10\pi}{8\pi}\right)^2 - 1\right]\left(\frac{\Omega-10\pi}{8\pi}\right)^2\right) \tag{1.71}$$

The signal, stationary phase approximation of the Fourier transform (its amplitude) and the numerical value of the Fourier transform amplitudes are shown in Fig. 1.8

Example 1.14. 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 using the stationary phase method.

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

$$t_0 = \left(\frac{\Omega}{2Na}\right)^{1/(2N-1)}$$

and

$$\phi''(t_0) = 2N(2N-1)a\left(\frac{\Omega}{2Na}\right)^{(2N-2)/(2N-1)}. \tag{1.72}$$

The amplitude and phase of $X(\Omega)$, according to (1.69), are

$$|X(\Omega)|^2 \simeq A^2(t_0)\left|\frac{2\pi}{\phi''(t_0)}\right| \tag{1.73}$$
$$= A^2\left(\left(\frac{\Omega}{2Na}\right)^{1/(2N-1)}\right)\left|\frac{2\pi}{(2N-1)\Omega}\left(\frac{\Omega}{2aN}\right)^{1/(2N-1)}\right|$$
$$\arg\{X(\Omega)\} \simeq \phi(t_0) - \Omega t_0 + \pi/4 = \frac{(1-2N)}{2N}\Omega\left(\frac{\Omega}{2aN}\right)^{1/(2N-1)} + \pi/4$$

for a large value of a.

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

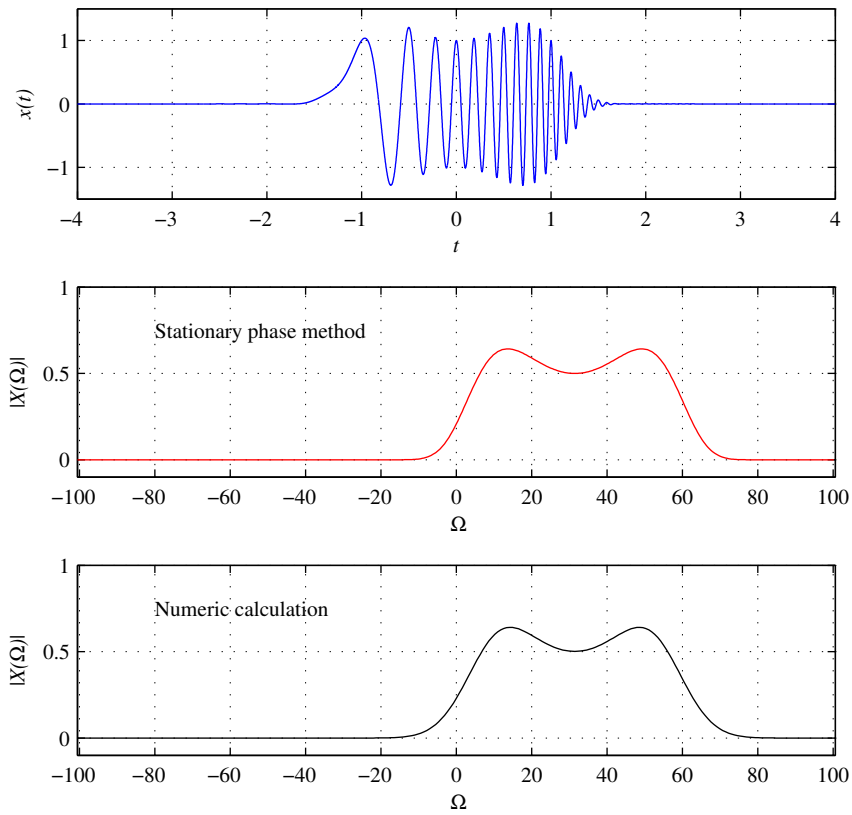


Figure 1.8 The signal (top), along with the stationary phase method approximation of its Fourier transform and the Fourier transform obtained by numeric calculation with a high precision (bottom).

The method of stationary phase may be defined in the frequency domain as well. For a Fourier transform

$$X(\Omega) = B(\Omega)e^{j\theta(\Omega)} \tag{1.74}$$

the method of stationary phase states that if the Fourier transform phase, $\theta(t)$, is monotonous and the amplitude, $B(t)$, is sufficiently smooth function, then

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} B(\Omega)e^{j\theta(\Omega)}e^{j\Omega t}d\Omega \simeq B(\Omega_0)e^{j\theta(\Omega_0)}e^{j\Omega_0 t}\sqrt{\frac{j}{2\pi\left|\theta''(\Omega_0)\right|}}\ , \tag{1.75}$$

where Ω_0 is the solution to

$$-\theta'(\Omega_0) = t,$$

and

$$t_g = -\theta'(\Omega)$$

is the group delay.

Example 1.15. Consider a system with the transfer function

$$H(\Omega) = \exp(-\Omega^2)\exp(-ja\Omega^2/2 - jb\Omega).$$

Find the impulse response of this system using the stationary phase method.

★According to the stationary phase method,

$$a\Omega_0 + b = t \quad \text{or} \quad \Omega_0 = \frac{t-b}{a}$$

and

$$\theta''(\Omega_0) = -a.$$

The impulse response is

$$h(t)) \simeq e^{-\Omega_0^2} e^{-ja\Omega_0^2/2 - jb\Omega_0 + j\Omega_0 t} \sqrt{\frac{j}{2\pi|\theta''(\Omega_0)|}} = e^{-\left(\frac{t-b}{a}\right)^2} e^{j((t-b)^2/(2a)+\pi/4)} \sqrt{\frac{1}{2\pi a}}.$$

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

Example 1.16. For a system with the frequency response $H(\Omega) = |H(\Omega)|e^{j0}$ the impulse response is $h(t)$. Find the impulse response of the systems whose transfer functions are:
(a) $H_a(\Omega) = |H(\Omega)|e^{-j4\Omega}$,
(b) $H_b(\Omega) = |H(\Omega)|e^{-j2\pi\Omega^2}$, and
(c) $H_c(\Omega) = |H(\Omega)|\left[\frac{3}{4} + \frac{1}{4}\cos(2\pi\Omega^2)\right]e^{j0}$.

★(a) The impulse response is

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

It is delayed with respect to $h(t)$ for $t_0 = 4$.
(b) In this case

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

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

$$h_b(t) = H\left(\frac{t}{4\pi}\right)e^{j(t^2/8\pi+\pi/4)}\sqrt{\frac{1}{8\pi^2}}.$$

The nonlinear group delay causes time scaling and form change of the impulse response.
Note: Check that $\int_{-\infty}^{\infty} h^2(t)dt = \int_{-\infty}^{\infty} h_b^2(t)dt$.
(c) Since $2\cos(2\pi\Omega^2) = \exp(j2\pi\Omega^2) + \exp(-j2\pi\Omega^2)$ we can write

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

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).

1.6 LAPLACE TRANSFORM

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

$$X(s) = \mathcal{L}\{x(t)\} = \int_{-\infty}^{\infty} x(t)e^{-st}dt, \tag{1.76}$$

where $s = \sigma + j\Omega$ is a complex number. It is obvious that the Fourier transform is a special case of a Laplace transform along the imaginary axis, where $\sigma = 0$ 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 ∞). A part of the complex s-plane where the Laplace transform exists (converges) is referred to as the *region of convergence* (ROC).

Example 1.17. Calculate the Laplace transform of $x(t) = e^{at}u(t)$, for a real-valued constant a.

★According to the definition

$$X(s) = \int_{0}^{\infty} e^{at}e^{-st}dt = -\frac{e^{-(s-a)t}\Big|_0^{\infty}}{s-a} = \frac{1}{s-a}$$

if $\lim_{t\to\infty} e^{-(s-a)t} = 0$ or $\sigma - a > 0$, that is, $\sigma > a$. The region of convergence of this Laplace transform, $X(s)$, is a part of the complex s-plane where $\sigma > a$. The point $s = a$ is the pole of the Laplace transform. The region of convergence cannot include any poles and it is limited by a vertical line in the complex s-plane passing through the pole, as shown in Fig. 1.9.

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

$$\mathrm{FT}\{x(t)e^{-\sigma t}\} = \int_{-\infty}^{\infty} x(t)e^{-\sigma t}e^{-j\Omega t}dt = \int_{-\infty}^{\infty} x(t)e^{-st}dt = X(s). \tag{1.77}$$

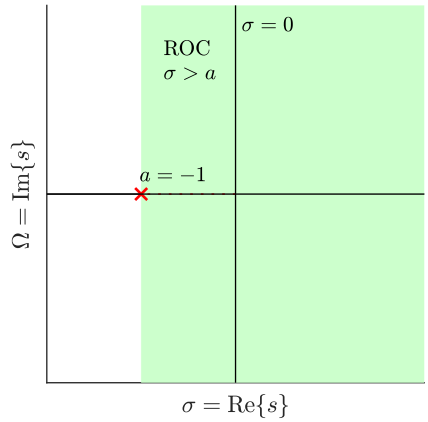


Figure 1.9 The region of convergence (ROC) of the Laplace transform of the signal $x(t) = e^{at}u(t)$ for $a = -1$.

In this way, we may calculate the Laplace transform of signals that are not absolutely integrable, that is, do not satisfy the condition for the Fourier transform convergence, $\int_{-\infty}^{\infty} |x(t)|\,dt < \infty$. For some values of σ, 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 sense of the generalized functions sense only. The 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.

Example 1.18. Find the Laplace transform of $x(t) = -e^{at}u(-t)$ and the region of its convergence.

★The Laplace transform of this signal is given by

$$X(s) = -\int_{-\infty}^{0} e^{at}e^{-st}dt = \frac{e^{-(s-a)t}\Big|_{-\infty}^{0}}{s-a} = \frac{1}{s-a}$$

if $\lim_{t\to-\infty} e^{-(s-a)t} = 0$ or $\sigma - \mathrm{Re}\{a\} < 0$, that is, $\sigma < \mathrm{Re}\{a\}$, where $\mathrm{Re}\{a\}$ is a real part of a. The Laplace transform $X(s)$ is this example has the same form as $X(s)$ in the previous example, with different regions of convergence. The Fourier transform of $x(t) = -e^{at}u(-t)$ exists if $\sigma = 0$ is within the region of convergence, that is, if $\mathrm{Re}\{a\} > 0$.

The inverse Laplace transform is defined by

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

where the integration is performed along a path, with γ within the region of convergence of $X(s)$.

1.6.1 Properties of the Laplace Transform

Properties of the Laplace transform may easily be generalized from those presented for the Fourier transform in Section 1.4.2, like for example, the linearity property and the convolution property, given by

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

and

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

Since the Laplace transform will be used to analyze linear systems described by linear differential equations we will pay a special attention to the relation of the signal derivatives to the corresponding forms in the Laplace domain. In general, the Laplace transform of the first derivative, $dx(t)/d(t)$, of a signal $x(t)$ is

$$\int_{-\infty}^{\infty} \frac{dx(t)}{dt} e^{-st} dt = x(t)e^{-st}\Big|_{-\infty}^{\infty} + s\int_{-\infty}^{\infty} x(t)e^{-st}dt = sX(s). \tag{1.78}$$

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

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

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

The region of convergence for the unilateral Laplace transform is the right-sided part of the s plane. This topic is discussed in Section 1.6.3.

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

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

The unilateral Laplace transform of the first derivative of a causal signal $x(t)$ is

$$\int_{0}^{\infty} \frac{dx(t)}{dt} e^{-st} dt = x(t)e^{-st}\Big|_{0}^{\infty} + s\int_{0}^{\infty} x(t)e^{-st}dt = sX(s) - x(0). \tag{1.79}$$

The value of signal $x(t)$ at $t = 0$, denoted by $x(0)$, is the initial condition.

The previous relation can easily be generalized to the higher-order signal derivatives of $x(t)$ and the corresponding Laplace transforms

$$\int_0^\infty \frac{d^n x(t)}{dt^n} e^{-st} dt = s^n \int_0^\infty x(t) e^{-st} dt - s^{n-1} x(0) - s^{n-2} x'(0) - \cdots - x^{(n-1)}(0)$$

$$= s^n X(s) - s^{n-1} x(0) - s^{n-2} x'(0) - \cdots - x^{(n-1)}(0) = s^n X(s) - \sum_{m=1}^{n} s^{n-m} x^{(m-1)}(0).$$

The unilateral Laplace transform of an integral with variable limit of the signal $x(t)$ is

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

since

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

The signal that corresponds to the derivative of the Laplace transform is obtained from

$$\frac{dX(s)}{ds} = \frac{d}{ds}\Big(\int_0^\infty x(t) e^{-st} dt\Big) = \int_0^\infty -t x(t) e^{-st} dt.$$

This means that

$$\mathcal{L}\{t x(t)\} = -\frac{dX(s)}{ds}.$$

Example 1.19. Find the Laplace transform of $x(t) = te^{at}u(t)$.

★The Laplace transform of $x(t)$ is obtained as

$$\mathcal{L}\{te^{at}u(t)\} = -\frac{d}{ds}\Big(\mathcal{L}\{e^{at}u(t)\}\Big) = -\frac{d}{ds}\Big(\frac{1}{s-a}\Big) = \frac{1}{(s-a)^2}.$$

This relation can easily be generalized to $\mathcal{L}\{t^n e^{at}u(t)\} = n!/(s-a)^{n+1}$.

Example 1.20. Find the Laplace transform of the signal $x(t) = e^{j\Omega_0 t}u(t)$.

★The Laplace transform of $e^{j\Omega_0 t}u(t)$ is

$$\mathcal{L}\{e^{j\Omega_0 t}u(t)\} = \int_0^\infty e^{j\Omega_0 t}e^{-st}dt = \frac{1}{s-j\Omega_0} = \frac{s+j\Omega_0}{s^2+\Omega_0^2},$$

for $\sigma > 0$. The Laplace transforms of $\cos(\Omega_0 t)u(t)$ and $\sin(\Omega_0 t)u(t)$ follow from the last relation as

$$\mathcal{L}\{\cos(\Omega_0 t)u(t)\} = \mathcal{L}\{e^{j\Omega_0 t}u(t)\}/2 + \mathcal{L}\{e^{-j\Omega_0 t}u(t)\}/2 = s/(s^2+\Omega_0^2),$$
$$\mathcal{L}\{\sin(\Omega_0 t)u(t)\} = \mathcal{L}\{e^{j\Omega_0 t}u(t)\}/2j - \mathcal{L}\{e^{-j\Omega_0 t}u(t)\}/2j = \Omega_0/(s^2+\Omega_0^2).$$

The initial value theorem and the final value theorem for the signal $x(t)$ are

$$x(0+) = \lim_{s\to\infty} sX(s) \text{ and } x(\infty) = \lim_{s\to 0} sX(s),$$

respectively. Both of them follow from (1.79). The requirement is that the Laplace transform of the signal, $x(t)$, and its derivative $dx(t)/dt$, exist. The final value of the signal does not exist if the poles of $sX(s)$ are: (a) on the right side of the s plane, (b) there is a pair of the conjugate-complex poles on the imaginary axis, (c) at the origin. The initial value theorem requires that the signal does not contain delta pulses at the origin.

1.6.2 Table of the Laplace Transform

Signal $x(t)$	Laplace transform $X(s)$
$\delta(t)$	1
$u(t)$	$1/s$
$e^{at}u(t)$	$\frac{1}{s-a}$
$tu(t)$	$1/s^2$
$e^{at}\cos(\Omega_0 t)u(t)$	$\frac{s-a}{(s-a)^2+\Omega_0^2}$
$e^{at}\sin(\Omega_0 t)u(t)$	$\frac{\Omega_0}{(s-a)^2+\Omega_0^2}$
$te^{at}u(t)$	$\frac{1}{(s-a)^2}$
$x'(t)u(t)$	$sX(s) - x(0)$
$tx(t)u(t)$	$-dX(s)/ds$
$x(t)u(t)/t$	$\int_s^\infty F(s)ds$
$e^{at}x(t)u(t)$	$X(s-a)$
$x(t) *_t u(t) = \int_0^t x(t)dt$	$X(s)/s$

1.6.3 Linear Systems Described by Differential Equations

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

$$a_N \frac{d^N y(t)}{dt^N} + \cdots + a_1 \frac{dy(t)}{dt} + a_0 y(t) = b_M \frac{d^M x(t)}{dt^M} + \cdots + b_1 \frac{dx(t)}{dt} + b_0 x(t), \quad (1.80)$$

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

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

The transfer function of this system is of the form

$$H(s) = \frac{Y(s)}{X(s)} = \frac{b_M s^M + \cdots + b_1 s + b_0}{a_N s^N + \cdots + a_1 s + a_0}. \quad (1.81)$$

Stability and causality. A linear time invariant system is stable if its impulse response $h(t)$ satisfies the condition $\int_{-\infty}^{\infty} |h(t)|\, dt < \infty$. Within the Laplace transform framework, this condition means that the line $\sigma = 0$, in the complex s-plane, belongs to the region of convergence of the transfer function, $H(s)$.

A system whose impulse response is of the form

$$h(t) = \sum_{n=1}^{N} A_n e^{a_n t} u(n)$$

is causal. Although, this is not a general form of causal systems, it is important in system analysis. The transfer function of this system is given by

$$H(s) = \sum_{n=1}^{N} \frac{A_n}{s - a_n}, \quad (1.82)$$

with the region of convergence defined by the set of inequalities $\sigma > \mathrm{Re}\{a_1\}$, $\sigma > \mathrm{Re}\{a_2\}$, $\ldots$, $\sigma > \mathrm{Re}\{a_N\}$. These inequalities can be written in a compact form as

$$\sigma > \max_n \mathrm{Re}\{a_n\}.$$

The region of convergence of the causal system in (1.82) is the right side of the line $\sigma = \max_n \mathrm{Re}\{a_n\}$, passing trough the pole with the largest real value part.

The system defined by (1.81) can be written in the form given by (1.82) if the polynomial order in the denominator is higher that the nominator polynomial order, that is, if $N > M$. This system is causal and stable if all poles a_n reside in the left side of the complex s-plane, that is, if $\mathrm{Re}\{a_n\} < 0$ for all $n = 1, 2, \ldots, N$ and the region of convergence is defined by

$\sigma > \max_n \mathrm{Re}\{a_n\}$, as illustrated in Fig. 1.10. Possible higher-order (multiple) poles in $H(s)$ would not change the conclusion about its causality.

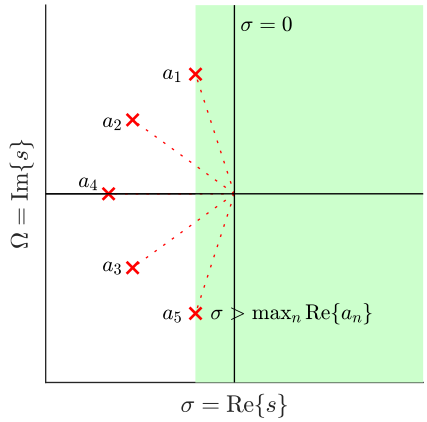


Figure 1.10 Poles of a stable and causal system, with its region of convergence, $\sigma > \max_n \mathrm{Re}\{a_n\}$, that includes the line $\sigma = 0$.

Example 1.21. A causal system with a proportional regulator is described with the transfer function,

$$H(s) = \frac{K}{s^2 + 4s + K},$$

where K is the constant of the regulator. Find the system response to the input signal $x(t) = u(t)$, for $K = 4$, $K = 3 < 4$, and $K = 20 > 4$.

★ The Laplace transform of the input signal is $X(s) = \mathcal{L}\{u(t)\} = 1/s$. The poles of the transfer function $H(s)$ are obtained from $s^2 + 4s + K = 0$ as

$$s_{1,2} = -2 \pm \sqrt{4 - K}.$$

For $K = 4$ the Laplace transform of the output signal is

$$Y(s) = \frac{4}{s(s+2)^2} = \frac{A}{s} + \frac{B}{s+2} + \frac{C}{(s+2)^2} = \frac{1}{s} - \frac{1}{s+2} - \frac{2}{(s+2)^2},$$

while the system response in the time domain is obtained as the inverse Laplace transform of $Y(s)$ and it is given by

$$y(t) = (1 - e^{-2t} - 2te^{-2t})u(t).$$

The convergence toward the steady-state is of the $e^{-2t}(1+2t)$ form (*critically dumped response*).

For $K = 3$ the Laplace transform, $Y(s)$, and the time domain signal, $y(t)$, are given by

$$Y(s) = \frac{3}{s(s+1)(s+3)} = \frac{A}{s} + \frac{B}{s+1} + \frac{C}{s+3} = \frac{1}{s} - \frac{3/2}{s+1} + \frac{1/2}{s+3},$$

and

$$y(t) = (1 - \frac{3}{2}e^{-t} + \frac{1}{2}e^{-3t})u(t).$$

In this case, the slowest converging term is of the form e^{-t}. The convergence of this term toward the steady-state is slower than in the case with $K = 4$ (*overdumped response*).

Finally, for $K = 20$ the outputs in the Laplace domain and the time domain are of the form

$$Y(s) = \frac{20}{s(s+2+j4)(s+2-j4)} = \frac{1}{s} - \frac{(2+j)/4}{s+2+j4} - \frac{(2-j)/4}{s+2-j4},$$

and

$$y(t) = (1 - \frac{2+j}{4}e^{-2t(1+2j)} - \frac{2-j}{4}e^{-2t(1-2j)})u(t) = (1 - e^{-2t}\cos(4t) - \frac{1}{2}e^{-2t}\sin(4t))u(t).$$

The convergence toward the steady-state is defined by the function e^{-2t}, with the oscillatory term $\sin(4t)$ (*underdumped response* that overshoots its final value).

The responses of the system for the three considered cases are shown in Fig. 1.11.

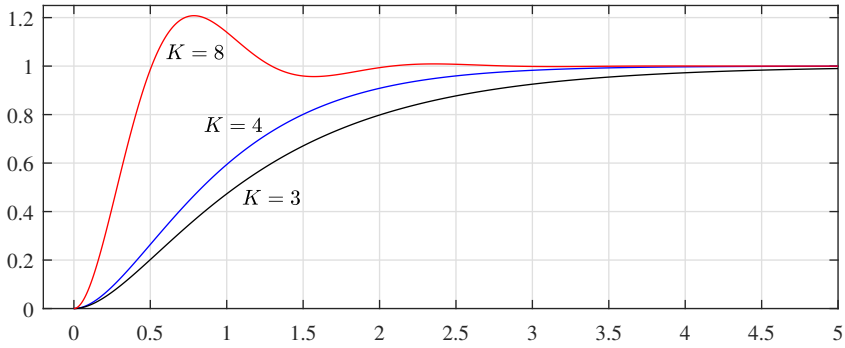


Figure 1.11 The responses of the system for the three considered cases: $K = 4$ (critically dumped), $K = 3$ (overdumped), and $K = 20$ (underdumped)

Example 1.22. Transfer function, $H(s)$, of the system is defined by

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

Find the impulse response if the system is stable. Is this system causal? Find the impulse response of a causal system with the same transfer function $H(s)$.

★ The poles of this transfer function are $s_1 = -2$, $s_2 = -1$, and $s_2 = 1$. The system is stable if the line $\sigma = 0$ belongs to the region of convergence. This region of convergence is defined by $-1 < \sigma < 1$. In order to find the inverse Laplace transform, the transfer function can be written

in the form

$$H(s) = \frac{1}{s+2} + \frac{1}{s+1} + \frac{1}{s-1}.$$

The impulse response for the region of convergence defined by $-1 < \sigma < 1$ is

$$h(t) = e^{-2t}u(t) + e^{-t}u(t) - e^{t}u(-t).$$

It is easy to check that the stability condition, $\int_{-\infty}^{\infty} |h(t)|\,dt = 5/2 < \infty$, holds for this system. This system is not causal, since $e^t u(-t)$ is not zero for $t < 0$.

Since the system is of form (1.81), with $N = 3 > M = 2$, it would be causal if the region of convergence is defined by $\sigma > \max_n \mathrm{Re}\{a_n\} = 1$. In this case, the impulse response would be of the form

$$h(t) = e^{-2t}u(t) + e^{-t}u(t) + e^{t}u(t).$$

This system is not stable since the last term tends to infinity as t increases.

Solution to the differential equations using the Laplace transform. The output of a linear-time invariant system described by (1.80) can be found by solving the corresponding differential equation. The Laplace transform approach to solve differential equations is of crucial importance in engineering. In general, if the initial conditions are included in (1.80), the corresponding Laplace transform domain equation is

$$\begin{aligned} a_N s^N Y(s) + \cdots + a_1 s Y(s) + a_0 Y(s) - \sum_{n=0}^{N} a_n \Big(\sum_{m=1}^{n} s^{n-m} x^{(m-1)}(0) \Big) \\ = b_M s^M X(s) + \cdots + b_1 s X(s) + b_0 X(s). \end{aligned}$$

The Laplace transform of the solution (output signal) can be written in the form

$$Y(s) = \frac{B(s)}{A(s)} X(s) + \frac{C(s)}{A(s)},$$

where $A(s) = a_N s^N + \cdots + a_1 s + a_0$, $B(s) = b_M s^M + \cdots + b_1 s + b_0$, and

$$C(s) = \sum_{n=0}^{N} a_n \Big(\sum_{m=1}^{n} s^{n-m} x^{(m-1)}(0) \Big).$$

The output consists of two parts, $Y(s) = Y_p(s) + Y_h(s)$, whose form is defined as follows:

- The part of the output caused by the input signal is given by

$$Y_p(s) = \frac{B(s)}{A(s)} X(s),$$

and it is called the *forced response* (in mathematics, the *particular* part of the differential equation solution).

- The output signal part due to the initial conditions is

$$Y_h(s) = \frac{C(s)}{A(s)}.$$

This part is independent of the input signal and it is called the *natural response* (in mathematics, the *homogeneous* part of the solution).

Example 1.23. A first-order causal system is described by the differential equation

$$\frac{dy(t)}{dt} + 2y(t) = 3u(t),$$

with the initial condition $y(0) = 0$. Find the system output, $y(t)$.

★The Laplace transform of both sides is

$$[sY(s) - y(0)] + 2Y(s) = \frac{3}{s}.$$

The Laplace transform $Y(s)$ of the output signal $y(t)$ is obtained in the form

$$Y(s) = \frac{3}{s(s+2)} = \frac{A}{s} + \frac{B}{s+2}.$$

The coefficients A and B are obtained from

$$3 = A(s+2) + Bs.$$

For $s = 0$, the coefficient A is obtained, $A = 3/2$, while $s = -2$ produces $B = -3/2$.

The output signal, $y(t)$, follows as the inverse Laplace transform of $Y(s)$,

$$y(t) = \frac{3}{2}(1 - e^{-2t})u(t).$$

Example 1.24. A causal system is described by the differential equation

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

with the initial conditions $y\prime(0) = 1$ and $y(0) = 0$. Find the system output, $y(t)$, for the input signal $x(t) = e^{-4t}u(t)$.

★The Laplace transform of both sides (including the initial conditions) is

$$[s^2Y(s) - sy(0) - y'(0)] + 3[sY(s) - y(0)] + 2Y(s) = X(s)$$

or

$$Y(s)(s^2+3s+2) = X(s) + sy(0) + y'(0) + 3y(0).$$

The Laplace transform of $x(t) = e^{-4t}u(t)$ is equal to $X(s) = 1/(s+4)$. The Laplace transform of the output signal is equal to

$$Y(s) = \frac{s+5}{(s+4)(s^2+3s+2)} = \frac{A_1}{s+4} + \frac{A_2}{s+2} + \frac{A_3}{s+1}.$$

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, $y(t)$, is the inverse Laplace transform of $Y(s)$, that is

$$y(t) = \frac{1}{6}e^{-4t}u(t) - \frac{3}{2}e^{-2t}u(t) + \frac{4}{3}e^{-t}u(t).$$

Note that $\frac{1}{6}e^{-4t}u(t) = y_p(t)$ is the forced response and $-\frac{3}{2}e^{-2t}u(t) + \frac{4}{3}e^{-t}u(t) = y_h(t)$ is the natural response in the time domain.

1.7 BUTTERWORTH FILTER

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

The squared frequency response of the *Butterworth lowpass filter* is defined by

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

It is shown in Fig. 1.12, for various N. 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 Ω_c. At the frequency Ω equal to the critical frequency, $\Omega = \Omega_c$, we get

$$|H(j\Omega_c)|^2 = \frac{1}{2}|H(0)|^2 = \frac{1}{2},$$

corresponding to $10\log(1/2) = -3[dB]$ gain, for any filter order N.

The squared frequency response may be written as

$$H(j\Omega)H(-j\Omega) = \frac{1}{1+\left(\frac{j\Omega}{j\Omega_c}\right)^{2N}}$$

The corresponding Laplace domain form is given by

$$H(s)H(-s) = \frac{1}{1+\left(\frac{s}{j\Omega_c}\right)^{2N}} \quad \text{for} \quad s = j\Omega.$$

Poles of the product of the transfer functions, $H(s)H(-s)$, are of the form

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

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

The poles of the product $H(s)H(-s)$ of the transfer function $H(s)$ of the Butterworth filter and its reversed version $H(-s)$ are located on the circle whose radius is Ω_c and at the positions defined by the phases

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

For a given filter order N and the critical frequency Ω_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 belongs to $H(-s)$. Since we want that the designed filter is stable and causal then we chose the poles

$$s_0,\ s_1,\ \ldots\ ,s_{N-1}$$

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

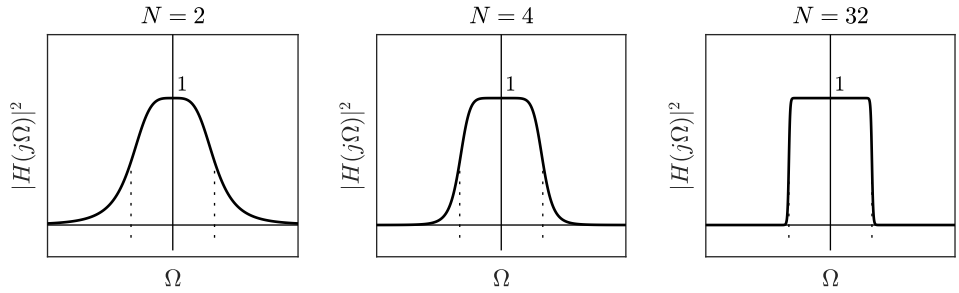


Figure 1.12 Squared amplitude of the frequency response of a Butterworth filter for various orders N.

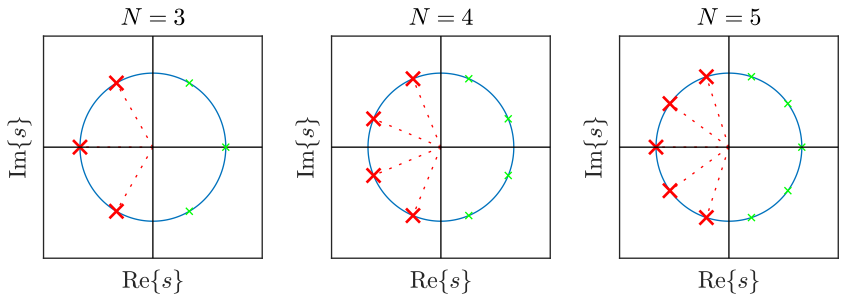


Figure 1.13 Poles of a stable Butterworth filter for $N=3$, $N=4$, and $N=5$.

Example 1.25. Design a lowpass Butterworth filter with the following filter order, N, and critical frequency, Ω_c,

(a) $N=3$ and $\Omega_c=1$,

(b) $N=4$ and $\Omega_c=3$.

★(a) The poles for the third-order filter, $N=3$, and critical frequency $\Omega_c=1$, have the phases

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

The pole values are

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

with the third-order Butterworth filter transfer function

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

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

(b) Poles for $N=4$ and $\Omega_c=3$ are with phases

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

Their values are

$$s_0=3\cos(\frac{\pi}{2}+\frac{\pi}{8})+j3\sin(\frac{\pi}{2}+\frac{\pi}{8}),\qquad s_1=3\cos(\frac{\pi}{2}+\frac{3\pi}{8})+j3\sin(\frac{\pi}{2}+\frac{3\pi}{8})$$
$$s_2=3\cos(\frac{\pi}{2}+\frac{5\pi}{8})+j3\sin(\frac{\pi}{2}+\frac{5\pi}{8}),\qquad s_3=3\cos(\frac{\pi}{2}+\frac{7\pi}{8})+j3\sin(\frac{\pi}{2}+\frac{7\pi}{8}),$$

with the fourth-oder Butterworth filter transfer function given by

$$H(s) = \frac{c}{(s^2+2.296s+9)(s^2+5.543s+9)} = \frac{81}{(s^2+2.296s+9)(s^2+5.543s+9)},$$

where $c = 81$ is used to make $H(0) = 1$.

In practice, we usually do not know the filter order N, but the passband frequency Ω_p and the stopband frequency Ω_s, of the filter, with the maximum attenuation in the passband $a_p[dB]$ and the minimum attenuation in the stopband $a_s[dB]$, as shown in Fig. 1.14. Based on these values we can calculate the filter order, N, and the critical frequency, Ω_c, needed for the filter design.

The passband and stopband relations for N and Ω_c are

$$\frac{1}{1+\left(\frac{\Omega_p}{\Omega_c}\right)^{2N}} \geq A_p^2 \tag{1.83}$$

$$\frac{1}{1+\left(\frac{\Omega_s}{\Omega_c}\right)^{2N}} \leq A_s^2, \tag{1.84}$$

where A_p and A_s are the required amplitudes of the frequency response at the respective passband and stopband frequencies, Ω_p and Ω_s. The relation

$$a = 20\log A$$

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

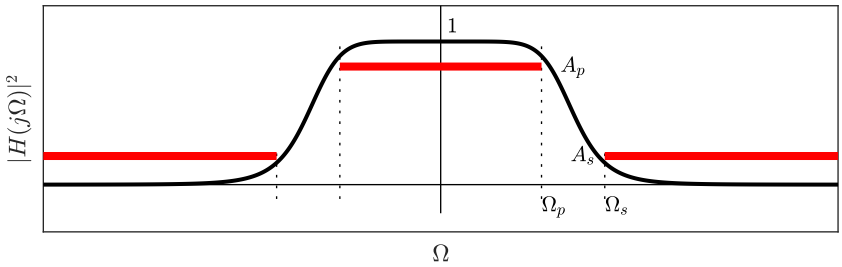


Figure 1.14 Specification of the Butterworth filter parameters in the passband and stopband.

Using the equality in both relations, (1.83) and (1.84), the order N follows

$$N = \frac{1}{2}\frac{\ln(\frac{1}{A_p^2}-1) - \ln(\frac{1}{A_s^2}-1)}{\ln\Omega_p - \ln\Omega_s}.$$

The nearest greater integer is assumed for the filter order N. Next, we can use any of the relations in (1.83) or (1.84) with the equality sign to calculate Ω_c. If we choose the first

one, then the critical frequency Ω_c will satisfy $\left|H(j\Omega_p)\right|^2 = A_p^2$, while if we use the second relation, the value of Ω_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 passband and stopband.

All other filter forms, like bandpass 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.

Part II

Deterministic Discrete-Time Signals and Systems

Chapter 2

Discrete-Time Signals and Transforms

THE first step in numerical processing of signals is in their discretization in time. A continuous-time signal is converted into a sequence of numbers, defining the discrete-time signal. The basic definitions of discrete-time signals and their transforms are presented in this chapter. The key fact in the conversion from a continuous-time signal into a sequence of numbers is that these two signal representations are equivalent under certain conditions. The discrete-time signal may contain the same information as the original continuous-time signal. The sampling theorem is fundamental for this relation between two signal forms. It is presented in this chapter, after basic definitions of discrete-time signals and systems are introduced.

2.1 DISCRETE-TIME SIGNALS

Discrete-time signals (discrete signals) are represented in a form of an ordered set of numbers $\{x(n)\}$. Commonly, they are obtained by sampling continuous-time signals. There exist discrete-time signals whose independent variable is inherently discrete in nature as well.

In the case that a discrete-time signal is obtained by sampling a continuous-time signal, we can write (Fig. 2.1),

$$x(n) = x(t)_{|t=n\Delta t}\,\Delta t. \tag{2.1}$$

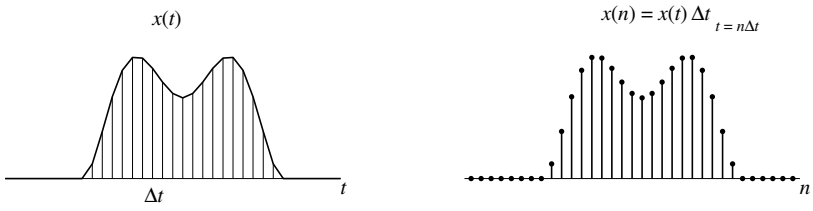


Figure 2.1 Signal discretization: continuous-time signal (left) and corresponding discrete-time signal (right).

Discrete-time signals are defined for an integer value of the argument n. 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 τ for continuous time and n and m for discrete time. Also, we hope that the context will always be clear, so that there is no doubt what kind of signal is considered. Notation $x[n]$ is sometimes used in literature for discrete-time signals, instead of $x(n)$.

Examples of discrete-time signals are presented next.

The discrete-time impulse signal is defined by

$$\delta(n) = \begin{cases} 1, \text{ for } n = 0 \\ 0, \text{ for } n \neq 0 \end{cases}. \tag{2.2}$$

It is presented in Fig. 2.2. In contrast to the continuous-time impulse signal, that cannot be

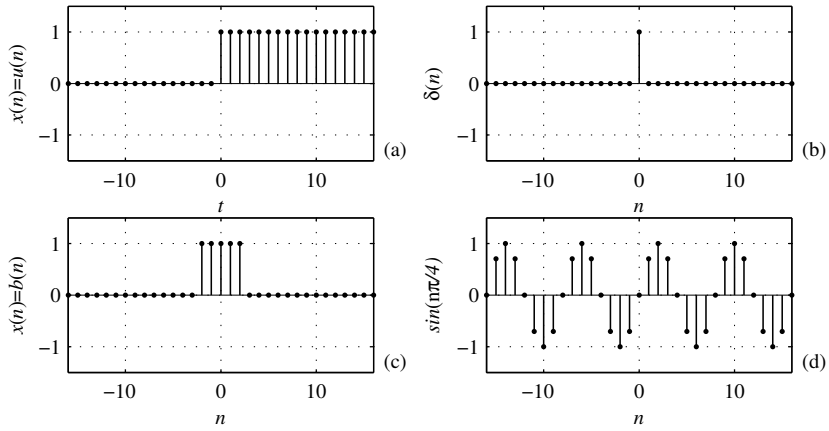


Figure 2.2 Illustration of discrete-time signals: (a) unit-step function, (b) discrete-time impulse signal, (c) boxcar signal $b(n) = u(n+2) - u(n-3)$, and (d) discrete-time sinusoid.

practically implemented and used, the discrete-time unit impulse is a signal that can easily be implemented and used in realizations. In mathematical notation, this signal corresponds to the Kronecker delta function

$$\delta_{m,n} = \begin{cases} 1, \text{ for } m = n \\ 0, \text{ for } m \neq n. \end{cases} \tag{2.3}$$

Any discrete-time signal can be written in a form of a sum of shifted and weighted discrete-time impulses,

$$x(n) = \sum_{k=-\infty}^{\infty} x(k)\delta(n-k), \tag{2.4}$$

as illustrated in Fig. 2.3.

The discrete unit-step signal is defined by

$$u(n) = \begin{cases} 1, \text{ for } n \geq 0 \\ 0, \text{ for } n < 0 \end{cases}. \tag{2.5}$$

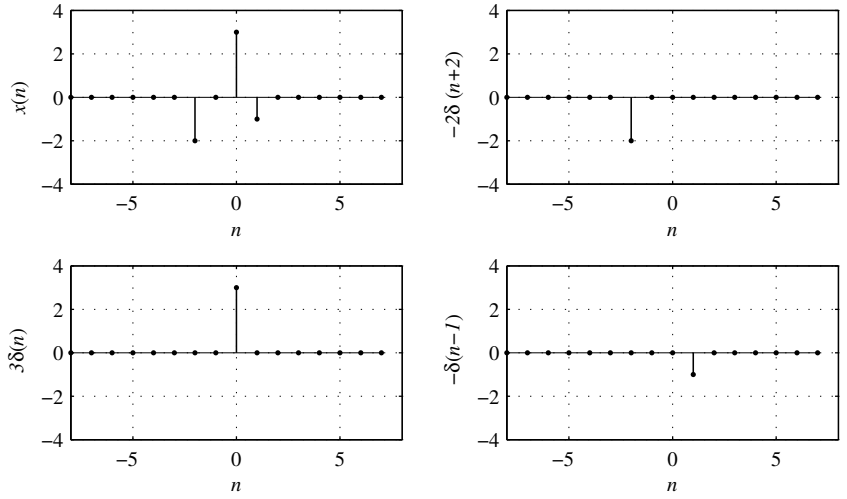


Figure 2.3 Signal $x(n)$ along with corresponding discrete-time impulses.

The discrete-time impulse and the unit-step signal are related as

$$\delta(n) = u(n) - u(n-1)$$
$$u(n) = \sum_{k=-\infty}^{n} \delta(k).$$

The discrete-time complex sinusoidal signal is defined by

$$x(n) = Ae^{j(\omega_0 n+\varphi)} = A\cos(\omega_0 n + \varphi) + jA\sin(\omega_0 n + \varphi). \tag{2.6}$$

A discrete-time signal is periodic if there exists an integer N such that

$$x(n+N) = x(n). \tag{2.7}$$

Smallest positive integer N that satisfies this equation is called the period of the discrete-time signal $x(n)$. Note that the signal $x(n)$ with a period N is also periodic with any integer multiple of N. Some basic discrete-time signals are presented in Fig. 2.2.

Example 2.1. Check the periodicity of discrete-time signals $x_1(n) = \sin(2\pi n/36)$, $x_2(n) = \cos(4\pi n/15+2)$, $x_3(n) = \exp(j0.1n)$, $x_4(n) = x_1(n) + x_2(n)$, and $x_5(n) = x_1(n) + x_3(n)$.

★Period of the discrete-time signal $x_1(n) = \sin(2\pi n/36)$ is obtained from $2\pi N_1/36 = 2\pi k$, where k is an integer. It is $N_1 = 36$, for $k = 1$. The period N_2 follows from $4\pi N_2/15 = 2\pi k$ as $N_2 = 15$ with $k = 2$. Period of signal $x_3(n)$ should be calculated from $0.1N_3 = 2\pi k$. Obviously,

there is no integer k such that N_3 is an integer. This signal is not periodic. The same holds for $x_5(n)$. The period of $x_4(n)$ is a common period for signals $x_1(n)$ and $x_2(n)$ with $N_1 = 36$ and $N_2 = 15$. It is $N_4 = 180$.

A discrete-time signal is even if

$$x(n) = x(-n).$$

For an odd signal holds

$$x(n) = -x(-n).$$

Example 2.2. Show that a discrete-time signal may be written as a sum

$$x(n) = x_e(n) + x_o(n)$$

where $x_e(n)$ and $x_o(n)$ are its even and odd part, respectively.

★For a signal $x(n)$ we can form its even and odd part as

$$x_e(n) = \frac{x(n) + x(-n)}{2} \quad \text{and} \quad x_o(n) = \frac{x(n) - x(-n)}{2}.$$

Summing these two parts, the signal $x(n)$ is reconstructed. Note that $x_o(0) = 0$.

A signal is Hermitian if

$$x(n) = x^*(-n).$$

Magnitude of a discrete-time signal is defined as the maximum value of the signal amplitude

$$M_x = \max_{-\infty<n<\infty} |x(n)|.$$

Energy of discrete-time signals is defined by

$$E_x = \sum_{n=-\infty}^{\infty} |x(n)|^2. \tag{2.8}$$

The instantaneous power of $x(n)$ is $P_x(n) = |x(n)|^2$, while the average signal power is

$$P_{AV} = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} |x(n)|^2 = \left\langle |x(n)|^2 \right\rangle, \tag{2.9}$$

where $\left\langle |x(n)|^2 \right\rangle$ is used to denote an average over large number of signal values, as $N \to \infty$. The average power of signals with a finite energy (energy signals) is $P_{AV} = 0$. For power signals (when $0 < P_{AV} < \infty$) the energy is infinite, $E_x \to \infty$.

Example 2.3. The energy of signal $x(n)$ is $E_x = 10$. The energy of its even part is $E_{x_e} = 3$. Find the energy of its odd part.

★The energy of signal is

$$E_x = \sum_{n=-\infty}^{\infty} |x(n)|^2 = \sum_{n=-\infty}^{\infty} |x_e(n) + x_o(n)|^2 = \sum_{n=-\infty}^{\infty} [x_e(n) + x_o(n)][x_e(n) + x_o(n)]^*$$
$$= \sum_{n=-\infty}^{\infty} |x_e(n)|^2 + \sum_{n=-\infty}^{\infty} |x_o(n)|^2 + \sum_{n=-\infty}^{\infty} [x_o(n)x_e^*(n) + x_e(n)x_o^*(n)].$$

The terms $x_o(n)x_e^*(n)$ and $x_e(n)x_o^*(n)$ in the last sum correspond to odd signals

$$x_o(-n)x_e^*(-n) = -x_o(n)x_e^*(n)$$
$$x_e(-n)x_o^*(-n) = -x_e(n)x_o^*(n).$$

Their sum is zero,

$$\sum_{n=-\infty}^{\infty} x_o(n)x_e^*(n) = \sum_{n=-\infty}^{\infty} x_e(n)x_o^*(n) = 0.$$

For the signals $x_e(n)$ and $x_o(n)$, satisfying the previous relation, we say that they are orthogonal. Therefore, for the energies E_x, E_{x_e}, and E_{x_o}, holds

$$E_x = E_{x_e} + E_{x_o}.$$

Obviously $E_{x_0} = E_x - E_{x_e} = 7$.

2.2 DISCRETE-TIME SYSTEMS

Discrete-time (discrete) system transforms one discrete-time signal (input) into the other (output signal)

$$y(n) = \mathbb{T}\{x(n)\}. \tag{2.10}$$

A discrete system $\mathbb{T}\{\cdot\}$ is linear if for any two signals $x_1(n)$ and $x_2(n)$ and any two constants a_1 and a_2 holds

$$y(n) = \mathbb{T}\{a_1x_1(n) + a_2x_2(n)\} = a_1\mathbb{T}\{x_1(n)\} + a_2\mathbb{T}\{x_2(n)\}. \tag{2.11}$$

A discrete system is time-invariant if for

$$y(n) = \mathbb{T}\{x(n)\} \tag{2.12}$$

holds

$$\mathbb{T}\{x(n - n_0)\} = y(n - n_0),$$

for any n_0.

For any input signal $x(n)$ the signal at the output of a linear time-invariant discrete system can be calculated if we know the output to the impulse signal. The output to the impulse signal, $h(n) = \mathbb{T}\{\delta(n)\}$, is the impulse response.

The output to an input signal $x(n)$ is

$$y(n) = \mathbb{T}\{x(n)\} = \mathbb{T}\left\{\sum_{k=-\infty}^{\infty} x(k)\delta(n-k)\right\}.$$

For a linear time-invariant discrete system we get

$$y(n) = \sum_{k=-\infty}^{\infty} x(k)\mathbb{T}\{\delta(n-k)\} = \sum_{k=-\infty}^{\infty} x(k)h(n-k). \tag{2.13}$$

This is a discrete-time convolution. Its notation is

$$x(n) *_n h(n) = \sum_{k=-\infty}^{\infty} x(k)h(n-k). \tag{2.14}$$

Discrete-time convolution is a commutative operation,

$$x(n) *_n h(n) = h(n) *_n x(n). \tag{2.15}$$

Example 2.4. Calculate discrete-time convolution of signals $x(n)$ and $h(n)$ shown in Fig. 2.4.

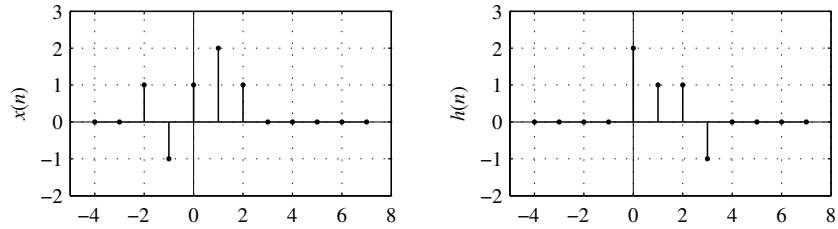


Figure 2.4 Input signal and impulse response.

★By definition, according to Fig. 2.5, we have

$$y(0) = \sum_{k=-\infty}^{\infty} x(k)h(-k) = 1 - 1 + 2 = 2,$$

$$y(1) = \sum_{k=-\infty}^{\infty} x(k)h(1-k) = -1 - 1 + 1 + 4 = 3.$$

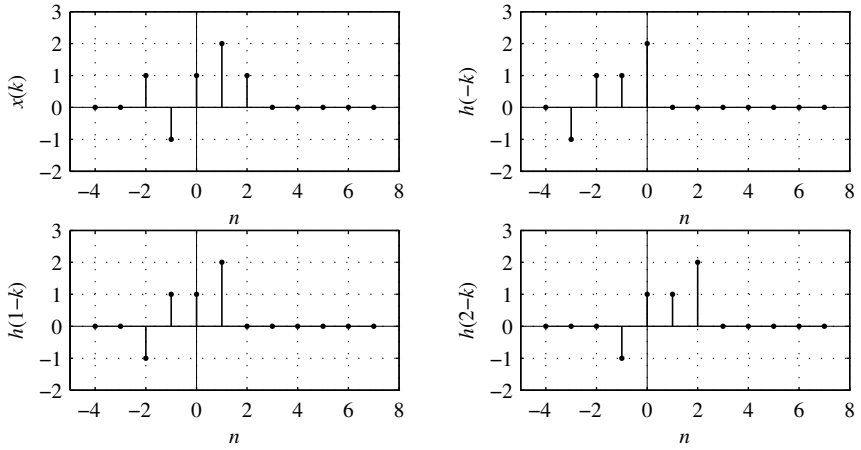


Figure 2.5 The signals $x(k)$, $h(-k)$, $h(1-k)$, and $h(2-k)$ used for the calculation of the output signal values $y(0)$, $y(1)$, and $y(2)$.

In a similar way $y(-2)=2$, $y(-1)=-1$, $y(2)=6$, $y(3)=2$, $y(4)=-1$, $y(5)=-1$, and $y(n)=0$, for all other n. The convolution $y(n)$ is shown in Fig. 2.6.

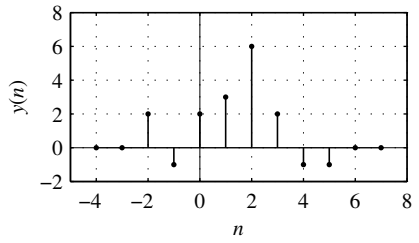


Figure 2.6 The output signal $y(n)=x(n)*h(n)$.

Example 2.5. Calculate the convolution of signals $x(n)=n[u(n)-u(n-10)]$ and $h(n)=u(n)$.

★The convolution of these two signals is

$$y(n)=\sum_{k=-\infty}^{\infty} k[u(k)-u(k-10)]u(n-k)$$

Using the fact that $u(k)-u(k-10)=1$ for $0\leq k\leq 9$ and $u(n-k)=1$ for $k\leq n$ we get

$$y(n)=\sum_{\substack{0\leq k\leq 9\\ k\leq n}} k=\begin{cases}\sum\limits_{k=0}^{n} k=n\frac{n+1}{2} & \text{for} \quad 0\leq n\leq 9\\ \sum\limits_{k=0}^{9} k=45 & \text{for} \quad n>9\end{cases}$$
$$=n\frac{n+1}{2}[u(n)-u(n-10)]+45u(n-10).$$

Example 2.6. If the response of a linear time-invariant system to the unit-step is $y(n)=\mathbb{T}\{u(n)\}=e^{-n}u(n)$ find the impulse response $h(n)$ of this system.

★The impulse response is

$$h(n)=\mathbb{T}\{\delta(n)\}=\mathbb{T}\{u(n)-u(n-1)\}=\mathbb{T}\{u(n)\}-\mathbb{T}\{u(n-1)\}$$
$$=e^{-n}u(n)-e^{-(n-1)}u(n-1)=e^{-n}[u(n)-e\,u(n-1)].$$

A discrete system is *causal* if there is no response before the input signal appears. For causal linear time-invariant discrete systems $h(n)=0$ for $n<0$ holds. For a signal that may be an impulse response of a causal system we say that it is a causal signal or one-sided signal.

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

$$\sum_{m=-\infty}^{\infty}|h(m)|<\infty. \tag{2.16}$$

The output of a linear time-invariant system is

$$|y(n)|=|\sum_{m=-\infty}^{\infty}x(n-m)h(m)|\leq\sum_{m=-\infty}^{\infty}|x(n-m)||h(m)|$$
$$\leq M_x\sum_{m=-\infty}^{\infty}|h(m)|.$$

Therefore $|y(n)|<\infty$ if (2.16) holds. It can be shown that the absolute sum convergence of the impulse response $h(n)$ is the necessary condition for a linear time-invariant discrete system to be stable as well.

2.3 FOURIER TRANSFORM OF DISCRETE-TIME SIGNALS

The Fourier transform of a discrete-time signal is defined by

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n}. \tag{2.17}$$

Notation $X(e^{j\omega})$ is used to emphasize the fact that it is a periodic function of the normalized frequency ω. The period is 2π.

In order to establish the relation between the Fourier transform of discrete-time signals and the Fourier transform of continuous-time signals,

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

we will write an approximation of the Fourier transform of continuous-time signal according to the rectangular rule of numerical integration,

$$X(\Omega) \cong \sum_{n=-\infty}^{\infty} x(n\Delta t)e^{-j\Omega n\Delta t}\Delta t. \tag{2.18}$$

By using the notation

$$x(n\Delta t)\Delta t \longrightarrow x(n) \quad \text{and} \quad \Omega\Delta t \longrightarrow \omega, \tag{2.19}$$

the previous approximation can be written as

$$\sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} = X(e^{j\omega}). \tag{2.20}$$

This is the Fourier transform of the discrete-time signal $x(n)$.

Later we will show that, under certain conditions, the Fourier transform $X(e^{j\omega})$ of discrete-time signals is not just an approximation of the Fourier transform $X(\Omega)$ of continuous-time signals, but the equality holds (that is, $X(\Omega) = X(e^{j\omega})$) with $\Omega\Delta t = \omega$ and $-\pi \leq \omega < \pi$.

The inverse Fourier transform of discrete-time signals is obtained by multiplying both sides of (2.20) by $e^{j\omega m}$ and integrating them over a period of $X(e^{j\omega})$

$$\sum_{n=-\infty}^{\infty} x(n) \int_{-\pi}^{\pi} e^{-j\omega(n-m)}d\omega = \int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega m}d\omega.$$

Since

$$\int_{-\pi}^{\pi} e^{-j\omega(n-m)}d\omega = 2\pi\delta(n-m),$$

we get the inverse Fourier transform relation

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega n}d\omega. \tag{2.21}$$

Example 2.7. Find the Fourier transform of the discrete-time signal

$$x(n) = Ae^{-\alpha|n|}$$

where $\alpha > 0$ is a real constant.

★The Fourier transform of this signal is

$$X(e^{j\omega}) = A + \sum_{n=-\infty}^{-1} Ae^{\alpha n - j\omega n} + \sum_{n=1}^{\infty} Ae^{-\alpha n - j\omega n}$$
$$= A\left(1 + \frac{e^{j\omega-\alpha}}{1-e^{j\omega-\alpha}} + \frac{e^{-j\omega-\alpha}}{1-e^{-j\omega-\alpha}}\right) = A\frac{1-e^{-2\alpha}}{1-2e^{-\alpha}\cos(\omega)+e^{-2\alpha}}. \tag{2.22}$$

Example 2.8. Find the inverse Fourier transform of a discrete-time signal if $X(e^{j\omega}) = 2\pi\delta(\omega)$ for $-\pi \le \omega < \pi$ and $X(e^{j\omega}) = 2\pi\sum_{k=-\infty}^{\infty}\delta(\omega + 2k\pi)$ for any ω.

★By definition

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} 2\pi\delta(\omega)e^{j\omega n}d\omega = 1.$$

Therefore, the Fourier transform of signal $x(n) = 1$ is

$$\sum_{n=-\infty}^{\infty} e^{-j\omega n} = 2\pi\sum_{k=-\infty}^{\infty}\delta(\omega + 2k\pi). \tag{2.23}$$

The equivalent form in the continuous-time domain is obtained (by using $\omega = \Omega T$ and $\delta(T\Omega) = \delta(\Omega)/T$) as

$$\sum_{n=-\infty}^{\infty} e^{j\Omega nT} = 2\pi\sum_{k=-\infty}^{\infty}\delta(\Omega T + 2k\pi) = \frac{2\pi}{T}\sum_{k=-\infty}^{\infty}\delta(\Omega + 2k\pi/T). \tag{2.24}$$

2.3.1 Properties

Linearity: The Fourier transform of discrete-time signals is linear

$$\mathrm{FT}\{ax(n)+by(n)\} = aX(e^{j\omega})+bY(e^{j\omega}), \tag{2.25}$$

where $X(e^{j\omega})$ and $Y(e^{j\omega})$ are the Fourier transforms of the discrete-time signals $x(n)$ and $y(n)$, respectively.

Shift and modulation: With respect to the signal shift and modulation the Fourier transform of discrete-time signals behaves in the same way as the Fourier transform of continuous-time signals,

$$\mathrm{FT}\{x(n-n_0)\} = X(e^{j\omega})e^{-jn_0\omega} \tag{2.26}$$

and

$$\mathrm{FT}\{x(n)e^{j\omega_0 n}\} = X(e^{j(\omega-\omega_0)}). \tag{2.27}$$

Example 2.9. The Fourier transform of a discrete-time signal $x(n)$ is $X(e^{j\omega})$. Find the Fourier transform of $y(n) = x(2n)$.

★For $y(n) = x(2n)$ the Fourier transform is

$$\begin{aligned}\mathrm{FT}\{x(2n)\} &= \sum_{n=-\infty}^{\infty} x(2n)e^{-j\omega n} = \sum_{n=-\infty}^{\infty} \frac{x(n)+(-1)^n x(n)}{2} e^{-j\omega n/2} \\ &= \frac{1}{2}\sum_{n=-\infty}^{\infty} [x(n)+e^{-jn\pi}x(n)]e^{-j\omega n/2} \\ &= \frac{1}{2}[X(e^{j\omega/2})+X(e^{j(\omega/2+\pi)})] = \frac{1}{2}[X(e^{j\omega/2})+X(e^{j(\omega+2\pi)/2})].\end{aligned} \tag{2.28}$$

The period of this Fourier transform is 2π. Period of $X(e^{j\omega/2})$ is 4π.

Example 2.10. Calculate the Fourier transform of the discrete-time signal (rectangular window),

$$w_R(n) = u(N+n)-u(n-N-1). \tag{2.29}$$

Write the Fourier transform of a Hann(ing) window

$$w_H(n) = \frac{1}{2}\left[1+\cos(n\pi/N)\right]\left[u(N+n)-u(n-N-1)\right].$$

★By definition

$$W_R(e^{j\omega}) = \sum_{n=-N}^{N} e^{-j\omega n} = e^{j\omega N}\frac{1-e^{-j\omega(2N+1)}}{1-e^{-j\omega}} = \frac{\sin(\omega\frac{2N+1}{2})}{\sin(\omega/2)}. \tag{2.30}$$

The Fourier transform of the Hann(ing) window can easily be written as

$$W_H(e^{j\omega}) = \frac{1}{2}\sum_{n=-N}^{N}\left(1+\frac{1}{2}e^{jn\pi/N}+\frac{1}{2}e^{-jn\pi/N}\right)e^{-j\omega n} = \tag{2.31}$$

$$= \frac{\sin(\omega\frac{2N+1}{2})}{2\sin(\omega/2)} + \frac{\sin((\omega-\frac{\pi}{N})\frac{2N+1}{2})}{4\sin((\omega-\frac{\pi}{N})/2)} + \frac{\sin((\omega+\frac{\pi}{N})\frac{2N+1}{2})}{4\sin((\omega+\frac{\pi}{N})/2)}. \tag{2.32}$$

As the window width increases in the time domain the main lobe width in the Fourier domain

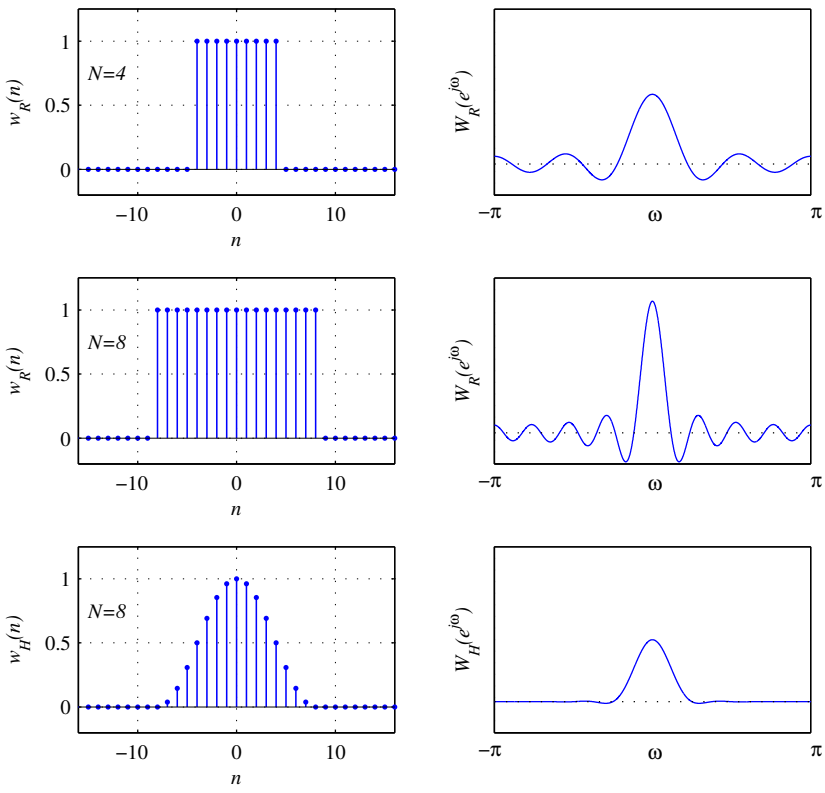


Figure 2.7 Discrete-time signal in a form of rectangular window of the widths $2N+1=9$ and $2N+1=17$ samples (top and middle), and a Hann(ing) window with $2N+1=17$ (bottom). The time domain values are on the left while the Fourier transforms of these discrete-time signals are on the right.

is narrowing. The first zero value of the Fourier transform of a rectangular window is at $\omega(2N+1)/2=\pi$, that is, at $\omega=2\pi/(2N+1)$ where $2N+1$ is the signal duration. In the case of a Hann(ing) window the main lobe is wider as compared to the rectangular window of the same width, but its convergence is much faster with very reduced oscillations in the Fourier transform, Fig. 2.7.

Convolution: The Fourier transform of a convolution of discrete-time signals,

$$\mathrm{FT}\{x(n) *_n h(n)\} = \sum_{n=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} x(k)h(n-k)e^{-jn\omega} = X(e^{j\omega})H(e^{j\omega}), \qquad (2.33)$$

is equal to the product of the Fourier transforms of corresponding discrete-time signals.

The Fourier transform of the impulse response

$$H(e^{j\omega}) = \sum_{n=-\infty}^{\infty} h(n)e^{-j\omega n}$$

is called *frequency response* of a discrete linear time-invariant system.

Example 2.11. Find the output of a discrete linear time-invariant system with frequency response $H(e^{j\omega})$ if the input signals are:
(a) $x(n) = Ae^{j\omega_0 n}$ and
(b) $x(n) = A\cos(\omega_0 n + \varphi)$. What is the output if the impulse response $h(n)$ is real-valued?

★(a) The output signal to the input $x(n)$ is

$$y(n) = \sum_{k=-\infty}^{\infty} h(k)x(n-k) = \sum_{k=-\infty}^{\infty} h(k)Ae^{j\omega_0(n-k)}$$
$$= Ae^{j\omega_0 n} \sum_{k=-\infty}^{\infty} h(k)e^{-j\omega_0 k} = Ae^{j\omega_0 n}H(e^{j\omega_0}) = A\left|H(e^{j\omega_0})\right| e^{j(\omega_0 n + \arg\{H(e^{j\omega_0})\})}.$$

(b) The input signal can be written as

$$x(n) = A\cos(\omega_0 n + \varphi) = \frac{A}{2}e^{j(\omega_0 n+\varphi)} + \frac{A}{2}e^{-j(\omega_0 n+\varphi)}.$$

According to the linearity property, the output is

$$y(n) = \frac{A}{2}\left|H(e^{j\omega_0})\right| e^{j(\omega_0 n+\varphi)+j\arg\{H(e^{j\omega_0})\}} + \frac{A}{2}\left|H(e^{-j\omega_0})\right| e^{-j(\omega_0 n+\varphi)+j\arg\{H(e^{-j\omega_0})\}}.$$

For a real-valued impulse response

$$H(e^{j\omega}) = \sum_{n=-\infty}^{\infty} h(n)\cos(\omega n) - j\sum_{n=-\infty}^{\infty} h(n)\sin(\omega n)$$

holds

$$H(e^{j\omega}) = H^*(e^{-j\omega})$$

with the even amplitude and odd phase of the transfer function

$$\left|H(e^{j\omega})\right| = \left|H(e^{-j\omega})\right|$$
$$\arg\{H(e^{j\omega}) = -\arctan\left\{\frac{\sum_{n=-\infty}^{\infty} h(n)\sin(\omega n)}{\sum_{n=-\infty}^{\infty} h(n)\cos(\omega n)}\right\} = -\arg\{H(e^{-j\omega})\}.$$

The output signal for a real-valued impulse response and $x(n) = A\cos(\omega_0 n + \varphi)$ is of the form

$$y(n) = A\left|H(e^{j\omega_0})\right|\cos(\omega_0 n + \varphi + \arg\{H(e^{j\omega_0})\}).$$

Example 2.12. Find the impulse response of an ideal discrete-time differentiator

$$H(e^{j\omega}) = j\omega \text{ for } -\pi \leq \omega < \pi.$$

★The impulse response is

$$h(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} j\omega e^{j\omega n} d\omega = \frac{j}{2\pi}\int_{-\pi}^{\pi} \omega\cos(\omega n)d\omega - \frac{1}{2\pi}\int_{-\pi}^{\pi} \omega\sin(\omega n)d\omega$$
$$= \frac{-1}{\pi}\int_{0}^{\pi} \omega\sin(\omega n)d\omega = \frac{1}{\pi}\omega\left.\frac{\cos(\omega n)}{n}\right|_0^{\pi} + \frac{1}{\pi n}\int_{0}^{\pi}\cos(\omega n)d\omega.$$

Since the last integral is equal to 0, the final result is

$$h(n) = \frac{\cos(\pi n)}{n} = \frac{(-1)^n}{n}$$

for $n \neq 0$ and $h(n) = 0$ for $n = 0$. Using samples $n = \pm 1, \pm 2, \ldots, \pm N$ the approximation of the frequency response is

$$H_N(e^{j\omega}) = \sum_{n=-N}^{N} h(n)e^{-j\omega n} = 2j\sum_{n=1}^{N}(-1)^{n-1}\frac{\sin(\omega n)}{n}.$$

Note that this system is not causal.

Product of signals: The Fourier transform of a product of the two discrete-time signals $x(n)$ and $h(n)$ is equal to the convolution of their Fourier transforms in the frequency domain,

$$\text{FT}\{x(n)h(n)\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\theta})H(e^{j(\omega-\theta)})d\theta = X(e^{j\omega}) *_{\omega} H(e^{j\omega}). \tag{2.34}$$

This convolution is periodic with period 2π (circular convolution).

2.3.2 Spectral Energy and Power Density

Parseval's theorem for discrete-time signals reads

$$\sum_{n=-\infty}^{\infty} x(n)y^*(n) = \sum_{n=-\infty}^{\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega}) e^{j\omega n} y^*(n) d\omega \tag{2.35}$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega}) \left(\sum_{n=-\infty}^{\infty} (e^{-j\omega n} y(n))^* \right) d\omega = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega}) Y^*(e^{j\omega}) d\omega.$$

For a signal $x(n)$, Parseval's theorem is

$$\sum_{n=-\infty}^{\infty} |x(n)|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| X(e^{j\omega}) \right|^2 d\omega = E_x.$$

Function $\left| X(e^{j\omega}) \right|^2$ is the spectral energy density of signal $x(n)$.

Since the average power of a signal $x(n)$ is defined by

$$P_{AV} = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} |x(n)|^2, \tag{2.36}$$

its power spectral density may be defined as

$$P_{xx}(e^{j\omega}) = \lim_{N\to\infty} \frac{1}{2N+1} \left| X_N(e^{j\omega}) \right|^2, \tag{2.37}$$

where the Fourier transform of $x(n)$ within $-N \le n \le N$ is denoted by $X_N(e^{j\omega})$. The power spectral density can be written as

$$P_{xx}(e^{j\omega}) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} \sum_{m=-N}^{N} x(n)x^*(m)e^{-j\omega(n-m)}. \tag{2.38}$$

For a very specific signal $x(n) = Ae^{j(\omega_0 n+\varphi)}$, which satisfies $r(k) = x(n)x^*(n-k) = A^2 e^{j\omega_0 k}$, the power spectral density is

$$P_{xx}(e^{j\omega}) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{k=-2N}^{2N} (2N+1-|k|) r(k) e^{-j\omega k},$$

since the value $r(0)$, for $n-m=k=0$, appears $2N+1$ times along the diagonal in n,m domain in (2.38). The value for $n-m=k=\pm 1$ appears $2N$ times, and so on. The value $r(k)$, for $n-m=k$, appears $2N+1-|k|$ times in double summation (2.38). Note that $P_{xx}(e^{j\omega})$, in this case, is the Fourier transform of $r(k)$ multiplied by the Bartlett window $1-|k|/(2N+1)$.

2.4 SAMPLING THEOREM IN THE TIME DOMAIN

Consider a continuous-time signal, $x(t)$, whose Fourier transform $X(\Omega)$ is nonzero within a limited frequency band $|\Omega| \leq \Omega_m$, that is

$$X(\Omega) = 0 \text{ for } |\Omega| > \Omega_m, \tag{2.39}$$

The signal $x(t)$ can be reconstructed at any t from the discrete-time samples, $x(n\Delta t)$, acquired at the instants $t = n\Delta t$ with the *sampling interval* Δt such that

$$\Delta t < \frac{\pi}{\Omega_m} = \frac{1}{2f_m},$$

where $\Omega_m = 2\pi f_m$. The discrete-time signal is commonly denoted by

$$x(n) = x(n\Delta t)\Delta t,$$

Now we will prove this fundamental statement in the analog to digital signal conversion. Since a limited duration of $X(\Omega)$ is assumed, we can make its periodic extension

$$X_p(\Omega) = \sum_{m=-\infty}^{\infty} X(\Omega + 2\Omega_0 m) \tag{2.40}$$

with the period in the frequency domain equal to $2\Omega_0$. For the reconstruction of the original aperiodic Fourier transform, $X(\Omega)$, from this periodic extension, $X_p(\Omega)$, it is of crucial importance that the basic period of $X_p(\Omega)$ contains undisturbed $X(\Omega)$, that is

$$X_p(\Omega) = X(\Omega) \quad \text{for} \quad |\Omega| < \Omega_0.$$

This condition is satisfied if the extension period $2\Omega_0$ and the maximum frequency in the Fourier transform of the signal, Ω_m, satisfy the inequality $\Omega_0 > \Omega_m$. In this case, it is possible to make transition from $X(\Omega)$ to $X_p(\Omega)$, and back, without losing any information. Of course, that would not be the case if $\Omega_0 > \Omega_m$ did not hold. By periodic extension of $X(\Omega)$, with $\Omega_0 \leq \Omega_m$ the overlapping (aliasing) would have occurred in $X_p(\Omega)$. It would not be reversible. A periodic extension of the Fourier transform with $\Omega_0 > \Omega_m$ is illustrated in Fig. 2.8.

The periodic function $X_p(\Omega)$ can be expanded into the Fourier series with coefficients

$$X_{-n} = \frac{1}{2\Omega_0} \int_{-\Omega_0}^{\Omega_0} X_p(\Omega) e^{j\pi\Omega n/\Omega_0} d\Omega = \frac{1}{2\Omega_0} \int_{-\infty}^{\infty} X(\Omega) e^{j\pi\Omega n/\Omega_0} d\Omega. \tag{2.41}$$

The integration limits are extended to the infinity since $X(\Omega) = X_p(\Omega)$ within the basic period interval and $X(\Omega) = 0$ outside this interval.

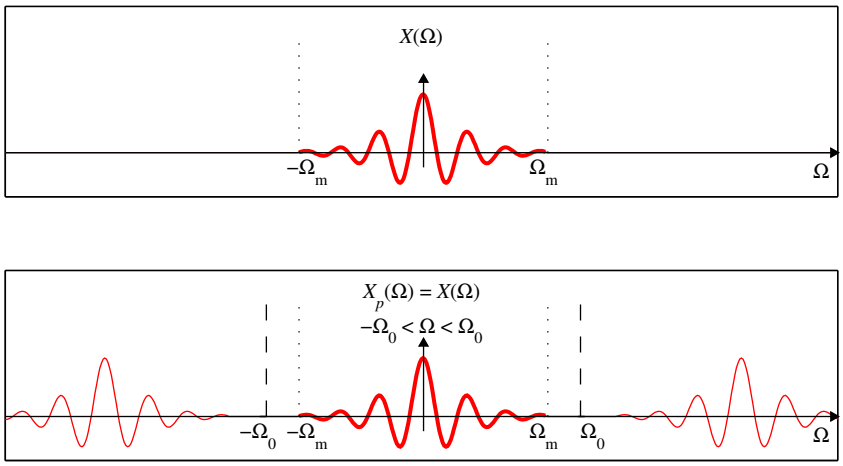


Figure 2.8 The Fourier transform of a signal, $X(\Omega)$, such that $X(\Omega) = 0$ for $|\Omega| > \Omega_m$ (top) and its periodically extended version, $X_p(\Omega)$, with the period $2\Omega_0 > 2\Omega_m$ (bottom).

The inverse Fourier transform of the continuous-time signal $x(t)$ is defined by

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\Omega) e^{j\Omega t} d\Omega. \tag{2.42}$$

By comparing equations (2.41) and (2.42), we easily conclude that

$$X_{-n} = \frac{\pi}{\Omega_0} x(t)_{|t=\pi n/\Omega_0} = x(n\Delta t)\Delta t \quad \text{with} \quad \Delta t = \frac{\pi}{\Omega_0},$$

meaning that the Fourier series coefficients of the periodically extended Fourier transform of $X(\Omega)$ are the samples of the signal $x(t)$, acquired at the instants $t = n\Delta t$, with the sampling interval $\Delta t = \pi/\Omega_0$. Therefore, the samples $x(n\Delta t)$ of the signal $x(t)$ and the periodically extended Fourier transform, $X_p(\Omega)$ are the Fourier series pair

$$x(n\Delta t)\Delta t = X_{-n} \quad \longleftrightarrow \quad X_p(\Omega) = \sum_{m=-\infty}^{\infty} X(\Omega + 2\Omega_0 m) \tag{2.43}$$

with $\Delta t = \pi/\Omega_0$.

The reconstruction formula for $x(t)$ form samples $x(n\Delta t)$ then follows from

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\Omega) e^{j\Omega t} d\Omega = \frac{1}{2\pi} \int_{-\Omega_0}^{\Omega_0} X_p(\Omega) e^{j\Omega t} d\Omega. \tag{2.44}$$

The periodic Fourier transform $X_p(\Omega)$ is now expanded into Fourier series to produce

$$x(t)=\frac{1}{2\pi}\int_{-\Omega_0}^{\Omega_0}\left(\sum_{n=-\infty}^{\infty}X_n e^{j\pi n\Omega/\Omega_0}\right)e^{j\Omega t}d\Omega. \tag{2.45}$$

With $X_n = x(-n\Delta t)\Delta t$ we get

$$x(t)=\frac{1}{2\pi}\int_{-\Omega_0}^{\Omega_0}\left(\sum_{n=-\infty}^{\infty}x(-n\Delta t)\Delta t e^{j\pi n\Omega/\Omega_0}\right)e^{j\Omega t}d\Omega. \tag{2.46}$$

Finally, the reconstruction formula

$$x(t)=\sum_{n=-\infty}^{\infty}x(n\Delta t)\frac{\sin(\frac{\pi}{\Delta t}(t-n\Delta t))}{\frac{\pi}{\Delta t}(t-n\Delta t)}, \tag{2.47}$$

follows by evaluating the simple integral over Ω. In this way, the signal $x(t)$ is expressed for any t, in terms of its samples $x(n\Delta t)$.

Example 2.13. The sampling theorem and relation (2.47) can be used to prove that $X(\Omega)=X(e^{j\omega})$ with $\Omega\Delta t=\omega$ for $|\omega|<\pi$ for the signals $x(t)$ sampled at the rate which satisfies the sampling theorem.

★Starting from the inverse Fourier transform definition

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

the signal $x(t)$, which satisfies the sampling theorem, can by written in terms of its samples, according to the third row of (2.46), as

$$x(t)=\frac{1}{2\pi}\int_{-\Omega_0}^{\Omega_0}\left(\sum_{n=-\infty}^{\infty}x(n\Delta t)\Delta t e^{-j\Delta t n\theta}\right)e^{j\theta t}d\theta.$$

The Fourier transform of such a signal $x(t)$ is given by

$$X(\Omega)=\int_{-\infty}^{\infty}\frac{1}{2\pi}\int_{-\Omega_0}^{\Omega_0}\left(\sum_{n=-\infty}^{\infty}x(n\Delta t)\Delta t e^{-j\Delta t n\theta}\right)e^{j\theta t}d\theta e^{-j\Omega t}dt$$

$$=\sum_{n=-\infty}^{\infty}x(n\Delta t)\Delta t\int_{-\Omega_0}^{\Omega_0}\delta(\theta-\Omega)e^{-j\Delta t n\theta}d\theta=\sum_{n=-\infty}^{\infty}x(n\Delta t)\Delta t e^{-j\Delta t n\Omega}\text{ for }|\Omega|<\Omega_0, \tag{2.48}$$

resulting in

$$X(\Omega) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} = X(e^{j\omega}) \quad \text{for} \quad |\omega| < \pi$$

with $\omega = \Omega\Delta t$ and $x(n) = x(n\Delta t)\Delta t$.

Example 2.14. If the highest frequency in a signal $x(t)$ is Ω_{m1} and the highest frequency in a signal $y(t)$ is Ω_{m2} what should be the sampling interval for the signals $x(t)y(t)$ and $x(t-t_1)y^*(t-t_2)$? The highest frequency Ω_m in a signal is used in the sense that the Fourier transform of the signal is zero for $|\Omega| > \Omega_m$.

★The Fourier transform of a product $x(t)y(t)$ is a convolution of the Fourier transforms $X(\Omega)$ and $Y(\Omega)$. Since these Fourier transforms are of limited duration within $|\Omega| < \Omega_{m1}$ and $|\Omega| < \Omega_{m2}$, respectively, in general, their convolution is limited to the frequency band $|\Omega| < \Omega_{m1} + \Omega_{m2}$. Therefore, the sampling interval for $y(t)$ should be

$$\Delta t < \frac{\pi}{\Omega_{m1} + \Omega_{m2}}.$$

Shifts in the time domain and the complex conjugate operation do not change the Fourier transform width. Therefore, the conclusion remains the same for the signal $x(t-t_1)y^*(t-t_2)$.

Example 2.15. If the signal

$$x(t) = e^{-|t|}$$

is sampled with $\Delta t = 0.1$, write the Fourier transform of the obtained discrete-time signal: (a) by a periodical extension of the continuous-time Fourier transform and (b) by a direct calculation based on the discrete-time signal. Comment on the expected error due to the discretization.

★The Fourier transform of this signal is

$$X(\Omega) = \int_{-\infty}^{\infty} x(t)e^{-j\Omega t}dt = \int_{-\infty}^{0} e^{t}e^{-j\Omega t}dt + \int_{0}^{\infty} e^{-t}e^{-j\Omega t}dt$$
$$= \frac{1}{1-j\Omega} + \frac{1}{1+j\Omega} = \frac{2}{1+\Omega^2}. \tag{2.49}$$

After sampling the signal in the time domain with the sampling interval $\Delta t = 0.1$, the Fourier transform is periodically extended with the period $2\Omega_0 = 2\pi/\Delta t = 20\pi$.

(a) The periodic Fourier transform is

$$X_p(\Omega) = \cdots + \frac{2}{1+(\Omega+20\pi)^2} + \frac{2}{1+\Omega^2} + \frac{2}{1+(\Omega-20\pi)^2} + \cdots$$

The value of $X_p(\Omega)$ at the period ending points $\pm 10\pi$ will approximately be equal to $X_p(\pm 10\pi) = 2/(1+100\pi^2) \cong 0.002$. By comparing this value with the maximum Fourier transform value $X(0) = 2$, we can conclude that the expected error due to the discretization of this signal (since it does not strictly satisfy the sampling theorem) will be of a 0.1% order.

(b) The discrete-time signal obtained by sampling $x(t) = \exp(-|t|)$ with $\Delta t = 0.1$ is $x(n) = 0.1e^{-0.1|n|}$. Its Fourier transform is already calculated with $A = 0.1$ and $\alpha = 0.1$ in equation (2.22). The result is

$$X(e^{j\omega}) = 0.1\frac{1-e^{-0.2}}{1-2e^{-0.1}\cos(\omega)+e^{-0.2}}. \tag{2.50}$$

Therefore, the exact value of an infinite sum in $X_p(\Omega)$ is $X(e^{j\omega})$ with $\omega = \Omega\Delta t = 0.1\Omega$

$$X_p(\Omega) = \sum_{k=-\infty}^{\infty} \frac{2}{1+(\Omega+20k\pi)^2} = 0.1\frac{1-e^{-0.2}}{1-2e^{-0.1}\cos(0.1\Omega)+e^{-0.2}}.$$

In this way, we have solved an interesting mathematical problem of finding a sum of an infinite series.

For $\Omega = 0$, the original value of the Fourier transform is $X(0) = 2$. In the signal that could be reconstructed based on the discretized signal $X_p(0) = 0.1(1+e^{-0.1})/(1-e^{-0.1}) = 2.00167$. The increase of 0.00167 is due to the periods overlapping. This overlapping manifests produces the aliasing error in $X(0)$. The value of error corresponds to our previous conclusion of about a 0.1% error order.

Example 2.16. A continuous-time signal

$$x(t) = \cos(25\pi t + \pi/4) + \sin(50\pi t - \pi/3)$$

is sampled with the sampling interval $\Delta t = 1/100$ and a discrete-time signal $x(n) = x(n\Delta t)\Delta t$ is formed. The discrete-time signal is processed using the system whose impulse response is

$$h(n) = \frac{1}{2}\delta(n) + \frac{1}{4}\delta(n-2) + \frac{1}{4}\delta(n+2)$$

Find the output signal $y(n)$ and the corresponding continuous-time signal $y_a(t)$.

★The discrete-time input signal is

$$x(n) = [\cos(n\pi/4 + \pi/4) + \sin(n\pi/2 - \pi/3)]\Delta t$$

with the Fourier transform

$$X(e^{j\omega}) = \frac{\pi}{100}\sum_{k=-\infty}^{\infty}[\delta(\omega+\frac{\pi}{4}+2k\pi)e^{-j\pi/4} + \delta(\omega-\frac{\pi}{4}+2k\pi)e^{j\pi/4}]$$
$$+ \frac{\pi}{100}\sum_{k=-\infty}^{\infty}[\delta(\omega+\frac{\pi}{2}+2k\pi)e^{-j\pi/6} + \delta(\omega-\frac{\pi}{2}+2k\pi)e^{j\pi/6}].$$

The frequency response of the discrete system is

$$H(e^{j\omega}) = \sum_{n=-\infty}^{\infty} h(n)e^{-j\omega n} = \frac{1}{2}(1+\cos(2\omega)).$$

The frequency response values at the frequencies of nonzero values in $X(e^{j\omega})$, within the basic period $-\pi \le \omega < \pi$, are $H(e^{\pm j\pi/4}) = 1/2$ and $H(e^{\pm j\pi/2}) = 0$. Therefore, the Fourier transforms of the output signal is

$$Y(e^{j\omega}) = H(e^{j\omega})X(e^{j\omega}) = \frac{\pi}{200}[\delta(\omega+\frac{\pi}{4})e^{-j\pi/4} + \delta(\omega-\frac{\pi}{4})e^{j\pi/4}] \quad \text{for} \quad -\pi \le \omega < \pi.$$

The output discrete-time signal is

$$y(n) = \frac{1}{2}\cos(n\pi/4+\pi/4)\Delta t.$$

The corresponding continuous-time output signal is given by

$$y(t) = \frac{1}{2}\cos(n\frac{\pi}{4\Delta t}+\pi/4) = \frac{1}{2}\cos(25\pi t+\pi/4).$$

Hint: Find the output signal for the same input and $h(n) = \sum_{i=-2}^{2}\delta(n-i)$.

2.5 PROBLEMS

Problem 2.1. Check the periodicity and find the period of signals:
(a) $x(n) = \sin(2\pi n/32)$,
(b) $x(n) = \cos(9\pi n/82)$,
(c) $x(n) = e^{jn/32}$, and
(d) $x(n) = \sin(\pi n/5) + \cos(5\pi n/6) - \sin(\pi n/4)$.

Problem 2.2. Check the linearity and time-invariance of the discrete system described by equation

$$y(n) = x(n) + 2.$$

Problem 2.3. The output of a linear time-invariant discrete system to the input signal $x(n) = u(n)$ is $y(n) = 2^{-n}u(n)$. Find the impulse response $h(n)$. Is the system stable?

Problem 2.4. Find the convolution

$$y(n) = x(n) *_n x(n)$$

for $x(n) = u(n) - u(n-5)$.

Problem 2.5. Find the convolution of discrete-time signals $x(n) = e^{-|n|}$ and $h(n) = u(n+5) - u(n-6)$.

Problem 2.6. A discrete system consists of systems with impulse responses $h_1(n) = e^{-an}u(n)$, $h_2(n) = e^{-bn}u(n)$, and $h_3(n) = u(n)$. Find the impulse response of the resulting system for:

(a) Systems $h_1(n)$, $h_2(n)$, and $h_3(n)$ connected in parallel,
(b) System $h_1(n)$ connected in parallel with a cascade of systems $h_2(n)$ and $h_3(n)$.

Problem 2.7. Consider three causal linear time-invariant systems in cascade. Impulse responses of these systems are $h_1(n)$, $h_2(n)$, and $h_2(n)$, respectively. The impulse response of the second and the third system is $h_2(n) = u(n) - u(n-2)$, while the impulse response of the whole system,

$$h(n) = h_1(n) *_n h_2(n) *_n h_2(n),$$

is shown in Fig. 2.9 (left).

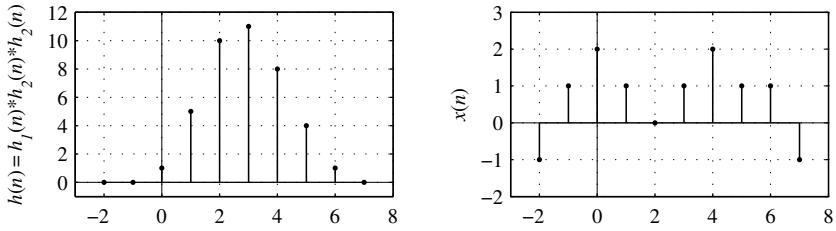


Figure 2.9 Problem 2.7, impulse response $h(n)$ (left) and Problem 2.14, discrete signal $x(n)$ (right).

Find $h_1(n)$ and $y(n) = h(n) *_n x(n)$, with $x(n) = \delta(n) - \delta(n-1)$.

Problem 2.8. Find the output of a discrete system whose impulse response is

$$h(n) = ne^{-n/2}u(n)$$

to the input signal

$$x(n) = 5\sin(\pi n/10) - 3\cos(\pi n/6 + \pi/6).$$

Find the sum

$$S = \sum_{n=0}^{\infty} ne^{-n/2}.$$

Problem 2.9. Find the Fourier transform of signals:
(a) $x(n) = u(n)$,
(b) $x(n) = 2\cos(\omega_0 n)u(n)$, and
(c) $y(n) = \sum_{k=-\infty}^{\infty} x(n+kN)$ if the Fourier transform of $x(n)$ is $X(e^{j\omega})$.

Problem 2.10. In order to design a system whose output will produce an approximation of the input signal derivative we may use a system with the impulse response

$$h(n) = a[\delta(n+1) - \delta(n-1)] + b[\delta(n+2) - \delta(n-2)].$$

Find the constants a and b such that

$$H(e^{j\omega}) \cong j\omega \text{ for small } \omega,$$

that is,

$$\left.\frac{dH(e^{j\omega})}{d\omega}\right|_{\omega=0} = j \quad \text{and} \quad \left.\frac{d^2H(e^{j\omega})}{d\omega^2}\right|_{\omega=0} = 0.$$

Problem 2.11. Find the Fourier transform of the following discrete-time signal (triangular window)

$$w_T(n) = \left(1 - \frac{|n|}{N+1}\right)[u(n+N) - u(n-N-1)].$$

with N being an even number.

Problem 2.12. Find the value of the definite integral

$$I = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\sin^2((N+1)\omega/2)}{\sin^2(\omega/2)} d\omega.$$

Problem 2.13. A window is formed as a sum of the three windows,

$$w(n) = w_H(n+N) + w_H(n) + w_H(n-N),$$

where $w_H(n)$ is the Hann(ing) window

$$w_H(n) = \frac{1}{2}[1 + \cos(n\pi/N)]\,[u(N+n) - u(n-N-1)].$$

Plot the window $w(n)$ and express its Fourier transform as a function of the Fourier transform of the Hann(ing) window $W_H(e^{j\omega})$. Generalize the results for

$$w(n) = \sum_{k=-K}^{K} w_H(n+kN).$$

Problem 2.14. A discrete-time signal $x(n)$ is shown in Fig. 2.9 (right). Without calculating its Fourier transform $X(e^{j\omega})$ find

$$X(e^{j0}), \quad X(e^{j\pi}), \quad \int_{-\pi}^{\pi} X(e^{j\omega})d\omega, \quad \int_{-\pi}^{\pi} \left|X(e^{j\omega})\right|^2 d\omega,$$

and a signal whose Fourier transform is the real part of $X(e^{j\omega})$, denoted by $\mathrm{Re}\{X(e^{j\omega})\}$.

Problem 2.15. Find the Fourier transform of the discrete-time signal

$$y(n) = ne^{-n/4}u(n).$$

Using this Fourier transform find the center of gravity of signal $x(n) = e^{-n/4}u(n)$, defined by

$$n_g = \frac{\sum_{n=-\infty}^{\infty} nx(n)}{\sum_{n=-\infty}^{\infty} x(n)}.$$

Problem 2.16. Impulse response of a discrete system is given by:

$$\text{(a)} \quad h(n) = \frac{\sin(n\pi/3)}{n\pi}, \text{ with } h(0) = 1/3,$$

$$\text{(b)} \quad h(n) = \frac{\sin^2(n\pi/3)}{(n\pi)^2}, \text{ with } h(0) = 1/9,$$

$$\text{(c)} \quad h(n) = \frac{\sin((n-2)\pi/4)}{(n-2)\pi}, \text{ with } h(2) = 1/4.$$

Show that the frequency response of the system with the impulse response $h(n) = \sin(n\pi/3)/n\pi$ is $H(e^{j\omega}) = 1$ for $|\omega| \leq \pi/3$ and $H(e^{j\omega}) = 0$ for $\pi/3 < |\omega| < \pi$. Find the frequency responses in other two cases (b) and (c). Find the output of these three systems to the input signal $x(n) = \sin(n\pi/6)$.

Problem 2.17. A continuous-time signal $x(t) = \cos(20\pi t + \pi/4) + \sin(90\pi t)$ is sampled with a step Δt and a discrete-time signal $x(n) = x(n\Delta t)\Delta t$ is formed. The signal is convolved with $h(n) = \sin(n\pi/2)/(n\pi)$. (a) What is the result of this convolution for $\Delta t = 1/100$? (b) If the signal is sampled with $\Delta t = 1/50$ what is the output signal? (c) Find the result of convolution for $\Delta t = 3/100$.

Problem 2.18. An analytic part $x_a(n)$ of a discrete-time signal $x(n)$ is defined in the frequency domain by

$$X_a(e^{j\omega}) = \begin{cases} 2X(e^{j\omega}) & \text{for} \quad 0 < \omega < \pi \\ X(e^{j\omega}) & \text{for} \quad \omega = 0 \\ 0 & \text{for} \quad -\pi \leq \omega < 0 \end{cases}.$$

In the time domain the analytic part can be written as

$$x_a(n) = x(n) + jx_h(n),$$

where $x_h(n)$ is the Hilbert transform of $x(n)$. Find the impulse response of the system that transforms a signal $x(n)$ into its Hilbert transform (Hilbert transformer).

Problem 2.19. The Fourier transform of a continuous signal $x(t)$ is nonzero only within $3\Omega_1 < \Omega < 5\Omega_1$. Find the maximum possible sampling interval Δt such that the signal can be reconstructed based on the samples $x(n\Delta t)$.

Problem 2.20. For a signal whose Fourier transform is zero-valued for frequencies $|\Omega| \geq \Omega_m = 2\pi f_m = \pi/\Delta t$ show that

$$x(t) = \int_{-\infty}^{\infty} x(\tau) \frac{\sin(\pi(t-\tau)/\Delta t)}{\pi(t-\tau)} d\tau.$$

Write a discrete-time version of this relation.

Problem 2.21. Sampling of a signal is done twice, with the sampling interval $\Delta t = 2\pi/\Omega_m$ that is twice larger than the sampling interval required by the sampling theorem ($\Delta t = \pi/\Omega_m$ is required). After the first sampling process, the discrete-time signal $x_1(n) = \Delta t x(n\Delta t)$ is formed, while after the second sampling process signal $x_2(n) = \Delta t x(n\Delta t + a)$ is formed. Show that the continuous-time signal $x(t)$ can be reconstructed based on $x_1(n)$ and $x_2(n)$ if $a \neq k\Delta t$, that is, if the samples $x_1(n)$ and $x_2(n)$ do not overlap in the continuous-time domain.

Problem 2.22. In general, a sinusoidal signal $x(t) = A\sin(\Omega_0 t + \varphi)$ is described with three parameters A, Ω_0 and φ. Thus, generally speaking, three points of $x(t)$ would be sufficient to find three signal parameters. If we know the signal $x(t)$ at $t = t_0$, $t = t_0 + \Delta t$ and $t = t_0 - \Delta t$ what is the relation and conditions to reconstruct, for example, Ω_0, which is usually the most important parameter of a sinusoid?

Problem 2.23. Show that the relation among the amplitudes of a signal $x(n)$ and its even and odd parts $x_e(n) = [x(n) + x(-n)]/2$ and $x_o(n) = [x(n) - x(-n)]/2$ is

$$A_s(n) \leq |x_e(n)| + |x_o(n)| \leq \sqrt{2} A_s(n)$$

where $A_s(n) > 0$ is defined by $A_s^2(n) = \left[|x(n)|^2 + |x(-n)|^2\right]/2$.

2.6 EXERCISE

Exercise 2.1. Calculate the convolution of the signals $x(n) = n[u(n) - u(n-3)]$ and $h(n) = \delta(n+1) + 2\delta(n) - \delta(n-2)$.

Exercise 2.2. Find the convolution of the signals $x(n) = e^{-|n|}$ and $h(n) = u(3-n)u(3+n)$.

Exercise 2.3. The output of a linear time-invariant discrete system to the input signal $x(n) = u(n)$ is $y(n) = (\frac{1}{3^n} + n)u(n)$. Find the impulse response $h(n)$. Is the system stable?

Exercise 2.4. For signal $x(n) = nu(5-n)u(n+5)$ find the values of $X(e^{j0})$, $X(e^{j\pi})$, $\int_{-\pi}^{\pi} X(e^{j\omega})d\omega$, and $\int_{-\pi}^{\pi} |X(e^{j\omega})|^2 d\omega$ without the Fourier transform calculation. Check the results by calculating the Fourier transform.

Exercise 2.5. For a signal $x(n)$ at an instant m a signal $y(n) = x(m-n)x^*(m+n)$ is formed. Show that the Fourier transform of $y(n)$ is real-valued. What is the Fourier

transform of $y(n)$ if $x(n) = A\exp(jan^2/4 + j2\omega_0 n)$? Find the Fourier transform of $z(m) = x(m-n)x^*(m+n)$ for a given n.

Note: The Fourier transform of $y(n)$ is the Fourier transform of $x(n)$ for a given m, while the Fourier transform of $z(m)$ is the Ambiguity function of $x(n)$ for a given n.

Exercise 2.6. For a signal $x(n)$ with the Fourier transform $X(e^{j\omega})$ find the Fourier transform of $x(2n)$. Find the Fourier transform of $y_1(2n) = x(2n)$ and $y_1(2n+1) = 0$. What is the Fourier transform of $x(2n+1)$ and what is the Fourier transform of the signal $y_2(n)$ defined by $y_2(2n) = 0$ and $y_2(2n+1) = x(2n+1)$. Check the result by showing that $Y_1(e^{j\omega}) + Y_2(e^{j\omega}) = X(e^{j\omega})$.

Exercise 2.7. For a real-valued signal $x(n)$ find the relation between its Fourier transform $X(e^{j\omega})$ and the corresponding Hartley transform

$$H(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)[\cos(\omega n) + \sin(\omega n)].$$

Write this relation if the signal is real-valued and even, that is, $x(n) = x(-n)$.

Exercise 2.8. Systems with impulse responses $h_1(n)$, $h_2(n)$ and $h_3(n)$ are connected in cascade. If the impulse responses $h_2(n) = h_3(n) = u(n) - u(n-2)$ and the resulting impulse response is $h(n) = \delta(n) + 5\delta(n-1) + 10\delta(n-2) + 11\delta(n-3) + 8\delta(n-4) + 4\delta(n-5) + \delta(n-6)$. Find the impulse response $h_1(n)$.

Exercise 2.9. Continuous-time signal $x(t) = \sin(100\pi t) + \cos(180\pi t) + \sin(200\pi t + \pi/4)$ is sampled with the sampling interval $\Delta t = 1/125$ and used as an input to the system with the transfer function $H(e^{j\omega}) = 1$ for $|\omega| < 3\pi/4$ and $H(e^{j\omega}) = 0$ for $|\omega| \geq 3\pi/4$. What is the discrete-time output of this system? What is the corresponding continuous-time output signal? What should be the sampling interval so that the continuous-time output signal $y(t)$ is equal to the input signal $x(t)$?

2.7 SOLUTIONS

Solution 2.1. (a) The signal shifted for N is given by $x(n+N) = \sin(2\pi(n+N)/32)$. The equality $x(n+N) = x(n)$ holds for $2\pi N/32 = 2k\pi$, $k = 1,2,\ldots$. The smallest integer N satisfying the previous condition is $N = 32$, with $k = 1$. The period of this signal is $N = 32$.

(b) For the signal $x(n) = \cos(9\pi n/82)$, the equality $x(n) = x(n+N) = \cos(9\pi n/82 + 9\pi N/82)$ holds for $9\pi N/82 = 2k\pi$, $k = 1,2,\ldots$. The period follows from $N = 164k/9$ and it is equal to $N = 164$ for $k = 9$.

(c) In this case $x(n+N) = e^{j(n/32+N/32)}$. The relation $N/32 = 2k\pi$, $k = 1,2,\ldots$, produces $N = 64k\pi$. This is not an integer for any k, meaning that the signal $x(n)$ is not periodic.

(d) The periods of the signal components are obtained from $N_1 = 10k$, $N_2 = 12k/5$, and $N_3 = 8k$. The smallest value of N when $N_1 = N_2 = N_3 = N$ is $N = 120$ containing 12 periods of $\sin(\pi n/5)$, 50 periods of $\cos(5\pi n/6)$, and 15 periods of $\sin(\pi n/4)$.

Solution 2.2. In order to establish if the linearity property holds we have to check the system output to a linear combination of the input signals $x_1(n)$ and $x_2(n)$,

$$\mathbb{T}\{a_1x_1(n) + a_2x_2(n)\} = a_1x_1(n) + a_2x_2(n) + 2.$$

This output is not equal to

$$a_1y_1(n) + a_2y_2(n) = a_1x_1(n) + 2a_1 + a_2x_2(n) + 2a_2.$$

Therefore, the system is not linear.

This system is time-invariant since

$$\mathbb{T}\{x(n-N)\} = x(n-N) + 2 = y(n-N).$$

Solution 2.3. The impulse response is defined by $h(n) = \mathbb{T}\{\delta(n)\}$, It can be written as

$$h(n) = \mathbb{T}\{u(n) - u(n-1)\}.$$

For a linear time-invariant discrete system holds

$$h(n) = \mathbb{T}\{u(n)\} - \mathbb{T}\{u(n-1)\}.$$

In this case, this relation means

$$\begin{aligned} h(n) &= \mathbb{T}\{x(n)\} - \mathbb{T}\{x(n-1)\} = y(n) - y(n-1) = 2^{-n}u(n) - 2^{-(n-1)}u(n-1) \\ &= \delta(n) + 2^{-n}u(n-1) - 2^{-(n-1)}u(n-1) = \delta(n) + 2^{-n}(1-2)u(n-1) \\ &= \delta(n) - 2^{-n}u(n-1). \end{aligned}$$

For this system

$$\sum_{n=-\infty}^{\infty}|h(n)| = 1 + \sum_{n=1}^{\infty} 2^{-n} = 1 + \frac{2^{-1}}{1-2^{-1}} = 2.$$

The system is stable since the sum of the impulse response absolute values is finite.

Solution 2.4. The convolution is calculated sample by sample as

$$\begin{aligned}
y(0) &= \sum_{k=-\infty}^{\infty} x(k)x(-k) = x(0)x(0) = 1\\
y(1) &= \sum_{k=-\infty}^{\infty} x(k)x(1-k) = x(0)x(1) + x(1)x(0) = 2\\
y(-1) &= \sum_{k=-\infty}^{\infty} x(k)x(-1-k) = 0\\
y(2) &= \sum_{k=-\infty}^{\infty} x(k)x(2-k) = 3\\
&\vdots
\end{aligned}$$

The calculation process is illustrated in Fig. 2.10, along with the final result $y(n)$.

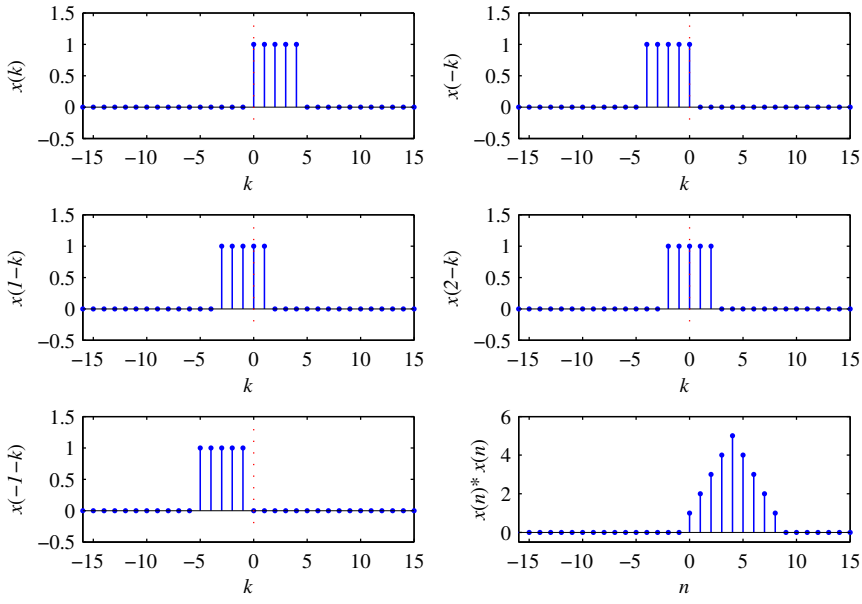


Figure 2.10 Illustration of the convolution calculation for a discrete-time signal $x(n) = u(n) - u(n-5)$.

Solution 2.5. Based on the convolution definition we an write

$$y(n) = x(n) *_n h(n) = \sum_{k=-\infty}^{\infty} x(k)h(n-k) = \tag{2.51}$$
$$= \sum_{k=-\infty}^{\infty} e^{-|k|}(u((n-k)+5) - u((n-k)-6))$$

with

$$u((n-k)+5) = \begin{cases} 1, \text{ for } k \le n+5 \\ 0, \text{ for } k > n+5 \end{cases} \quad \text{and} \quad u((n-k)-6) = \begin{cases} 1, \text{ for } k \le n-6 \\ 0, \text{ for } k > n-6 \end{cases}$$

we get

$$(u((n-k)+5) - u((n-k)-6)) = \begin{cases} 1, \text{ for } n-6 < k \le n+5 \\ 0, \text{ elsewhere.} \end{cases}$$

The infinite sum in (2.51) reduces to the terms for $n-5 \le k \le n+5$

$$y(n) = \sum_{k=n-5}^{n+5} e^{-|k|}.$$

Since $|k| = k$, for $k \ge 0$, and $|k| = -k$, for $k < 0$, we have three cases:
(1) For $n+5 \le 0$, that is, $n \le -5$, we have $k \le 0$ for all terms. Therefore $|k| = -k$, and

$$y(n) = \sum_{k=n-5}^{n+5} e^{k} = e^{n-5}\frac{1-e^{11}}{1-e} = e^{n}\frac{e^{-5}-e^{6}}{1-e} = e^{n}\frac{e^{0.5}}{e^{0.5}}\frac{e^{-5.5}-e^{5.5}}{e^{-0.5}-e^{0.5}} = e^{n}\frac{\sinh 5.5}{\sinh 0.5}.$$

(2) For $n-5 \ge 0$, the lowest $k = n-5$ is greater than 0. Then, $k \ge 0$ for all terms and

$$y(n) = \sum_{k=n-5}^{n+5} e^{-k} = e^{-n+5}\frac{1-e^{-11}}{1-e^{-1}} = e^{-n}\frac{e^{-0.5}}{e^{-0.5}}\frac{e^{5.5}-e^{-5.5}}{e^{0.5}-e^{-0.5}} = e^{-n}\frac{\sinh 5.5}{\sinh 0.5}.$$

(3) For $-5 < n < 5$, the index k can take positive and negative values. The convolution is split into two sums as

$$y(n) = \sum_{k=n-5}^{n+5} e^{-|k|} = \sum_{k=n-5}^{-1} e^{k} + \sum_{k=0}^{n+5} e^{-k} = \sum_{k=1}^{5-n} e^{-k} + \sum_{k=0}^{n+5} e^{-k}$$
$$= e^{-1}\frac{1-e^{-(5-n)}}{1-e^{-1}} + \frac{1-e^{-(n+6)}}{1-e^{-1}} = e^{-1/2}\frac{1-e^{n-5}}{e^{1/2}-e^{-1/2}} + e^{1/2}\frac{1-e^{-(n+6)}}{e^{1/2}-e^{-1/2}}$$
$$= \frac{\cosh(1/2) - e^{1/2}\cosh(n)}{\sinh(1/2)}.$$

Finally, we can write

$$y(n) = e^{-|n|}\frac{\sinh 5.5}{\sinh 0.5} \text{ for } |n| \geq 5 \quad \text{and} \quad y(n) = \frac{\cosh 0.5 - e^{-5.5}\cosh(n)}{\sinh 0.5} \text{ for } |n| < 5.$$

Solution 2.6. (a) For a parallel connection of systems the output signal is given by

$$\begin{aligned} y(n) &= y_1(n) + y_2(n) + y_3(n) \\ &= \sum_{k=-\infty}^{\infty} h_1(k)x(n-k) + \sum_{k=-\infty}^{\infty} h_2(k)x(n-k) + \sum_{k=-\infty}^{\infty} h_3(k)x(n-k) \\ &= \sum_{k=-\infty}^{\infty} [h_1(k) + h_2(k) + h_3(k)]x(n-k). \end{aligned}$$

The resulting impulse response of systems connected in parallel is

$$h(n) = h_1(k) + h_2(k) + h_3(k) = [e^{-an} + e^{-bn} + 1]u(n).$$

(b) For a cascade of systems with the impulse responses $h_2(n)$ and $h_3(n)$, the output from the first system is

$$y_2(n) = \sum_{k=-\infty}^{\infty} h_2(k)x(n-k) = \sum_{k=-\infty}^{\infty} h_2(n-k)x(k).$$

The input to the second system is equal to the output of the first system, while the output of the second system is

$$\begin{aligned} y_3(n) &= \sum_{m=-\infty}^{\infty} h_3(m)y_2(n-m) = \sum_{m=-\infty}^{\infty} h_3(m) \sum_{k=-\infty}^{\infty} h_2(n-m-k)x(k) \\ &= \sum_{k=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} h_3(m)h_2(n-m-k)x(k) = \sum_{k=-\infty}^{\infty} h_{23}(n-k)x(k) \end{aligned}$$

where

$$h_{23}(n) = \sum_{m=-\infty}^{\infty} h_3(m)h_2(n-m) = h_2(n) *_n h_3(n).$$

The impulse response of the whole system is

$$h(n) = h_1(n) + h_{23}(n) = h_1(n) + h_2(n) *_n h_3(n),$$

with

$$h_2(n) *_n h_3(n) = \sum_{m=-\infty}^{\infty} e^{-b(n-m)}u(n-m)u(m) = u(n)\sum_{m=0}^{n} e^{-b(n-m)} = \frac{e^{-bn} - e^{b}}{1 - e^{b}}u(n).$$

Solution 2.7. Since we know the impulse response $h_2(n)$, we can calculate

$$h_2(n) *_n h_2(n) = \delta(n) + 2\delta(n-1) + \delta(n-2).$$

Therefore, the total impulse response is obtained as

$$\begin{aligned} h(n) &= h_1(n) *_n [h_2(n) *_n h_2(n)] = h_1(n) + 2h_1(n-1) + h_1(n-2) \\ h_1(n) &= h(n) - 2h_1(n-1) - h_1(n-2). \end{aligned}$$

From the last relation follows: $h_1(n) = 0$ for $n < 0$, $h_1(0) = h(0) = 1$, $h_1(1) = h(1) - 2h_1(0) = 3$, $h_1(2) = h(2) - 2h_1(1) - h_1(0) = 3$, $h_1(3) = 2$, $h_1(4) = 1$, $h_1(5) = 0$, and $h_1(n) = 0$ for $n > 5$. The output signal for the input $x(n) = \delta(n) - \delta(n-1)$ can be easily calculated as

$$y(n) = h(n) - h(n-1).$$

Solution 2.8. Instead of a direct convolution we will calculate the frequency response, $H(e^{j\omega})$, of the discrete system. First, we will find the Fourier transform of the signal $e^{-n/2}u(n)$,

$$H_1(e^{j\omega}) = \sum_{n=0}^{\infty} e^{-n/2}e^{-j\omega n} = \frac{1}{1 - e^{-(1/2+j\omega)}}$$

and differentiate both sides with respect to ω

$$-j\sum_{n=0}^{\infty} ne^{-n/2}e^{-j\omega n} = \frac{-je^{-(1/2+j\omega)}}{(1 - e^{-(1/2+j\omega)})^2}.$$

The frequency response $H(e^{j\omega})$ is then obtained in the form

$$H(e^{j\omega}) = \sum_{n=0}^{\infty} ne^{-n/2}e^{-j\omega n} = \frac{e^{-(1/2+j\omega)}}{(1 - e^{-(1/2+j\omega)})^2}.$$

The output for a real-valued $h(n)$ is

$$\begin{aligned} y(n) =& 5\left|H(e^{j\pi/10})\right|\sin(\pi n/10 + \arg\{H(e^{j\pi/10}\}) \\ & -3\left|H(e^{j\pi/6})\right|\cos(\pi n/6 + \pi/6 + \left|H(e^{j\pi/6})\right|) \\ =& 14.1587\sin(\pi n/5 - 1.1481) - 5.7339\cos(\pi n/3 + \pi/6 - 1.6605). \end{aligned}$$

Value of the sum S is

$$S = \sum_{n=0}^{\infty} ne^{-n/2} = H(e^{j0}) = \frac{\sqrt{e}}{(\sqrt{e}-1)^2}.$$

Solution 2.9. (a) The unit step signal can be written as

$$x(n)=u(n)=\lim_{a\to 0}\left[\frac{1}{2}e^{-an}u(n)+\frac{1}{2}-\frac{1}{2}e^{-an}u(-n-1)\right]=\lim_{a\to 0}x_a(n).$$

The Fourier transform of $x_a(n)$ is

$$\begin{aligned}X_a(e^{j\omega})&=\sum_{n=-\infty}^{\infty}\left[\frac{1}{2}e^{-an}u(n)+\frac{1}{2}-\frac{1}{2}e^{-an}u(-n-1)\right]e^{-j\omega n}\\&=\frac{\frac{1}{2}}{1-e^{-a-j\omega}}+\sum_{k=-\infty}^{\infty}\pi\delta(\omega+2k\pi)-\frac{\frac{1}{2}e^{a+j\omega}}{1-e^{a+j\omega}}\end{aligned}$$

$$X(e^{j\omega})=\lim_{a\to 0}X_a(e^{j\omega})=\frac{1}{1-e^{-j\omega}}+\sum_{k=-\infty}^{\infty}\pi\delta(\omega+2k\pi).$$

The result from (2.23) is used to transform the constant signal equal to $1/2$.

(b) This signal can be written in the form

$$x(n)=2\cos(\omega_0 n)u(n)=(e^{j\omega_0 n}+e^{-j\omega_0 n})u(n).$$

Its Fourier transform is

$$\begin{aligned}X(e^{j\omega})&=\frac{1}{1-e^{-j(\omega-\omega_0)}}+\sum_{k=-\infty}^{\infty}\pi\delta(\omega-\omega_0+2k\pi)\\&+\frac{1}{1-e^{-j(\omega+\omega_0)}}+\sum_{k=-\infty}^{\infty}\pi\delta(\omega+\omega_0+2k\pi)\\&=2\frac{1-e^{-j\omega}\cos(\omega_0)}{1-2\cos(\omega_0)e^{-j\omega}+e^{-j2\omega}}\\&+\sum_{k=-\infty}^{\infty}\pi\left[\delta(\omega-\omega_0+2k\pi)+\delta(\omega+\omega_0+2k\pi)\right].\end{aligned}$$

(c) For a periodic signal $y(n)$ the Fourier transform is

$$\begin{aligned}Y(e^{j\omega})&=\sum_{k=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}x(n+kN)e^{-j\omega n}=\sum_{k=-\infty}^{\infty}X(e^{j\omega})e^{j\omega kN}\\&=X(e^{j\omega})\sum_{k=-\infty}^{\infty}e^{j\omega kN}.\end{aligned}$$

Using (2.23), we get

$$Y(e^{j\omega})=X(e^{j\omega})2\pi\sum_{k=-\infty}^{\infty}\delta(\omega N+2k\pi)=X(e^{j\omega})\frac{2\pi}{N}\sum_{k=-\infty}^{\infty}\delta(\omega+\frac{2k\pi}{N}).$$

Solution 2.10. For the impulse response $h(n)$ the frequency response is

$$H(e^{j\omega}) = 2aj\sin(\omega) + 2jb\sin(2\omega).$$

The first derivative of $H(e^{j\omega})$, at $\omega = 0$, is

$$\left.\frac{dH(e^{j\omega})}{d\omega}\right|_{\omega=0} = 2aj + 4jb = j,$$

while the second derivative, at $\omega = 0$, if of the form

$$\left.\frac{d^2H(e^{j\omega})}{d\omega^2}\right|_{\omega=0} = -2aj - 8jb = 0.$$

The constants a and b follow from the system

$$\begin{aligned} a + 2b &= 1/2 \\ a + 4b &= 0 \end{aligned}$$

as $b = -1/4$ and $a = 1$ with the resulting impulse response

$$h(n) = \delta(n+1) - \delta(n-1) - \frac{1}{4}(\delta(n+2) - \delta(n-2)).$$

Solution 2.11. Note that

$$w_T(n) = \frac{1}{N+1} w_R(n) *_n w_R(n)$$

where $w_R(n) = u(n + N/2) - u(n - N/2 - 1)$ is the rectangular window. Since

$$W_R(e^{j\omega}) = \frac{\sin(\omega\frac{N+1}{2})}{\sin(\omega/2)},$$

we have

$$W_T(e^{j\omega}) = \frac{1}{N+1} W_R(e^{j\omega}) W_R(e^{j\omega}) = \frac{1}{N+1} \frac{\sin^2(\omega\frac{N+1}{2})}{\sin^2(\omega/2)}.$$

Solution 2.12. This integral represents the energy of a discrete-time signal whose Fourier transform is defined by

$$X(e^{j\omega}) = \frac{\sin(\omega\frac{N+1}{2})}{\sin(\omega/2)}.$$

This signal is a rectangular window, $x(n) = u(n + N/2) - u(n - N/2 - 1)$. Its energy is

$$I = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\sin^2((N+1)\omega/2)}{\sin^2(\omega/2)} d\omega = \sum_{n=-N/2}^{N/2} x^2(n) = \sum_{n=-N/2}^{N/2} 1 = N + 1.$$

This integral is also equal to the value of $w_T(0)$ multiplied by $N + 1$.

Solution 2.13. The Hann(ing) window

$$w_H(n) = \frac{1}{2}\left[1 + \cos(n\pi/N)\right]\left[u(N + n) - u(n - N - 1)\right]$$

is nonzero within $-N \leq n \leq N - 1$. Thus, the windows $w_H(n)$ and $w_H(n - N)$ overlap within $0 \leq n \leq N - 1$. The new window, within this interval, is of the form

$$\begin{aligned} w(n) &= w_H(n) + w_H(n - N) \\ &= \frac{1}{2}\left[1 + \cos(n\pi/N)\right] + \frac{1}{2}\left[1 + \cos((n - N)\pi/N)\right] \\ &= 1 + \frac{1}{2}\cos(n\pi/N) + \frac{1}{2}\cos(n\pi/N - \pi) = 1. \end{aligned}$$

The same holds for $-N \leq n \leq -1$, when

$$w(n) = w_H(n + N) + w_H(n) = 1.$$

The resulting window is

$$w(n) = \begin{cases} 0 & \text{for} \quad n < -2N \\ \frac{1}{2}\left[1 - \cos(n\pi/N)\right] & \text{for} \quad -2N + 1 \leq n \leq -N + 1 \\ 1 & \text{for} \quad -N \leq n \leq N - 1 \\ \frac{1}{2}\left[1 - \cos(n\pi/N)\right] & \text{for} \quad N \leq n \leq 2N - 1 \\ 0 & \text{for} \quad n > 2N - 1, \end{cases}$$

since $\frac{1}{2}\left[1 + \cos((n \pm N)\pi/N) = \frac{1}{2}\left[1 - \cos(n\pi/N)\right]\right.$. The Fourier transform of the resulting window, in terms of the Fourier transform of the Hann(ing) window $W_H(e^{j\omega})$, is

$$\begin{aligned} W(e^{j\omega}) &= W_H(e^{j\omega})e^{-j\omega N} + W_H(e^{j\omega}) + W_H(e^{j\omega})e^{j\omega N} \\ &= W_H(e^{j\omega})[1 + 2\cos(\omega N)]. \end{aligned}$$

For a sum of $2K + 1$ windows

$$w(n) = \sum_{k=-K}^{K} w_H(n + kN)$$

we get

$$w(n)=\begin{cases} 0 & \text{for} \quad n<-(K+1)N \\ \frac{1}{2}\left[1+\cos((n+KN)\frac{\pi}{N})\right] & \text{for} \quad -(K+1)N+1\le n\le -KN+1 \\ 1 & \text{for} \quad -KN\le n\le KN-1 \\ \frac{1}{2}\left[1+\cos((n-KN)\frac{\pi}{N})\right] & \text{for} \quad KN\le n\le (K+1)N-1 \\ 0 & \text{for} \quad n>(K+1)N-1 \end{cases}$$

with

$$W(e^{j\omega})=W_H(e^{j\omega})\sum_{k=-K}^{K}e^{-j\omega kN}=W_H(e^{j\omega})\frac{\sin(\omega(2K+1)N/2)}{\sin(\omega N/2)}.$$

Similar results hold for the Hamming and the triangular window. These results can be generalized for shifts of $N/2$, $N/4$,...

For very large K the second term variations in $W(e^{j\omega})$ are much faster than the variations of $W_H(e^{j\omega})$. Thus, for large K the Fourier transform $W(e^{j\omega})$ approaches to the Fourier transform of a rectangular window of the width $(2K+1)N$.

Solution 2.14. Based on the definition of the Fourier transform of discrete-time signals,

$$X(e^{j0})=\sum_{n=-\infty}^{\infty}x(n)=7, \qquad X(e^{j\pi})=\sum_{n=-\infty}^{\infty}x(n)(-1)^n=1,$$

$$\int_{-\pi}^{\pi}X(e^{j\omega})d\omega=2\pi x(0)=4\pi, \qquad \int_{-\pi}^{\pi}\left|X(e^{j\omega})\right|^2d\omega=2\pi\sum_{n=-\infty}^{\infty}|x(n)|^2=30\pi.$$

Finally, $X(e^{j\omega})=\mathrm{Re}\{X(e^{j\omega})\}+j\,\mathrm{Im}\{X(e^{j\omega})\}$ and $X^*(e^{j\omega})=\mathrm{Re}\{X(e^{j\omega})\}-j\,\mathrm{Im}\{X(e^{j\omega})\}$. Thus,

$$\mathrm{Re}\{X(e^{j\omega})\}=\frac{1}{2}\left(X(e^{j\omega})+X^*(e^{j\omega})\right).$$

The inverse Fourier transform of $\mathrm{Re}\{X(e^{j\omega})\}$ is

$$y(n)=\frac{1}{2}(x(n)+x^*(-n)).$$

Solution 2.15. The Fourier transform of the signal $y(n)$ is

$$\begin{aligned} Y(e^{j\omega}) &= \sum_{n=-\infty}^{\infty}ne^{-n/4}u(n)e^{-j\omega n}=j\frac{d}{d\omega}\left[\sum_{n=0}^{\infty}e^{-n/4}e^{-j\omega n}\right] \\ &= j\frac{d}{d\omega}\frac{1}{1-e^{-1/4-j\omega}}=\frac{e^{-1/4-j\omega}}{(1-e^{-1/4-j\omega})^2}. \end{aligned}$$

The center of gravity of $x(n) = e^{-n/4}u(n)$ is

$$n_g = \frac{\sum_{n=-\infty}^{\infty} nx(n)}{\sum_{n=-\infty}^{\infty} x(n)} = \frac{Y(e^{j0})}{X(e^{j0})} = \frac{\frac{e^{-1/4-j\omega}}{(1-e^{-1/4-j\omega})^2}|_{\omega=0}}{\frac{1}{1-e^{-1/4-j\omega}}|_{\omega=0}} = \frac{1}{e^{1/4}-1} = 3.52.$$

Solution 2.16. (a) The inverse Fourier transform of

$$H(e^{j\omega}) = \begin{cases} 1 & \text{for} \quad |\omega| \leq \pi/3 \\ 0 & \text{for} \quad \pi/3 < |\omega| < \pi \end{cases}$$

is

$$h(n) = \frac{1}{2\pi}\int_{-\pi/3}^{\pi/3} e^{j\omega n} d\omega = \frac{e^{j\omega n}}{2j\pi n}\Big|_{-\pi/3}^{\pi/3} = \frac{\sin(\pi n/3)}{\pi n}.$$

The frequency response value at the input signal frequency $\omega = \pm\pi/6$ is $H(e^{\pm j\pi/6}) = 1$. The output signal is given by $y(n) = \sin(n\pi/6)$.
(b) The frequency response is $H(e^{j\omega}) *_\omega H(e^{j\omega})$, resulting in $y(n) = 0.25\sin(n\pi/6)$.
(c) The output signal is equal to $y(n) = \sin((n-2)\pi/6) = \sin(n\pi/6 - \pi/3)$.

Solution 2.17. For the signal $x(t) = \cos(20\pi t + \pi/4) + \sin(90\pi t)$, the corresponding discrete-time signal is

$$x(n) = \cos(20\pi n\Delta t + \pi/4)\Delta t + \sin(90\pi n\Delta t)\Delta t.$$

(a) For $\Delta t = 1/100$

$$x(n) = \cos(0.2\pi n + \pi/4)/100 + \sin(0.9\pi n)/100,$$

with the Fourier transform

$$X(e^{j\omega}) = \frac{\pi}{100}\sum_{k=-\infty}^{\infty}\left[\delta(\omega - 0.2\pi + 2k\pi)e^{j\pi/4} + \delta(\omega + 0.2\pi + 2k\pi)e^{-j\pi/4}\right]$$
$$+ \frac{\pi}{j100}\sum_{k=-\infty}^{\infty}\Big[\delta(\omega - 0.9\pi + 2k\pi) - \delta(\omega + 0.9\pi + 2k\pi)\Big].$$

Since the Fourier transform of $h(n) = \sin(n\pi/2)/(n\pi)$ is $H(e^{j\omega}) = 1$ for $|\omega| \leq \pi/2$ and $H(e^{j\omega}) = 0$ for $\pi/2 < |\omega| < \pi$, the result of the convolution is equal to the output of system with the transfer function $H(e^{j\omega})$ to the input signal $x(n)$. In this case

$$x(n) = \cos(0.2\pi n + \pi/4)/100.$$

A continuous-time signal corresponding to the output discrete-time signal is given by $y(t) = \cos(20\pi t + \pi/4)$, as shown in Fig. 2.11(top).

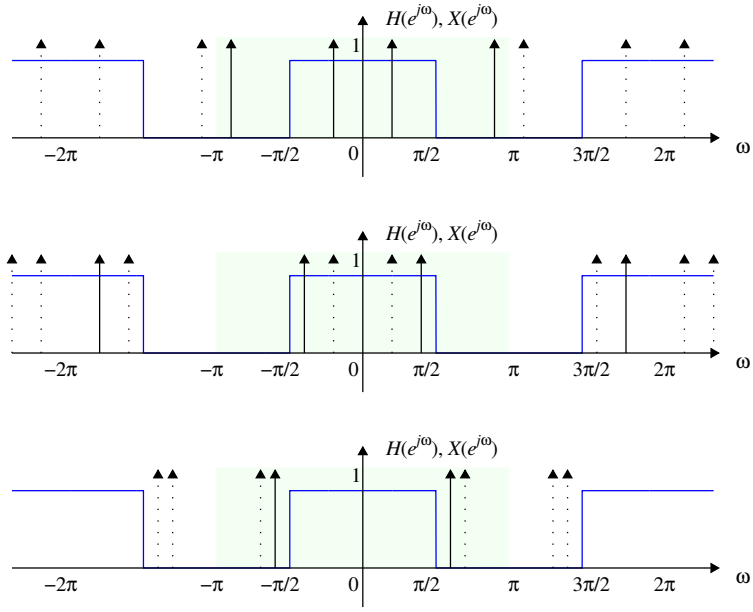


Figure 2.11 Illustration of the system output with various sampling intervals (a)-(c).

(b) If the signal is sampled with the sampling interval $\Delta t = 1/50$ the discrete-time signal is

$$x(n) = \cos(0.4\pi n + \pi/4)/50 + \sin(1.8\pi n)/50,$$

with the Fourier transform

$$X(e^{j\omega}) = \frac{\pi}{50}\sum_{k=-\infty}^{\infty}[\delta(\omega - 0.4\pi + 2k\pi)e^{j\pi/4} + \delta(\omega + 0.4\pi + 2k\pi)e^{-j\pi/4}]$$
$$+ \frac{\pi}{j50}\sum_{k=-\infty}^{\infty}[\delta(\omega - 1.8\pi + 2k\pi) - \delta(\omega + 1.8\pi + 2k\pi)].$$

The Fourier transform components within the basic period, $-\pi \le \omega < \pi$, are

$$X(e^{j\omega}) = \frac{\pi}{50}[\delta(\omega - 0.4\pi)e^{j\pi/4} + \delta(\omega + 0.4\pi)e^{-j\pi/4}]$$
$$+\frac{\pi}{j50}[\delta(\omega - 1.8\pi + 2\pi) - \delta(\omega + 1.8\pi - 2\pi)]$$
$$= \frac{\pi}{50}[\delta(\omega - 0.4\pi)e^{j\pi/4} + \delta(\omega + 0.4\pi)e^{-j\pi/4}] + \frac{\pi}{j50}[\delta(\omega + 0.2\pi) - \delta(\omega - 0.2\pi)].$$

The result of convolution is

$$x(n) = \cos(0.4\pi n + \pi/4)/50 - \sin(0.2\pi n)/50,$$

with the corresponding continuous-time signal

$$x(t) = \cos(20\pi t + \pi/4) - \sin(10\pi t).$$

The component $-\sin(10\pi t)$ does not correspond to any frequency in the input signal, Fig. 2.11(middle). This effect is illustrated in Fig. 2.12.

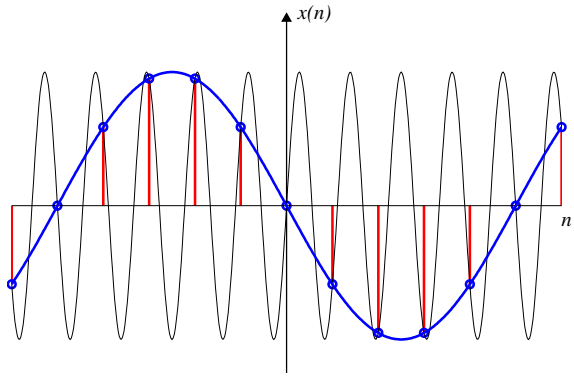


Figure 2.12 Illustration of the aliasing caused frequency change, from signal $\sin(90\pi t)$ to signal $-\sin(10\pi t)$.

(c) For the sampling interval $\Delta t = 3/100$ the discrete-time signal is of the form

$$x(n) = 3\cos(0.6\pi n + \pi/4)/100 + 3\sin(2.7\pi n)/100.$$

The Fourier transform components within the basic period, $-\pi \leq \omega < \pi$, are given by

$$\begin{aligned} X(e^{j\omega}) = {} & \frac{3\pi}{100}[\delta(\omega - 0.6\pi)e^{j\pi/4} + \delta(\omega + 0.6\pi)e^{-j\pi/4}] \\ & + \frac{3\pi}{j100}[\delta(\omega - 2.7\pi + 2\pi) - \delta(\omega + 2.7\pi - 2\pi)]. \end{aligned}$$

The result of the convolution is $y(n) = 0$, Fig. 2.11(bottom).

Solution 2.18. The Fourier transform of an analytic part of a signal is

$$\begin{aligned} X_a(e^{j\omega}) &= \begin{cases} 2X(e^{j\omega}) & \text{for} \quad 0 < \omega < \pi \\ X(e^{j\omega}) & \text{for} \quad \omega = 0 \\ 0 & \text{for} \quad -\pi \leq \omega < 0 \end{cases} \\ &= X(e^{j\omega}) + \text{sign}(\omega)(X(e^{j\omega}) = X(e^{j\omega}) + X_h(e^{j\omega}). \end{aligned}$$

The frequency response of the discrete Hilbert transformer is

$$H(e^{j\omega}) = \begin{cases} 1 & \text{for} \quad 0 < \omega < \pi \\ 0 & \text{for} \quad \omega = 0 \\ -1 & \text{for} \quad -\pi \leq \omega < 0 \end{cases} = \text{sign}(\omega)$$

for $-\pi \leq \omega < \pi$. The impulse response is

$$h(n) = \int_{-\pi}^{\pi} \text{sign}(\omega) e^{j\omega n} d\omega = \frac{2\sin^2(\pi n/2)}{\pi n}.$$

For $n = 0$ the impulse response is $h(0) = 0$. The discrete-time Hilbert transformer, in the frequency and the time domain, is shown in Fig. 2.13.

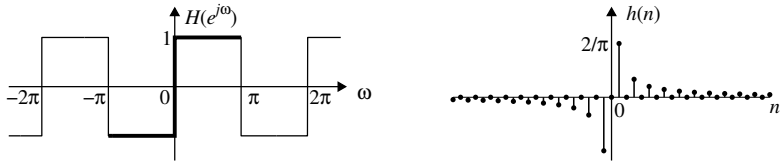


Figure 2.13 Frequency and impulse response of the discrete-time Hilbert transformer.

Solution 2.19. From the definition and conditions for the sampling theorem we could conclude that the maximum sampling interval should be related to the maximum frequency $5\Omega_1$ as $\Delta t = \pi/(5\Omega_1)$, corresponding to the periodical extension of the Fourier transform $X(\Omega)$ with period $10\Omega_1$. However, in this case, there is no need to use such a large period in order to achieve that two periods do not overlap. It is sufficient to use the period equal to $2\Omega_1$, as shown in Fig. 2.14. In this case, we will be able to reconstruct the signal, with some additional processing.

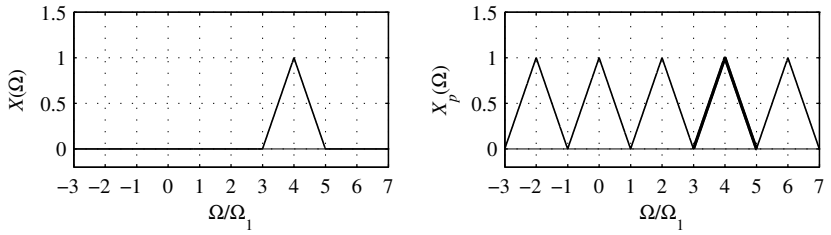


Figure 2.14 Problem 2.19: illustration of the Fourier transform periodic extension.

It is obvious that, after the signal sampling with $\Delta t = \pi/\Omega_1$ (periodic extension of Fourier transform with $2\Omega_1$), the basic period $-\Omega_1 < \Omega < \Omega_1$ will contain the original

Fourier transform shifted for $4\Omega_1$. The reconstructed signal is

$$x(t) = e^{j4\Omega_1 t} \sum_{n=-\infty}^{\infty} x(n\Delta t) \frac{\sin(\pi(t-n\Delta t)/\Delta t)}{\pi(t-n\Delta t)/\Delta t} \text{ with } \Delta t = \pi/\Omega_1.$$

Solution 2.20. For signal whose Fourier transform is zero for frequencies $|\Omega| \geq \Omega_m = 2\pi f_m = \pi/\Delta t$ hods

$$X(\Omega) = X(\Omega)H(\Omega)$$

where

$$H(\Omega) = \begin{cases} 1 & \text{for} \quad |\Omega| < \Omega_m = \pi/\Delta t \\ 0 & \text{for} \quad |\Omega| \geq \Omega_m = \pi/\Delta t \end{cases}.$$

The impulse response of $H(\Omega)$ is equal to

$$h(t) = \frac{1}{2\pi} \int_{-\pi/\Delta t}^{\pi/\Delta t} e^{j\Omega t} d\Omega = \frac{\sin(\pi t/\Delta t)}{\pi t}.$$

Then $x(t) = x(t) * h(t)$ produces the signal

$$x(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau = \int_{-\infty}^{\infty} x(\tau) \frac{\sin(\pi(t-\tau)/\Delta t)}{\pi(t-\tau)} d\tau.$$

In order to write this relation in the discrete-time form note that

$$X(\Omega) = X_p(\Omega)H(\Omega) \tag{2.52}$$

holds if the Fourier transform of signal $X(\Omega)$ is periodically extended with the period $2\pi/\Delta t \geq 2\Omega_m$, to produce

$$X(\Omega) *_\Omega \sum_{k=-\infty}^{\infty} 2\pi\delta\left(\Omega - \frac{2\pi}{\Delta t}k\right) = X_p(\Omega).$$

Convolution in the frequency domain of two Fourier transforms corresponds to the product of signals in the time domain, that is

$$x(t) \sum_{n=-\infty}^{\infty} \delta(t+n\Delta t)\Delta t = \text{IFT}\{X_p(\Omega)\} = x_p(t). \tag{2.53}$$

Relation (1.67), that reads

$$\frac{2\pi}{\Delta t} \sum_{k=-\infty}^{\infty} \delta\left(\Omega - \frac{2\pi}{\Delta t}k\right) = \text{FT}\left\{\sum_{n=-\infty}^{\infty} \delta(t+n\Delta t)\right\} = \text{FT}\left\{\sum_{n=-\infty}^{\infty} \delta(t-n\Delta t)\right\}$$

is used.

From (2.52) and then (2.53) the discrete-time form follows

$$x(t) = x_p(t) *_t h(t) = \int_{-\infty}^{\infty} x(\tau) \sum_{n=-\infty}^{\infty} \delta(\tau - n\Delta t) h(t-\tau)\Delta t d\tau$$

$$= \sum_{n=-\infty}^{\infty} x(n\Delta t) h(t - n\Delta t)\Delta t = \sum_{n=-\infty}^{\infty} x(n\Delta t) \frac{\sin(\frac{\pi}{\Delta t}(t - n\Delta t))}{\frac{\pi}{\Delta t}(t - n\Delta t)}. \tag{2.54}$$

The convergence of function $\sin(t)/t$ is very slow.

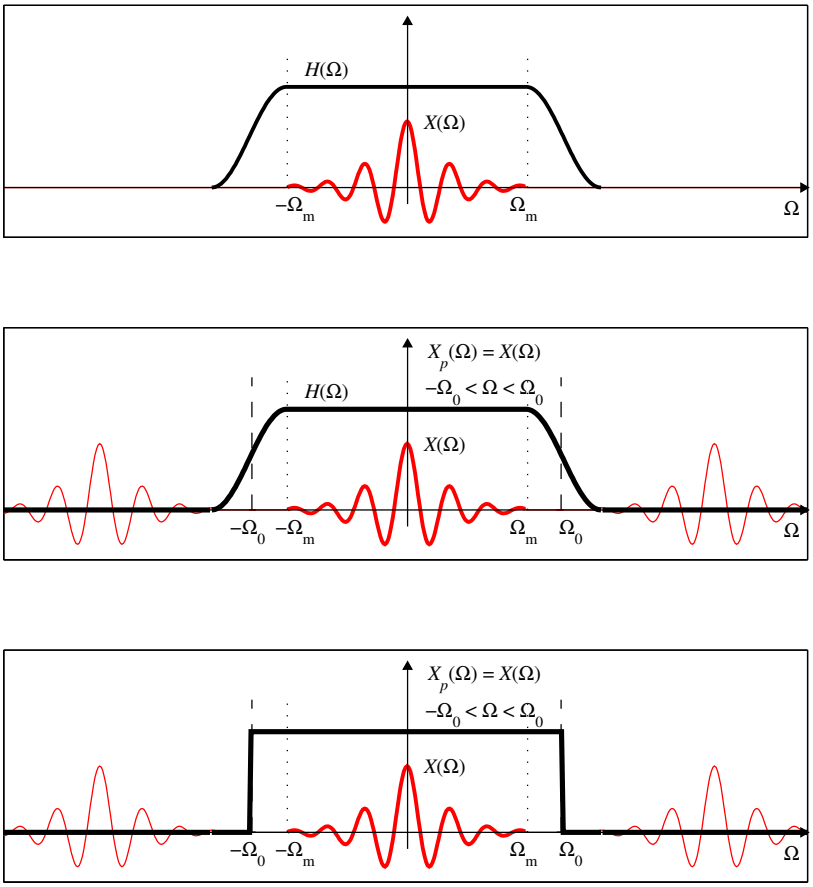


Figure 2.15 Smoothed filter in the sampling theorem illustration (first two graphs) versus original sampling theorem relation within filtering framework.

The previous derivation provides a possibility that a smooth transition in $H(\Omega)$ is used for $\Omega_m \leq |\Omega| \leq \Omega_m + \Delta\Omega_m$. This region of smooth changes from $H(\Omega) = 1$, for $|\Omega| < \Omega_m$, to $H(\Omega) = 0$, for $|\Omega| \geq \Omega_m + \Delta\Omega_m$, improves the convergence of $h(t)$, as

illustrated in Fig. 2.15. The sampling step should be $(\Omega_m + \frac{\Delta\Omega_m}{2}) = \pi/\Delta t$ so that the periodic extension of $X(\Omega)H(\Omega)$ does not include overlapped $X(\Omega)$ values. The impulse response $h(t)$ can be then used in the reconstruction formula

$$x(t) = \sum_{n=-\infty}^{\infty} x(n\Delta t)h(t - n\Delta t),$$

with a reduction of the sampling interval to $\Delta t = \pi/(\Omega_m + \frac{\Delta\Omega_m}{2})$ with respect to $\Delta t = \pi/\Omega_m$.

Solution 2.21. The Fourier transforms of discrete-time signals, in a continuous frequency notation, are periodically extended versions of $X(\Omega)$ with the period $2\pi/\Delta t$,

$$X_1(\Omega) = \sum_{n--\infty}^{\infty} X(\Omega + 2\pi n/\Delta t),$$
$$X_2(\Omega) = \sum_{n--\infty}^{\infty} X(\Omega + 2\pi n/\Delta t)e^{j(\Omega+2\pi n/\Delta t)a}.$$

Within the basic period (considering the positive frequencies $0 \leq \Omega < \Omega_m$), only two periods overlap

$$X_1(\Omega) = X(\Omega) + X(\Omega - 2\pi/\Delta t),$$
$$X_2(\Omega) = X(\Omega)e^{j\Omega a} + X(\Omega - 2\pi/\Delta t)e^{j(\Omega-2\pi/\Delta t)a}.$$

The second term $X(\Omega - 2\pi/\Delta t)$ in these relations is the overlapped period (aliasing) of the Fourier transform, that should be eliminated using these two equations. The Fourier transform $X(\Omega)$ of the original signal follows in the form

$$X(\Omega) = \frac{X_1(\Omega)e^{-j2\pi a/\Delta t} - X_2(\Omega)e^{-j\Omega a}}{e^{-j2\pi a/\Delta t} - 1} \text{ for } a \neq k\Delta t.$$

Similarly for negative frequencies, within the basic period $-\Omega_m < \Omega < 0$, follows

$$X(\Omega) = \frac{X_1(\Omega)e^{j2\pi a/\Delta t} - X_2(\Omega)e^{-j\Omega a}}{e^{j2\pi a/\Delta t} - 1} \text{ for } a \neq k\Delta t.$$

Therefore, the signal can be reconstructed from two independent discrete-time signals undersampled with factor of two.

A similar result could be derived for N independently sampled, N times undersampled signals.

Solution 2.22. It is easy to show that

$$\frac{x(t_0 + \Delta t) + x(t_0 - \Delta t)}{2x(t_0)} = \frac{A\sin(\Omega_0 t_0 + \varphi + \Omega_0\Delta t) + A\sin(\Omega_0 t_0 + \varphi - \Omega_0\Delta t)}{2A\sin(\Omega_0 t_0 + \varphi)}$$

or

$$\frac{x(t_0+\Delta t)+x(t_0-\Delta t)}{2x(t_0)}=\frac{2\sin(\Omega_0 t_0+\varphi)\cos(\Omega_0\Delta t)}{2\sin(\Omega_0 t_0+\varphi)}=\cos(\Omega_0\Delta t),$$

with

$$\Omega_0=\frac{1}{\Delta t}\arccos\left(\frac{x(t_0+\Delta t)+x(t_0-\Delta t)}{2x(t_0)}\right).$$

The condition for a unique solution is that the argument of cosine is $0\leq\Omega_0\Delta t\leq\pi$, limiting the approach to small values of Δt.

We will discuss the discrete complex-valued signal as well. For a complex sinusoid $x(n)=A\exp(j2\pi k_0 n/N+\phi_0)$, with available two samples $x(n_1)=A\exp(j\varphi(n_1))$ and $x(n_2)=A\exp(j\varphi(n_2))$, from

$$\frac{x(n_1)}{x(n_2)}=\exp(j2\pi k_0(n_1-n_2)/N)$$

follows

$$2\pi k_0(n_1-n_2)/N=\varphi(n_1)-\varphi(n_2)+2k\pi,$$

where k is an arbitrary integer. Then

$$k_0=\frac{\varphi(n_1)-\varphi(n_2)}{2\pi(n_1-n_2)}N+\frac{k}{n_1-n_2}N. \tag{2.55}$$

Let us analyze the ambiguous term $kN/(n_1-n_2)$ role in the determination of k_0. For $n_1-n_2=1$, this term is kN, meaning that any frequency k_0 would be ambiguous with kN. Any value k_0+kN for $k\neq 0$, in this case, will be outside the basic period $0\leq k\leq N-1$. Thus, we may find k_0 in a unique way, within $0\leq k_0\leq N-1$. However, for $n_1-n_2=L>1$, the terms $kN/(n_1-n_2)=kN/L$ produce shifts within the frequency basic period. Then several possible solutions for the frequency k_0 are obtained. For example, for $N=16$ and $k_0=5$ if we use $n_1=1$ and $n_2=5$, a possible solution to (2.55) is $k_0=5$, but also

$$k_0=5+16k/4,$$

or $k_0=9$, $k_0=13$, and $k_0=1$, for k_0 within $0\leq k_0\leq 15$, are possible solutions for frequency of the considered discrete-time signal.

Solution 2.23. For the absolute values of an even and odd part of the signal holds

$$|x_e(n)|^2+|x_o(n)|^2=\left|\frac{x(n)+x(-n)}{2}\right|^2+\left|\frac{x(n)-x(-n)}{2}\right|^2.$$

From this relation we can write

$$|x_e(n)|^2+|x_o(n)|^2=\frac{|x(n)|^2+|x(-n)|^2}{2}=A_s^2(n).$$

Obviously, $|x_e(n)|^2 \leq A_s^2(n)$ and $|x_o(n)|^2 \leq A_s^2(n)$. Replacing $|x_o(n)| = \sqrt{A_s^2(n) - |x_e(n)|^2}$ into $|x_e(n)| + |x_o(n)|$ we get

$$|x_e(n)| + |x_o(n)| = |x_e(n)| + \sqrt{A_s^2(n) - |x_e(n)|^2}.$$

Now, we have to check the function

$$f(\chi) = \chi + \sqrt{A_s^2(n) - \chi^2}$$

for variable χ within $0 \leq \chi \leq |A_s(n)|$. Variable χ stands for $|x_e(n)|$. By differentiating $f(\chi)$ with respect to χ we get $df(\chi)/d\chi = 1 - \chi/\sqrt{A_s^2(n) - \chi^2}$. The stationary point is obtained from $df(\chi)/d\chi = 0$ for $\chi = A_s(n)/\sqrt{2}$. The derivative $df(\chi)/d\chi$ is positive for $\chi < A_s(n)/\sqrt{2}$ and negative for $\chi > A_s(n)/\sqrt{2}$. Thus, the stationary point is the position of the function maximum. The maximum function value is $\sqrt{2}A_s(n)$ since

$$|x_e(n)| + |x_o(n)| \leq \frac{A_s(n)}{\sqrt{2}} + \sqrt{A_s^2(n) - \frac{A_s^2(n)}{2}} = \sqrt{2}A_s(n).$$

The minimum value is achieved at the interval ending points for $\chi = 0$ or $\chi = A_s(n)$, producing the final result

$$A_s(n) \leq |x_e(n)| + |x_o(n)| \leq \sqrt{2}A_s(n).$$

Chapter 3

Discrete Fourier Transform

DISCRETE-TIME signals can be processed on digital computers in the time domain. Their Fourier transform is a function of continuous frequency. For numeric processing of discrete-time signals in the frequency domain their Fourier transform should be discretized as well. Discretization in the frequency domain will enable numeric processing of discrete-time signals in both time and frequency domain.

3.1 DFT DEFINITION

The discrete Fourier transform (DFT) is defined by

$$\text{DFT}\{x(n)\} = X(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi kn/N} \tag{3.1}$$

for $k = 0,1,2,\ldots,N-1$.

In order to establish the relation between the DFT with the Fourier transform of discrete-time signals, consider a discrete-time signal $x(n)$ of limited duration. Assume that nonzero samples of $x(n)$ are within $0 \leq n \leq N_0 - 1$. The Fourier transform of this discrete-time signal is given by

$$X(e^{j\omega}) = \sum_{n=0}^{N_0-1} x(n)e^{-j\omega n}.$$

The DFT values can be considered as the frequency domain samples of the Fourier transform of discrete-time signals, taken with the sampling interval $\Delta\omega = 2\pi/N$ in the frequency domain, where N is the number of samples of $X(e^{j\omega})$ within the period $-\pi \leq \omega < \pi$,

$$X(k) = X(e^{j2\pi k/N}) = X(e^{j\omega})\Big|_{\omega=k\Delta\omega=2\pi k/N}. \tag{3.2}$$

In order to examine how the Fourier transform sampling in the frequency domain influences the signal in the time domain, we will form a periodic extension of the discrete-time signal $x(n)$ with the period N equal to the number of the samples within the basic frequency domain period, such that $N \geq N_0$, Fig. 3.1.

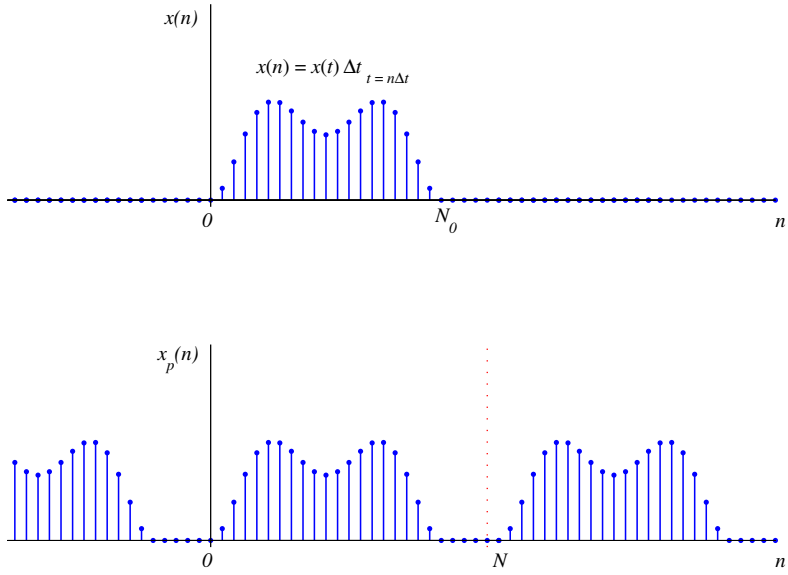


Figure 3.1 Periodic extension of a discrete-time signal.

With N being greater or equal to the signal duration N_0, we will be able to reconstruct the original signal $x(n)$ from its periodic extension $x_p(n)$. Furthermore, we will assume that the periodic signal $x_p(n)$ is formed from the samples of periodic continuous-time signal $x_p(t)$ with a period T (corresponding to N signal samples within the period, $T = N\Delta t$, and $x_p(n) = x_p(n\Delta t)\Delta t$). The Fourier series coefficients of $x_p(t)$ are defined by

$$X_k = \frac{1}{T}\int_0^T x_p(t)e^{-j2\pi kt/T}dt.$$

Assuming that the sampling theorem is satisfied, the integral can be replaced by a sum (in the sense of Example 2.13)

$$X_k = \frac{1}{T}\sum_{n=0}^{N-1} x(n\Delta t)e^{-j2\pi kn\Delta t/T}\Delta t$$

with $x_p(t) = x(t)$ within $0 \leq t < T$. Using $T/\Delta t = N$, $x(n\Delta t)\Delta t = x(n)$, and $X(k) = TX_k$ this sum can be written in the form

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi kn/N}. \tag{3.3}$$

Therefore, the relation between the DFT and the Fourier series coefficients is

$$X(k) = TX_k. \tag{3.4}$$

Sampling the Fourier transform of a discrete-time signal corresponds to the periodical extension of the original discrete-time signal in time by the period N (equivalent to the period $T = N\Delta t$ in the continuous-time domain). The period N in the discrete-time domain is equal to the number of samples of the Fourier transform within one period in the frequency domain. We can conclude that this periodic extension in time (discretization in frequency) will not influence the possibility to recover the original signal if the original discrete-time signal duration is not longer than N (the number of samples in the Fourier transform of discrete-time signal).

The inverse DFT is obtained by multiplying both sides of the DFT definition (3.1) by $e^{j2\pi km/N}$ and summing over k

$$\sum_{k=0}^{N-1} X(k)e^{j2\pi mk/N} = \sum_{n=0}^{N-1} x(n) \sum_{k=0}^{N-1} e^{j2\pi k(m-n)/N}$$

with

$$\sum_{k=0}^{N-1} e^{j2\pi k(m-n)/N} = \frac{1 - e^{j2\pi(m-n)}}{1 - e^{j2\pi(m-n)/N}} = N\delta(m-n),$$

for $0 \leq m,n \leq N-1$. The inverse discrete Fourier transform (IDFT) of a signal $x(n)$ is

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k)e^{j2\pi nk/N}. \tag{3.5}$$

for $0 \leq n \leq N-1$.

The signal calculated using the IDFT is, by definition, periodic with the period N since

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

Therefore the DFT of a signal $x(n)$ calculated using the signal samples within $0 \leq n \leq N-1$ assumes that the signal $x(n)$ is periodically extended with the period N, that is

$$\text{IDFT}\{\text{DFT}\{x(n)\}\} = \sum_{m=-\infty}^{\infty} x(n+mN)$$

$$\text{with} \sum_{m=-\infty}^{\infty} x(n+mN) = x(n) \text{ for } 0 \leq n \leq N-1.$$

The values of this periodical extension within the basic period are equal to $x(n)$. This is a circular extension of signal $x(n)$. The following notations are also used for this kind of the

signal $x(n)$ extension

$$\text{IDFT}\{\text{DFT}\{x(n)\}\} = x(n \bmod N) = x((n))_N.$$

If the original signal $x(n)$, used in the DFT calculation, is aperiodic, then

$$x(n) = \text{IDFT}\{\text{DFT}\{x(n)\}\}\,(u(n) - u(n-N)),$$

assuming that the initial DFT was calculated for signal samples $x(n)$ within $0 \leq n \leq N-1$.

In literature it is quite common to use the same notation for both $x(n)$ and $\text{IDFT}\{\text{DFT}\{x(n)\}\}$ having in mind that any DFT calculation with N signal samples *implicitly assumes a periodic extension* of the original signal $x(n)$ with period N. Thus, we will use this kind of notation, except in the cases when we want to emphasize a difference in the results when the inherent periodicity in the signal (when the DFT is used) is not properly taken into account.

Example 3.1. For the signals $x(n) = 2\cos(2\pi n/8)$ for $0 \leq n \leq 7$ and $x(n) = 2\cos(2\pi n/16)$ for $0 \leq n \leq 7$ plot the periodic signals $\text{IDFT}\{\text{DFT}\{x(n)\}\}$ with $N = 8$ without calculating the DFTs.

★The periodic extensions of these signals resulting from

$$\text{IDFT}\{\text{DFT}\{x(n)\}\} = \sum_{m=-\infty}^{\infty} x(n+8m)$$

are shown in Fig. 3.2.

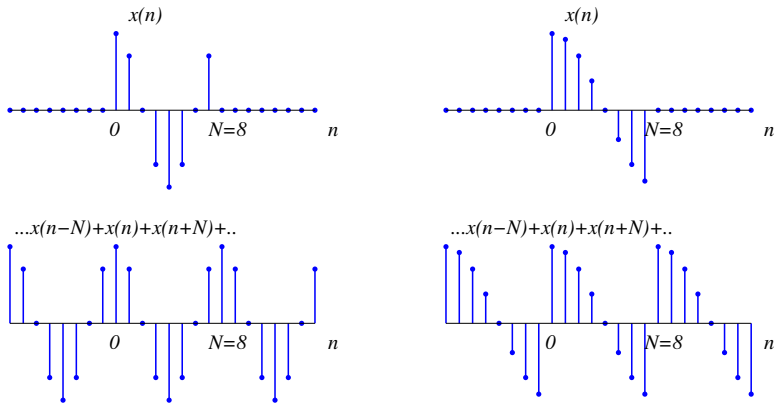


Figure 3.2 Signals $x(n) = 2\cos(2\pi n/8)$ for $0 \leq n \leq 7$ (left) and $x(n) = 2\cos(2\pi n/16)$ for $0 \leq n \leq 7$ (right) along with their periodic extensions $\text{IDFT}\{\text{DFT}\{x(n)\}\}$ with $N = 8$.

Example 3.2. For the signal $x(n)$, whose values are $x(0) = 1$, $x(1) = 1/2$, $x(2) = -1$, and $x(3) = 1/2$, find the DFT with $N = 4$. What is the IDFT for $n = -2$?

★The DFT of this signal is

$$X(k) = \sum_{n=0}^{3} x(n)e^{-j2\pi nk/4} = 1 + \frac{1}{2}e^{-j2\pi k/4} - e^{-j\pi k} + \frac{1}{2}e^{j2\pi k/4}$$
$$= 1 + (-1)^{k+1} + \cos(2\pi k/4).$$

The IDFT is

$$x(n) = \frac{1}{4}\sum_{k=0}^{3}[1 + \cos(2\pi k/4) + (-1)^{k+1}]e^{j2\pi nk/4},$$

for $0 \le n \le 3$. The DFT and IDFT inherently assume the signal and its Fourier transform periodicity. Thus, the result for $n = -2$ is

$$x(-2) = \frac{1}{4}\sum_{k=0}^{3} X(k)e^{j2\pi(-2)\frac{k}{4}} = \frac{1}{4}\sum_{k=0}^{3} X(k)e^{j2\pi(4-2)\frac{k}{4}} = x(4-2) = x(2) = -1.$$

Matrix notation and calculation complexity: The DFT can be written in a matrix form as

$$\begin{bmatrix} X(0) \\ X(1) \\ \vdots \\ X(N-1) \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & e^{-j\frac{2\pi}{N}} & \cdots & e^{-j\frac{2\pi(N-1)}{N}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{-j\frac{2\pi(N-1)}{N}} & \cdots & e^{-j\frac{2\pi(N-1)(N-1)}{N}} \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(N-1) \end{bmatrix} \tag{3.6}$$

or

$$\mathbf{X} = \mathbf{W}\mathbf{x}, \tag{3.7}$$

where $\mathbf{X}$ and $\mathbf{x}$ are the vectors containing the signal and its DFT values

$$\mathbf{X}=[X(0)\ X(1)\ \ldots\ X(N-1)]^T$$
$$\mathbf{x}=[x(0)\ x(1)\ \ldots\ x(N-1)]^T,$$

respectively, while $\mathbf{W}$ is the discrete Fourier transform matrix with the coefficients

$$\mathbf{W} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & W_N^1 & \cdots & W_N^{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & W_N^{(N-1)} & \cdots & W_N^{(N-1)(N-1)} \end{bmatrix}, \tag{3.8}$$

where

$$W_N^k = e^{-j2\pi k/N}$$

is used to simplify the notation.

The number of additions to calculate a DFT is $N-1$ for every $X(k)$ in (3.1). Since there are N DFT coefficients, the total number of additions is $N(N-1)$. From the matrix in (3.6) we can see that the multiplications are not needed for calculation of $X(0)$. There is no need for a multiplication in the first term of every coefficient calculation as well. If we neglect the fact that some other terms in matrix (3.6) may also take values 1, -1, j, or $-j$ then the number of multiplications is $(N-1)^2$. The order of the number of multiplications and the number of additions for the DFT calculation is N^2.

The inverse DFT in a matrix form is

$$\mathbf{x} = \mathbf{W}^{-1}\mathbf{X}, \tag{3.9}$$

with $\mathbf{W}^{-1} = \frac{1}{N}\mathbf{W}^*$, where * denotes complex-conjugate operation. The same calculation complexity analysis holds for the inverse DFT as in the case of the DFT.

3.2 DFT PROPERTIES

Most of the DFT properties can be derived in the same way as for the Fourier transform and the Fourier transform of discrete-time signals.

1. Consider a **signal** $x(n)$ **shifted in time** $x(n-n_0)$. If the DFT of signal $x(n)$ is $X(k) = \text{DFT}\{x(n)\}$ then $X(k)e^{-j2\pi kn_0/N}$ will represent a signal

$$\begin{aligned}\text{IDFT}\{X(k)e^{-j2\pi kn_0/N}\} &= \frac{1}{N}\sum_{k=0}^{N-1} X(k)e^{-j2\pi kn_0/N}e^{j\frac{2\pi}{N}kn} \\ &= \frac{1}{N}\sum_{k=0}^{N-1} X(k)e^{j\frac{2\pi}{N}k(n-n_0)} = x(n-n_0).\end{aligned} \tag{3.10}$$

   Here $x(n-n_0)$ is the signal obtained when $x(n)$ is periodically extended with N first and then this periodic signal is shifted for n_0. The basic period of the original signal $x(n)$ is now within $n_0 \le n \le N+n_0-1$.

   This kind of shift in periodic signals, used in the above relation, is also referred to as a *circular shift*. Thus, with the circular shift

$$\text{DFT}\{x(n-n_0)\} = X(k)e^{-j2\pi kn_0/N}. \tag{3.11}$$

2. For a **modulated signal** $x(n)e^{j2\pi nk_0/N}$ we easily get

$$\text{DFT}\left\{x(n)e^{j2\pi nk_0/N}\right\} = X(k-k_0). \tag{3.12}$$

3. The DFT **is real-valued** if

$$x^*(n) = x(N-n).$$

For a real-valued DFT holds

$$X(k) = X^*(k)$$

or

$$\sum_{n=0}^{N-1} x(n)e^{-j2\pi nk/N} = \sum_{n=0}^{N-1} x^*(n)e^{j2\pi nk/N} = \sum_{n=0}^{N-1} x^*(N-n)e^{j2\pi(N-n)k/N},$$

where $x^*(N)e^{j2\pi Nk/N} = x^*(0)e^{j2\pi 0k/N}$ is used. Since $e^{j2\pi k(N-n)/N} = e^{-j2\pi nk/N}$ we get

$$\sum_{n=0}^{N-1} x(n)e^{-j2\pi nk/N} = \sum_{n=0}^{N-1} x^*(N-n)e^{-j2\pi nk/N}.$$

It means that if $X(k) = X^*(k)$, then $x^*(n) = x(N-n) = x(-n)$.

In the same way, for a real-valued signal $x(n)$ the DFT satisfies the following property

$$X^*(k) = X(N-k).$$

4. **Parseval's theorem** of discrete-time periodic signals relates the energy in the time domain and the frequency domain

$$\begin{aligned}\sum_{n=0}^{N-1} |x(n)|^2 &= \sum_{n=0}^{N-1} x(n)x^*(n) = \sum_{n=0}^{N-1} x(n)\Big(\frac{1}{N}\sum_{k=0}^{N-1} X(k)e^{j2\pi nk/N}\Big)^* \\ &= \frac{1}{N}\sum_{k=0}^{N-1} X^*(k)\sum_{n=0}^{N-1} x(n)e^{-j2\pi nk/N} = \frac{1}{N}\sum_{k=0}^{N-1} |X(k)|^2.\end{aligned}$$

5. **Convolution of two periodic signals** $x(n)$ and $h(n)$, whose period is N, is defined by

$$y(n) = \sum_{m=0}^{N-1} x(m)h(n-m).$$

The DFT of this signal is

$$Y(k) = \mathrm{DFT}\{y(n)\} = \sum_{n=0}^{N-1}\sum_{m=0}^{N-1} x(m)h(n-m)e^{-j2\pi nk/N} = X(k)H(k). \quad (3.13)$$

Thus, the DFT of the convolution of two periodic signals is equal to the product of the DFTs of individual signals. Since the convolution is performed on periodic signals (the DFT inherently assumes signals periodicity), a circular shift of signals is assumed in the calculation. This kind of convolution is called *circular convolution*.

Relation (3.13) indicates that we can calculate convolution of two aperiodic discrete-time signals of a limited duration in the following way:

- Calculate the DFTs of signals $x(n)$ and $h(n)$ with N nonzero samples, to obtain $X(k)$ and $H(k)$. At this point, inherently, we make periodic extensions of $x(n)$ and $h(n)$, with a period N.
- Multiply these two DFTs to obtain the DFT of the output signal $Y(k) = X(k)H(k)$.
- Calculate the inverse DFT to get the convolution (the output signal with N samples)
$$y(n) = \text{IDFT}\{Y(k)\}.$$

This procedure looks computationally more complex than the direct calculation of convolution, by definition. However, due to very efficient and fast routines for the DFT and the IDFT calculation, this way of calculating the convolution could be more efficient than the direct one.

In using this procedure, we have to take care about the length of signals and their DFTs that assume periodic extensions.

Example 3.3. Consider a discrete-time signal
$$x(n) = u(n) - u(n-5).$$
Calculate the convolution $x(n) * x(n)$. Periodically extend signals with period $N = 7$ and calculate the circular convolution (corresponding to the DFT based convolution calculation with $N = 7$). This value of N satisfies the condition that it is larger than the signal duration. Compare the results. What value of N should be used for the period so that the direct convolution corresponds to one period of the circular convolution?

★The signal $x(n)$ and its reversed version $x(-n)$, along with some shifted signals used in the convolution calculation, are presented in Fig. 3.3.

In the circular (DFT) calculation, for example, at $n = 0$, the convolution value is
$$x_p(n) *_n x_p(n) = \sum_{m=0}^{6} x_p(m)x_p(0-m) = 1+0+0+1+1+0+0 = 3.$$

In addition to the term $x(0)x(0) = 1$ which exists in the aperiodic convolution, two terms for $m = 3$ and $m = 4$ appeared due to the periodic extension of the signal. They made that the circular convolution value differs from the convolution of original aperiodic signals. The same situation occurred for $n = 1$ and $n = 2$. For $n = 3$, 4, and 5 the correct result for the aperiodic convolution is obtained using the circular convolution.

From the previous calculation, it could be concluded that if the signal periods in the calculation of the circular convolution were separated by at least two more zero signal samples (if the period N were $N \geq 9$) this difference would not occur (overlapping of the signal samples in the basic period with the extended period samples would be avoided), as shown in Fig. 3.4 for $N = 9$. Then one period of the circular convolution, for $0 \leq n \leq N-1$, would correspond to the original aperiodic convolution.

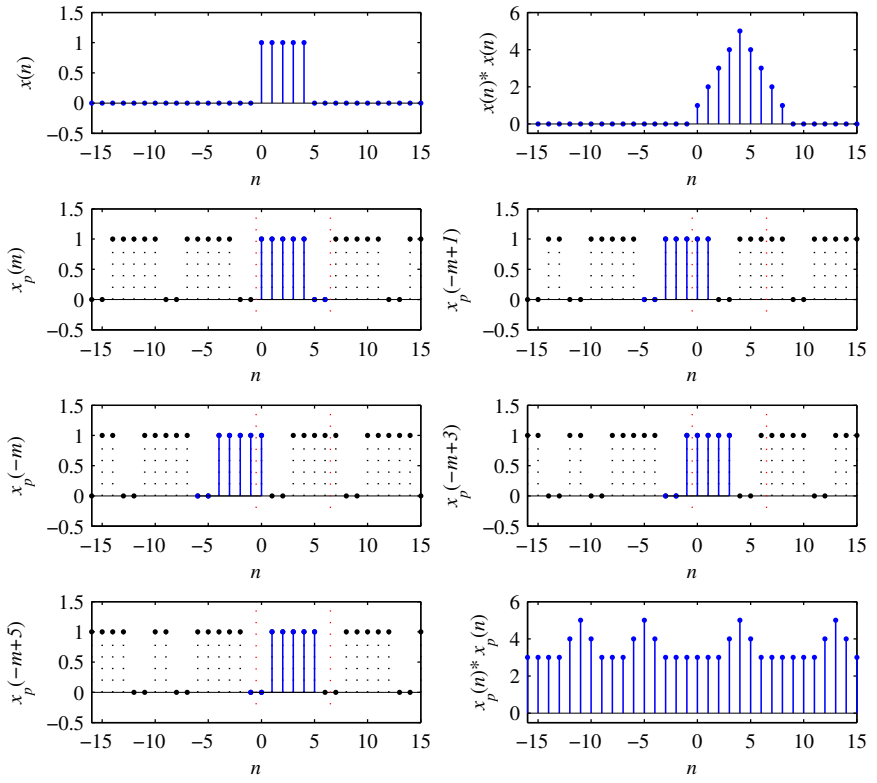


Figure 3.3 Illustration of the discrete-time signal convolution and circular convolution for signals whose length is 5 and the circular convolution is calculated with $N = 7$.

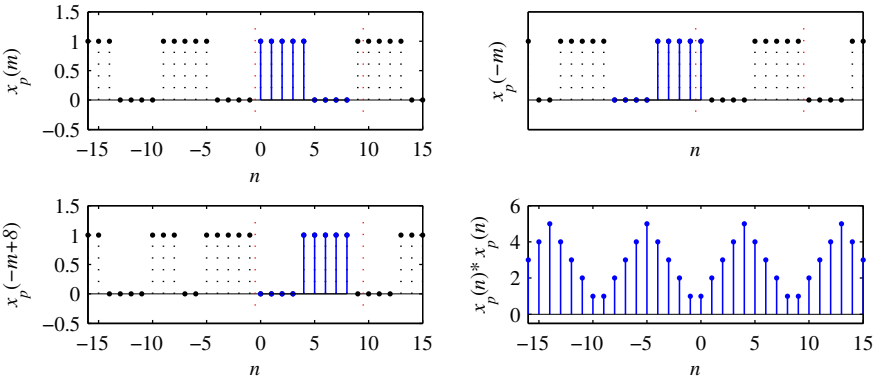


Figure 3.4 Illustration of the discrete-time signal circular convolution for signals whose length is 5 and the circular convolution is calculated with $N = 9$.

Generalization: If a signal $x(n)$ is of the length M, then we can calculate its DFT with any $N \geq M$, and the signal will not overlap with its extended periods, implicitly added using the DFT. If a signal $h(n)$ is of the length L, then we can calculate its DFT with any $N \geq L$. However, if we want to use their DFTs for the convolution calculation (to use the circular convolution), then from the previous example we see that the length of the convolution $y(n)$ is equal $M+L-1$. Therefore, for the DFT-based calculation of the convolution $y(n)$, we have to use at least

$$N \geq M+L-1.$$

This means that both DFTs, $X(k)$ and $H(k)$, whose product results in $Y(k)$, must be at least of $N \geq M+L-1$ duration (period). Otherwise, aliasing (overlapping of the periods) will appear and the circular convolution calculated in this way would not correspond (within the basic period) to the convolution of the original discrete-time (aperiodic) signals.

Duration of an input signal $x(n)$ may be much longer that the duration of the impulse response $h(n)$. For example, an input signal may have tens of thousands of samples, while the impulse response of a discrete system duration is, for example, tens of samples, that is $M \gg L$. The direct convolution of these two signals could be calculated (after first $L-1$ output samples) as

$$y(n) = \sum_{m=n-L+1}^{n} x(m)h(n-m).$$

For every output sample, L multiplications would be used. For a direct DFT application in the convolution calculation we should wait until the end of the signal and then zero-pad both the input signal and the impulse response up to $M+L-1$. This kind of calculation is not efficient. Instead of using the direct DFT calculation, the signal can be split into nonoverlapping sequences whose duration N is of the order of the impulse response duration L,

$$x(n) = \sum_{k=0}^{K-1} x_k(n),$$

where

$$x_k(n) = x(n)[u(n-kN) - u(n-(k+1)N]$$

and $M = KN$ (the input signal can always be zero-padded up to the nearest KN duration, where K is an integer). The output signal is

$$y(n) = \sum_{k=0}^{K-1} \left(\sum_{m=n-L+1}^{n} x_k(m)h(n-m) \right) = \sum_{k=0}^{K-1} y_k(n). \tag{3.14}$$

For the calculation of the convolutions $y_k(n) = x_k(n) *_n h(n)$, the signals $x_k(n)$ and $h(n)$ should be of duration $N+L-1$ only. These convolutions can be calculated after every $N \ll M$ input signal samples. The output sequence $y_k(n)$ duration is $N+L-1$. Since $y_k(n)$, $k = 0,1,\ldots,K-1$, are calculated with the step N in the time-domain, they overlap, although the input signals $x_k(n)$ are nonoverlapping. For the two successive $y_k(n)$ and $y_{k+1}(n)$ and $L \leq N$, $L-1$, the samples within $kN+N \leq n < kN+N+L-1$ overlap.

This effect should be taken into account, by summing the overlapped output samples in $y(n)$, after the individual convolutions $y_k(n) = x_k(n) *_n h(n)$ are calculated using the DFTs, as shown in Fig. 3.5.

3.3 ZERO-PADDING AND INTERPOLATION

The basic period of the DFT $X(k)$, calculated for $k = 0,1,2,\ldots,N-1$, should be considered as having two parts:

- One part of the DFT values for $0 \leq k \leq N/2-1$, which corresponds to the positive frequencies

$$\omega = \frac{2\pi}{N}k \quad \text{or} \quad \Omega = \frac{2\pi}{N\Delta t}k, \quad \text{for} \quad 0 \leq k \leq N/2-1, \tag{3.15}$$

  and the

- Other part being a shifted version of the DFT corresponding to the negative frequencies (in the original aperiodic signal)

$$\omega = \frac{2\pi}{N}(k-N) \quad \text{or} \quad \Omega = \frac{2\pi}{N\Delta t}(k-N), \quad \text{for} \quad N/2 \leq k \leq N-1. \tag{3.16}$$

Illustration of the frequency value correspondence to the frequency index in the DFT is given in Fig. 3.6

We have seen that the DFT of a signal whose duration is limited to M samples can be calculated using any $N \geq M$. In practice, this means that we can add (use) as many zeros, after the nonzero signal $x(n)$ values, as we like. By doing this, we increase the calculation complexity, but we also increase the number of samples within the same frequency range of the Fourier transform.

If we recall that

$$X(k) = X(e^{j\omega})_{|\omega=k\Delta\omega=2\pi k/N} = X(\Omega)_{|\Omega=k\Delta\Omega=2\pi k/(N\Delta t)}, \tag{3.17}$$

holds in the case when the sampling theorem is satisfied, then we see that by increasing N in the DFT calculation, the density of sampling (interpolation) in the Fourier transform of the original signal increases. The DFT interpolation by zero padding the signal in the time domain is illustrated in Fig. 3.7.

The same holds for the frequency domain. If we calculate the DFT of a signal with N samples and then add, for example, N zeros after the region corresponding to the highest frequencies, then by the IDFT of this $2N$ point DFT, we will interpolate the original signal in time. All zero values in the frequency domain should be inserted between two parts (regions) of the original DFT corresponding to positive and negative frequencies.

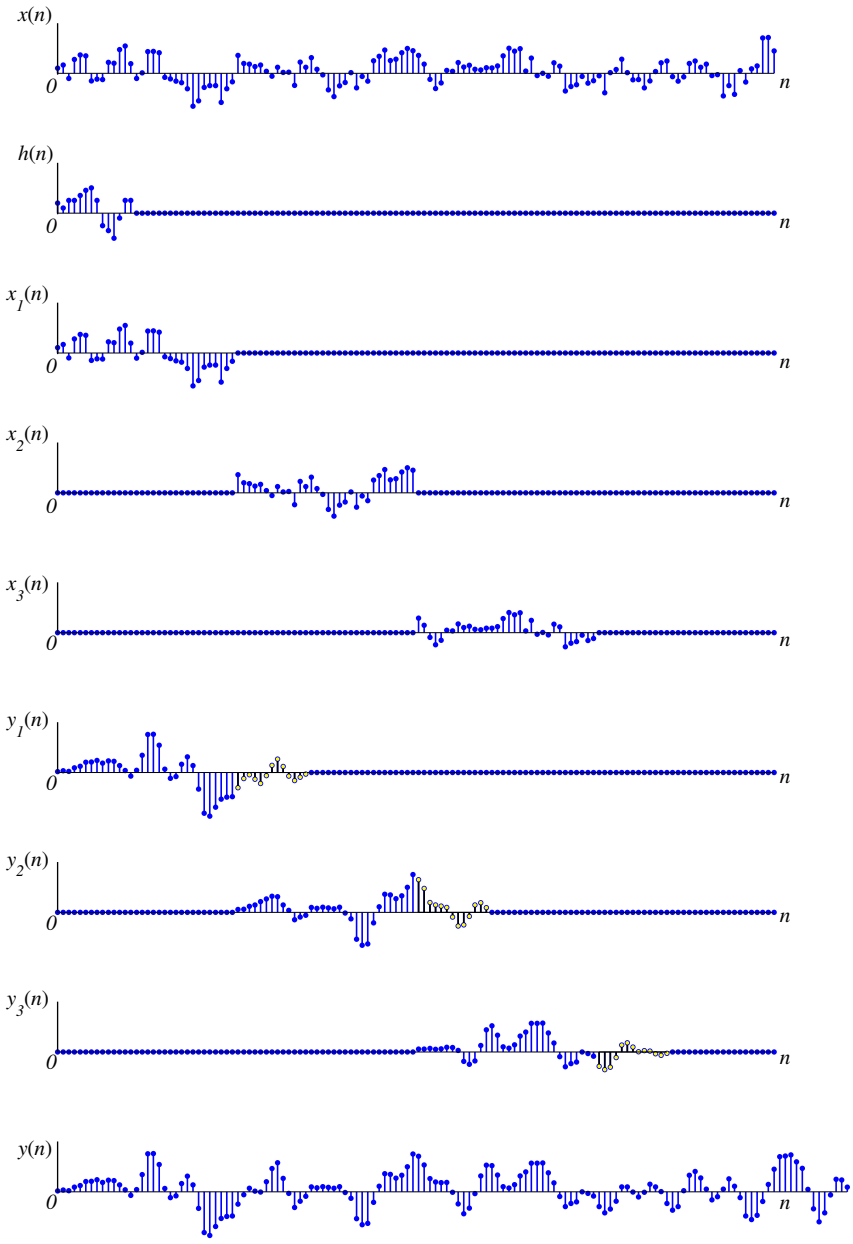


Figure 3.5 Illustration of the convolution, $y(n)$, calculation when the input signal, $x(n)$, duration is much longer than the duration of the system impulse response, $y(n)$.

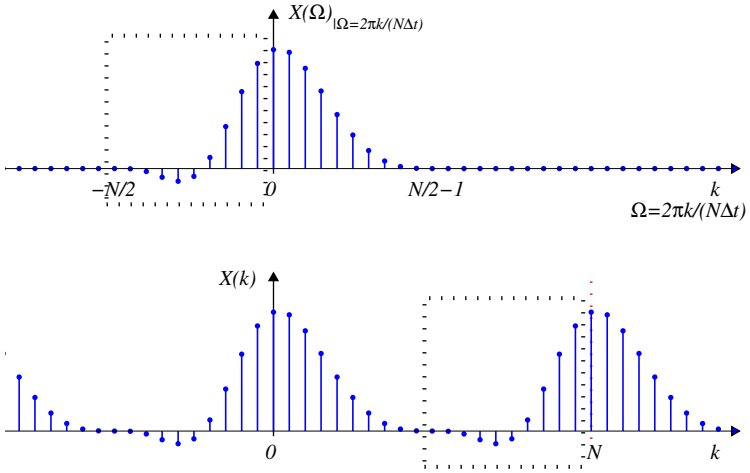


Figure 3.6 Relation between the frequency in the continuous-time and the DFT frequency index.

Example 3.4. The Hann(ing) window for a signal truncation within $-N/2 \le n \le N/2-1$, is

$$w(n) = \frac{1}{2}[1+\cos(\frac{2\pi n}{N})], \quad \text{for } -N/2 \le n \le N/2-1. \tag{3.18}$$

If the original signal values are within $0 \le n \le N-1$ then the Hann(ing) window form is

$$w(n) = \frac{1}{2}[1-\cos(\frac{2\pi n}{N})], \quad \text{for } 0 \le n \le N-1. \tag{3.19}$$

Present the zero-padded forms of the Hann(ing) windows with $2N$ samples.

★The zero-padded form of the Hann(ing) windows used for windowing data within the intervals $-N/2 \le n \le N/2-1$ and $0 \le n \le N-1$ are shown in Fig. 3.8. The DFTs of windows (3.18) and (3.19) are

$$W(k) = N[\delta(k) + \delta(k-1)/2 + \delta(k+1)/2]/2$$

and

$$W(k) = N[\delta(k) - \delta(k-1)/2 - \delta(k+1)/2]/2,$$

respectively. After the presented zero-padding the window DFT realness property $w_{pz}(n) = w_{pz}(n-2N)$ is preserved (for an even N in the case $-N/2 \le n \le N/2-1$ and for an odd N for data within $0 \le n \le N-1$).

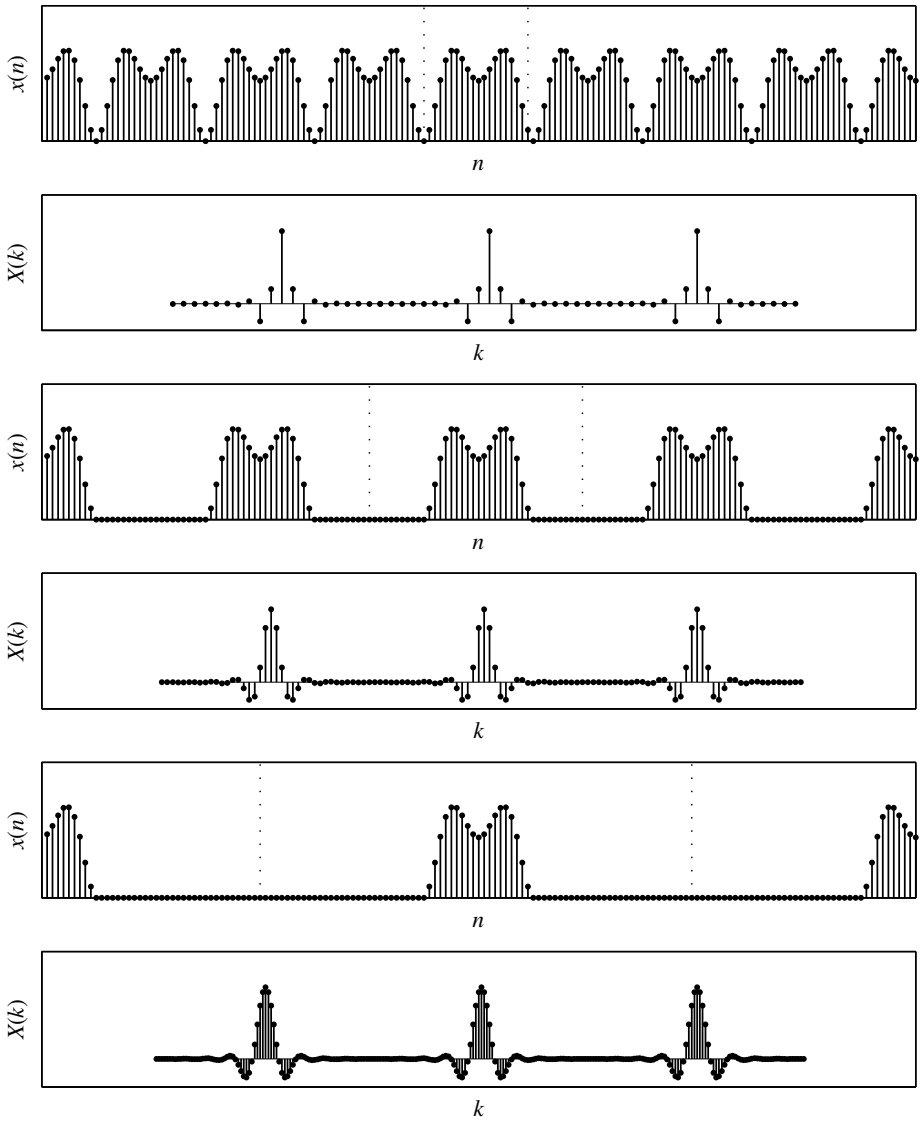


Figure 3.7 Discrete-time signal and its DFT (top two subplots). Discrete-time signal zero-padded and its DFT interpolated (two subplots in the middle). Zero-padding (interpolation) factor was 2. Discrete-time signal zero-padded and its DFT interpolated (two bottom subplots). Zero-padding (interpolation) factor was 4. According to the duality property, the same holds if $X(k)$ were a signal in the discrete-time and $x(-n)$ was its Fourier transform.

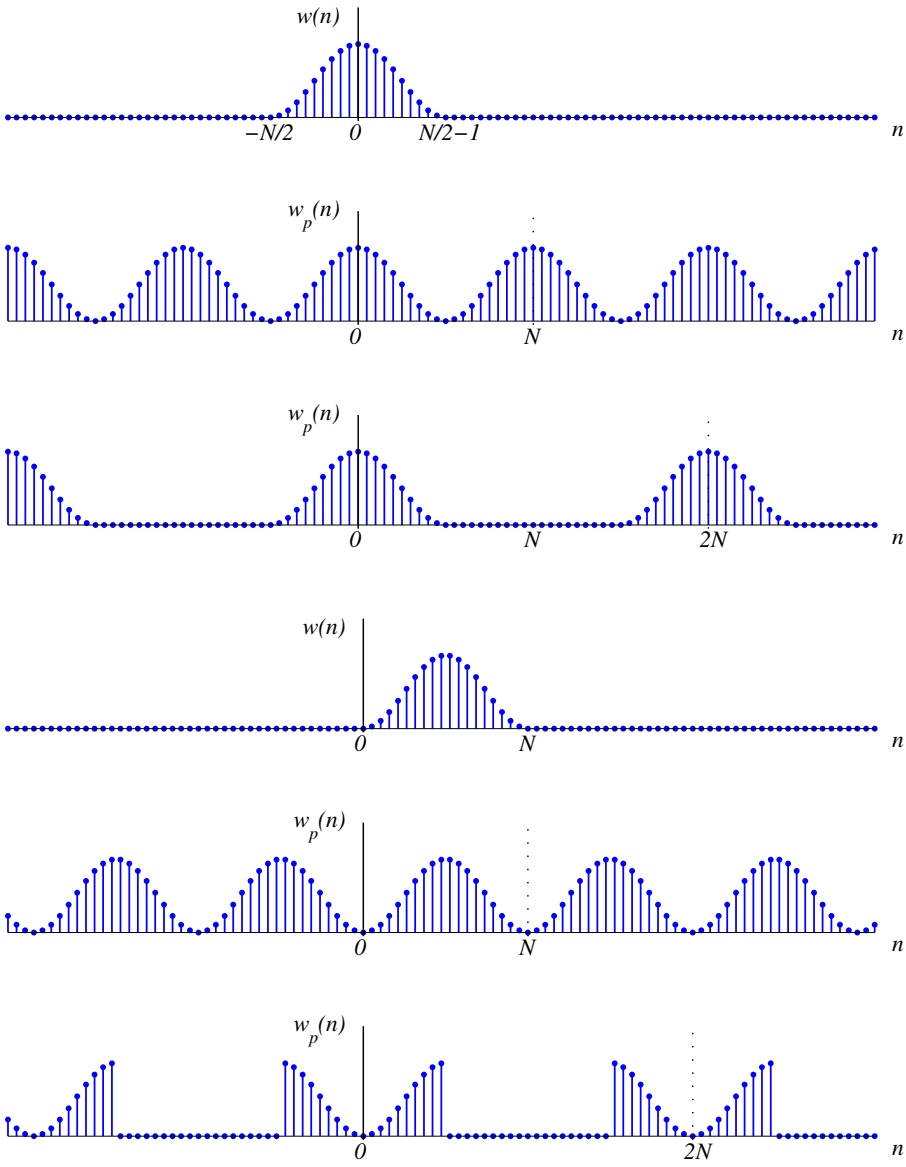


Figure 3.8 Zero-padding of the Hann(ing) windows used to window data within $-N/2 \leq n \leq N/2-1$ and $0 \leq n \leq N-1$.

3.4 RELATION AMONG THE FOURIER REPRESENTATIONS

Presentation of the DFT will be concluded with an illustration (Fig. 3.9) of the relation among the four forms of the Fourier domain signal representations for the cases of:

1. Continuous-time aperiodic signal (**Fourier transform**):

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\Omega)e^{j\Omega t}d\Omega, \qquad X(\Omega) = \int_{-\infty}^{\infty} x(t)e^{-j\Omega t}dt.$$

2. Continuous-time periodic signal (**Fourier series**):

$$x_p(t) = \sum_{m=-\infty}^{\infty} x(t+mT)$$

$$x_p(t) = \sum_{n=-\infty}^{\infty} X_n e^{j2\pi nt/T}, \qquad X_n = \frac{1}{T}\int_{-T/2}^{T/2} x(t)e^{-j2\pi nt/T}dt,$$

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

If the periodic signal is formed by a periodic extension of an aperiodic signal $x(t)$ then there is no signal overlapping (aliasing) in the periodic signal if the original aperiodic signal duration is shorter than the extension period T and the previous relation holds.

3. Discrete-time aperiodic signal (**Fourier transform of discrete-time signals**)

$$x(n) = x(n\Delta t)\Delta t,$$

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega n}d\omega, \qquad X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n},$$

$$X(e^{j\omega}) = \sum_{m=-\infty}^{\infty} X(\Omega + m\frac{2\pi}{\Delta t})_{|\Omega=\omega/\Delta t}.$$

The Fourier transform $X(e^{j\omega})$ of the discrete-time signal $x(n)$ is a periodic extension of the Fourier transform $X(\Omega)$, $\omega = \Omega\Delta t$, of a continuous-time signal $x(t)$. There is no overlapping (aliasing) if the width of the Fourier transform of the original continuous-time signal is shorter than the extension period equal to $2\pi/\Delta t$.

4. Discrete-time periodic signal (**discrete Fourier transform -DFT**)

$$x_p(n) = \sum_{m=-\infty}^{\infty} x(n+mN) = x_p(t)_{|t=n\Delta t},$$

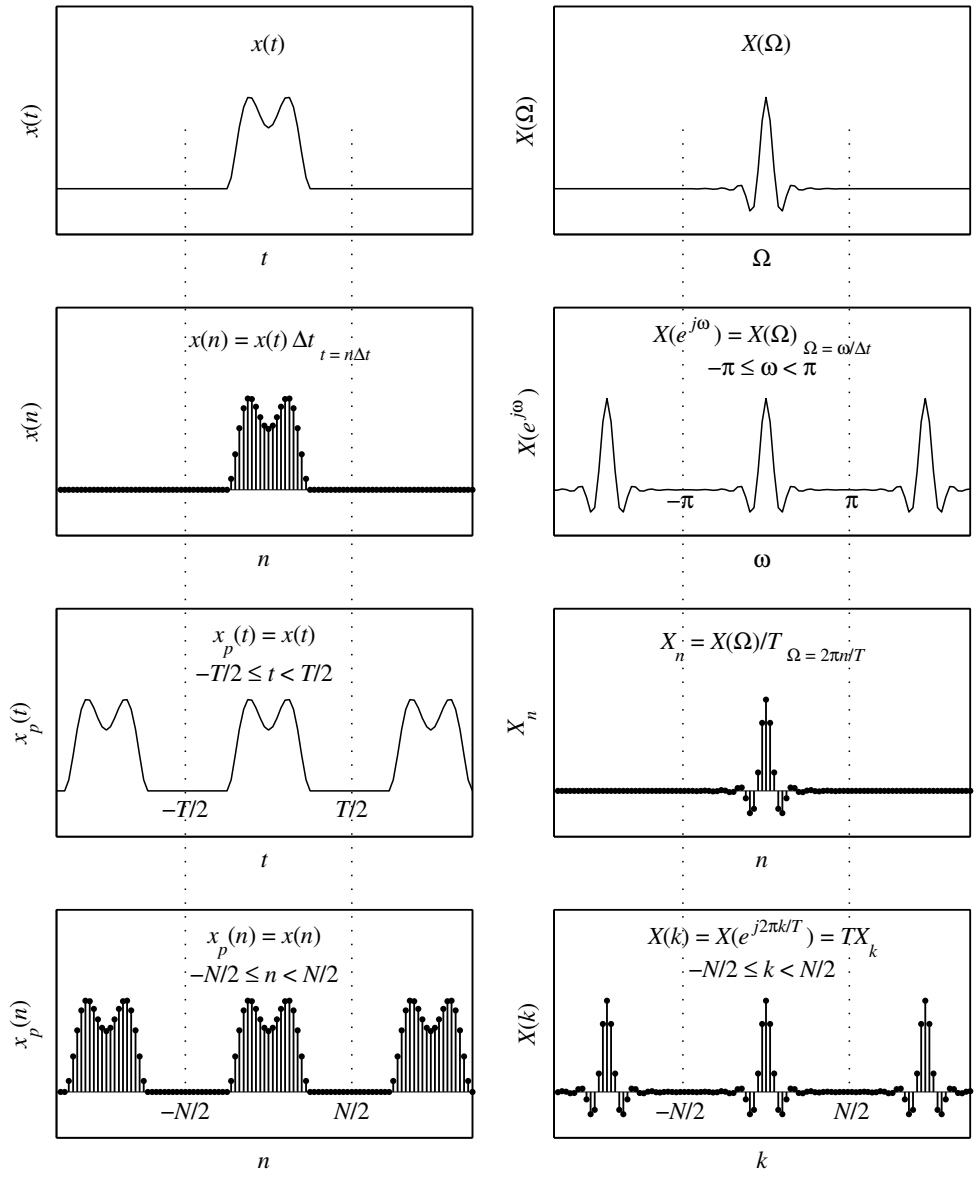


Figure 3.9 Aperiodic continuous-time signal and its Fourier transform (first row). Discrete-time signal and its Fourier transform (second row). Periodic continuous-time signal and its Fourier series coefficients (third row). Periodic discrete-time signal and its discrete Fourier transform (DFT), (fourth row).

$$x_p(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(k)e^{j2\pi nk/N}, \qquad X(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi nk/N},$$

$$X(k) = X(e^{j\omega})_{|\omega=2\pi k/N} = X(\Omega)_{|\Omega=2\pi k/(N\Delta t)} = TX_k.$$

For the periodic discrete-time signal $x_p(n)$, it has been assumed that there is no overlapping of the original aperiodic discrete-time signal $x(n)$ samples, that is, its duration is shorter than the extension period N, $x(n) = x_p(n)$ for $0 \le n \le N-1$.

All forms of the Fourier representations are related and shown in Fig. 3.9.

3.5 FAST FOURIER TRANSFORM

Algorithms that provide efficient calculation of the DFT, with a reduced number of arithmetic operations, are called the fast Fourier transform (FFT). A unified approach to the DFT and the inverse DFT, (3.5), is used. The only differences between the DFT and inverse DFT calculation are in the sign of the exponent and a division of the final result by N.

Here we will present an algorithm based on splitting the signal $x(n)$, with N samples, into two signals $x(n)$ for $0 \le n \le N/2-1$ and $x(n)$ for $N/2 \le n \le N-1$, whose duration is $N/2$. It is assumed that N is an even number. By definition, the DFT of a signal with N samples is

$$\begin{aligned}
\mathrm{DFT}_N\{x(n))\} = X(k) &= \sum_{n=0}^{N-1} x(n)e^{-j2\pi nk/N} \\
&= \sum_{n=0}^{N/2-1} x(n)e^{-j2\pi nk/N} + \sum_{n=N/2}^{N-1} x(n)e^{-j2\pi nk/N} \\
&= \sum_{n=0}^{N/2-1} \left[x(n) + x(n+N/2)(-1)^k\right] e^{-j2\pi nk/N}
\end{aligned}$$

since

$$e^{-j2\pi(n+N/2)k/N} = e^{-j2\pi nk/N}e^{-j\pi k} = (-1)^k e^{-j2\pi nk/N}.$$

For an even number, $k = 2r$, we have

$$\mathrm{DFT}_{N/2}\{g(n)\} = X(2r) = \sum_{n=0}^{N/2-1} g(n)e^{-j2\pi nr/(N/2)}$$

with

$$g(n) = x(n) + x(n+N/2).$$

For an odd number, $k = 2r+1$, follows

$$\mathrm{DFT}_{N/2}\{h(n)\} = X(2r+1) = \sum_{n=0}^{N/2-1} h(n)e^{-j2\pi nr/(N/2)},$$

where

$$h(n) = (x(n) - x(n + N/2))e^{-j2\pi n/N}.$$

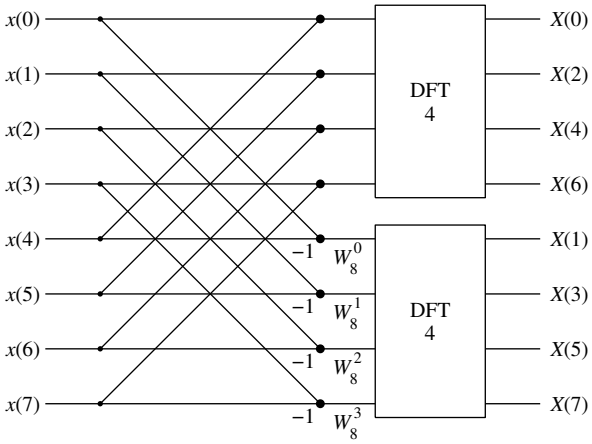


Figure 3.10 DFT of length 8 calculation using two DFTs of length 4.

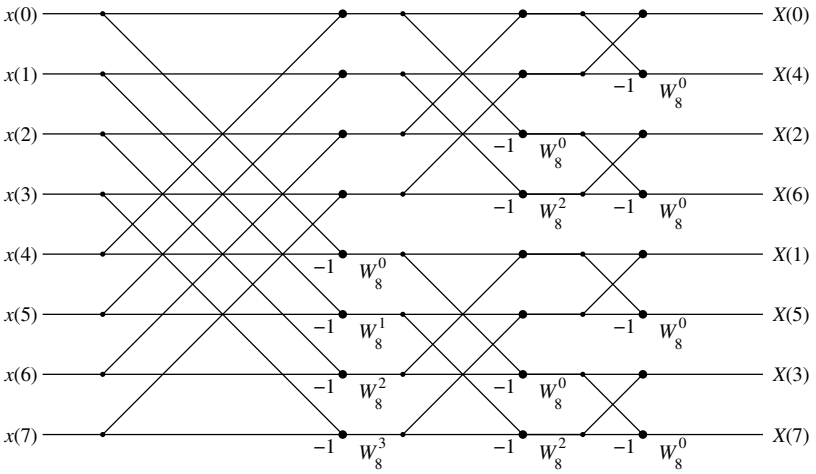


Figure 3.11 FFT calculation scheme obtained by decimation-in-frequency for $N = 8$.

In this way, one DFT of N elements is split into two DFTs of $N/2$ elements. Having in mind that the direct calculation of a DFT with N elements requires an order of N^2 operations, it means that we have reduced the calculation complexity, since

$$N^2 > (N/2)^2 + (N/2)^2.$$

An illustration of the DFT calculation, with $N = 8$, using two DFT with $N/2 = 4$ is shown in Fig. 3.10.

We can continue in the same way and split every DFT with $N/2$ elements into two DFTs with $N/4$, and so on. A complete calculation scheme is shown in Fig. 3.11.

This kind of the DFT calculation is referred to as the decimation-in-frequency algorithm. We can conclude that in this FFT algorithm an order of $N\log_2 N$ of operations is required. Here, it has been assumed that $\log_2 N = p$ is an integer, that is, $N = 2^p$.

If we want to be precise the number of additions is exactly

$$N_{\text{additions}} = N\log_2 N.$$

In the first stage, there are $(N/2-1)$ multiplications. In the second stage, there are $2(N/4-1)$ multiplications. In the next stage would be $4(N/8-1)$ multiplications. Finally in the last stage would be $2^{p-1}\left(\frac{N}{2^p}-1\right) = \frac{N}{2}(\frac{N}{N}-1) = 0$ multiplications ($N = 2^p$ or $p = \log_2 N$). The total number of multiplications, in this FFT algorithm, is

$$\begin{aligned} N_{\text{multiplicat.}} &= \left(\frac{N}{2}-1\right) + 2\left(\frac{N}{4}-1\right) + 4\left(\frac{N}{8}-1\right) + \cdots + 2^{p-1}\left(\frac{N}{2^p}-1\right) \\ &= \frac{N}{2} - 1 + \frac{N}{2} - 2 + \frac{N}{2} - 4 + \cdots + \frac{N}{2} - \frac{N}{2} \\ &= \frac{N}{2}p - (1+2+2^2+\cdots+2^{p-1}) = \frac{N}{2}p - \frac{1-2^p}{1-2} \\ &= \frac{N}{2}\log_2 N - (N-1) = \frac{N}{2}\left[\log_2 N - 2\right] + 1. \end{aligned}$$

If the multiplications by j and $-j$ were excluded the number of multiplications would be additionally reduced.

Example 3.5. Consider a signal $x(n)$ within $0 \le n \le N-1$. Assume that N is an even number. Show that the DFT of $x(n)$ can be calculated as two DFTs, one DFT calculated using the even samples of $x(n)$ and the other DFT obtained using the odd samples of $x(n)$.

★By the DFT definition

$$\begin{aligned} X(k) &= \sum_{n=0}^{N-1} x(n)e^{-j2\pi kn/N} = \sum_{m=0}^{N/2-1} x(2m)e^{-j2\pi k2m/N} + \sum_{m=0}^{N/2-1} x(2m+1)e^{-j2\pi k(2m+1)/N} \\ &= \sum_{m=0}^{N/2-1} x_e(m)e^{-j2\pi km/(N/2)} + e^{-j2\pi k/N}\sum_{m=0}^{N/2-1} x_o(m)e^{-j2\pi km/(N/2)}, \end{aligned} \tag{3.20}$$

where $x_e(m) = x(2m)$ and $x_o(m) = x(2m+1)$ are the signal samples with even and odd indices, respectively. If we use the notation $X_e(k) = \text{DFT}\{x_e(n)\}$ and $X_o(k) = \text{DFT}\{x(n_o)\}$, for $k = 0,1,\ldots,N/2-1$, then

$$X(k) = X_e(k) + e^{-j2\pi k/N}X_o(k) \quad \text{for} \quad k = 0,1,\ldots,N/2-1$$

and

$$X(k) = X_e(k - N/2) - e^{-j2\pi k/N} X_o(k - N/2) \quad \text{for} \quad k = N/2, \ldots, N-1$$

since $X_e(k)$ and $X_o(k)$ are periodic with period $N/2$. Thus, the DFT of N elements is split into two DFTs of $N/2$ elements. If $N/2$ is an even number, we can continue and split two DFTs of $N/2$ elements into four DFTs of $N/4$ elements, and so on. This is a decimation-in-time algorithm, Fig. 3.12.

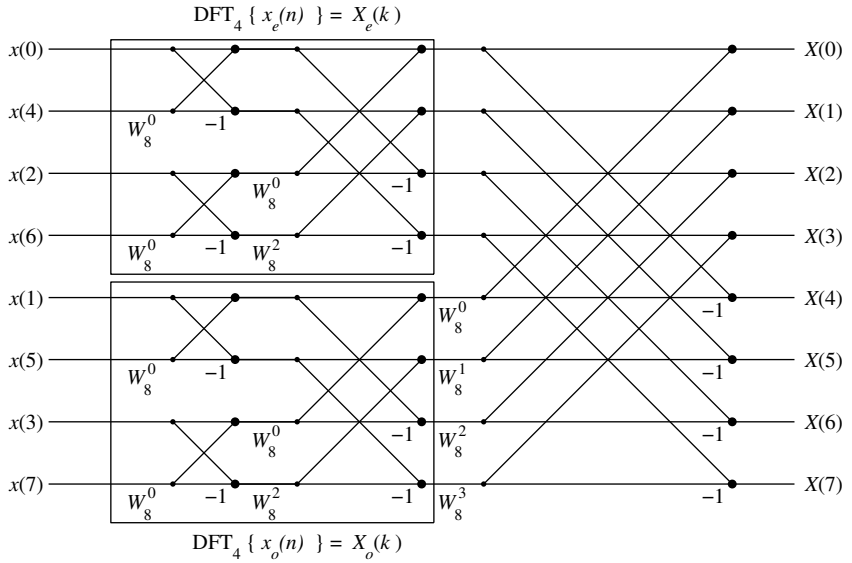


Figure 3.12 Decimation-in-time FFT algorithm for $N = 8$.

Example 3.6. Consider a signal $x(n)$ within $0 \le n \le N-1$. Assume that $N = 3M$. Show that the DFT of $x(n)$ can be calculated using three DFTs of M samples.

★The DFT of $x(n)$ is

$$X(k) = \sum_{n=0}^{3M-1} x(n) e^{-j2\pi kn/(3M)}$$

$$= \sum_{m=0}^{M-1} x(m) e^{-j2\pi km/(3M)} + \sum_{m=M}^{2M-1} x(m) e^{-j2\pi km/(3M)} + \sum_{m=2M}^{3M-1} x(m) e^{-j2\pi km/(3M)}$$

$$= \sum_{m=0}^{M-1} \left[x(m) + x(m+M) e^{-j\frac{2\pi kM}{3M}} + x(m+2M) e^{-j\frac{2\pi k2M}{3M}} \right] e^{-j\frac{2\pi mk}{3M}}.$$

Now we can consider three cases for frequency index k:

- For $r = 3k$, when

$$X(3k) = \sum_{m=0}^{M-1} g(n) e^{-j2\pi mk/M}$$

with $g(n) = x(m) + x(m+M) + x(m+2M)$.

- When $r = 3k+1$, we have

$$X(3k+1) = \sum_{m=0}^{M-1} r(n) e^{-j2\pi mk/M}$$

with $r(n) = \left[x(m) + ax(m+M) + a^2 x(m+2M)\right] e^{-j2\pi m/(3M)}$.

- For $r = 3k+2$, follows

$$X(3k+2) = \sum_{m=0}^{M-1} p(n) e^{-j2\pi mk/M}$$

with $p(n) = \left[x(m) + a^2 x(m+M) + ax(m+2M)\right] e^{-j2\pi 2m/(3M)}$, where $a = e^{-j2\pi/3}$.

Thus, a DFT of $N = 3M$ elements is split into three DFTs of $N/3 = M$ elements. Three DFTs of $N/3$ elements require an order of $3(N/3)^2 = N^2/3$ operations. If, for example, $M = N/3$ is an even number, we can continue and split three DFTs of $N/3$ elements into six DFTs of $N/6$ elements, and so on.

3.6 SAMPLING OF PERIODIC SIGNALS

A periodic signal $x(t)$, with a period T, can be reconstructed if its Fourier series is with limited number of nonzero coefficients, so that $X_k = 0$ for $k > k_m$. This means that the Fourier series coefficients corresponding to frequencies greater than $\Omega_m = 2\pi k_m/T$ are zero-valued. The periodic signal $x(t)$ can be reconstructed from the samples taken with the sampling interval $\Delta t < \pi/\Omega_m = 1/(2f_m)$. The number of samples within the period is $N = T/\Delta t$.

The reconstructed signal is

$$x(t) = \sum_{n=0}^{N-1} x(n\Delta t) \frac{\sin[(n - \frac{t}{\Delta t})\pi]}{N \sin[(n - \frac{t}{\Delta t})\pi/N]}$$

for and odd N and

$$x(t) = \sum_{n=0}^{N-1} x(n\Delta t) e^{j(n - t/\Delta t)\pi/N} \frac{\sin[(n - \frac{t}{\Delta t})\pi]}{N \sin[(n - \frac{t}{\Delta t})\pi/N]}$$

for an even N.

Example 3.7. Samples of a periodic signal $x(t)$ are taken with the sampling interval $\Delta t = 1$. The obtained discrete-time signal samples $x(n)$ are the elements of the signal vector $\mathbf{x} = [0, 2.8284, -2, 2.8284, 0, -2.8284, 2, -2.8284]^T$ for $0 \le n \le N-1$ with $N = 8$. Assuming that the signal satisfies the sampling theorem find its value at $t = 1.5$. Check the accuracy if the original signal values were known, $x(t) = 3\sin(3\pi t/4) + \sin(\pi t/4)$.

★Using the reconstruction formula for an even number of samples, N, within the period we get

$$x(1.5) = \sum_{n=0}^{7} x(n) e^{j(n-1.5)\pi/8} \frac{\sin[(n-1.5)\pi]}{8\sin[(n-1.5)\pi/8]} = -0.2242.$$

This result is equal to the original signal value. Calculation is repeated with $0 \le t \le 8$, with a step 0.01. The reconstructed values of $x(t)$ are presented in Fig. 3.13.

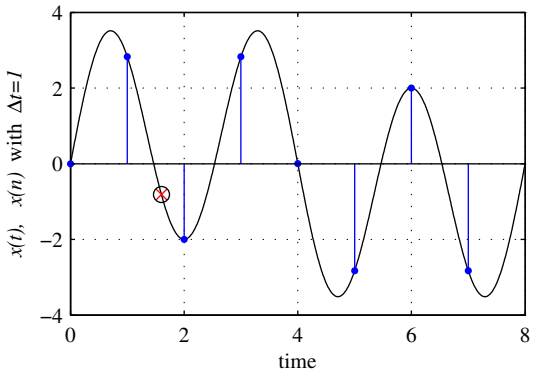


Figure 3.13 Periodic signal reconstructed from its samples at $\Delta t = 1$.

In order to prove the sampling theorem of periodic signals write the signal $x(t)$ in the form of the Fourier series expansion

$$x(t) = \sum_{k=-k_m}^{k_m} X_k e^{j2\pi kt/T}. \tag{3.21}$$

Using N samples of the signal $x(t)$ within the period (assuming that N is an odd number), that is, by sampling the signal at instants $n\Delta t = nT/N$, we get

$$x(n\Delta t) = \sum_{k=-k_m}^{k_m} X_k e^{j2\pi kn/N}.$$

With $(N-1)/2 \geq k_m$ we can write

$$x(n\Delta t)\Delta t = \frac{T}{N}\sum_{k=-k_m}^{k_m} X_k e^{j2\pi kn/N} = \frac{T}{N}\sum_{k=-(N-1)/2}^{(N-1)/2} X_k e^{j2\pi kn/N}.$$

With $x(n\Delta t)\Delta t = x(n)$ and $TX_k = X(k)$ this form of the Fourier series reduces to the DFT and the inverse DFT

$$x(n) = \frac{1}{N}\sum_{k=-(N-1)/2}^{(N-1)/2} X(k)e^{j2\pi kn/N}, \qquad X(k) = \sum_{n=0}^{N-1} x(n)^{-j2\pi kn/N}.$$

Substituting the Fourier series coefficients X_k, expressed in terms of $X(k)$ and $x(n)$, into signal (3.21), with $k_m = (N-1)/2$, we get

$$\begin{aligned} x(t) &= \frac{1}{T}\sum_{k=-\frac{N-1}{2}}^{\frac{N-1}{2}}\sum_{n=0}^{N-1} x(n)e^{-j2\pi k\frac{n}{N}}e^{j2\pi k\frac{t}{T}} = \frac{1}{N}\sum_{n=0}^{N-1}\sum_{k=-\frac{N-1}{2}}^{\frac{N-1}{2}} x(n\Delta t)e^{j2\pi k(\frac{t}{T}-\frac{n}{N})} \\ &= \frac{1}{N}\sum_{n=0}^{N-1} x(n\Delta t)e^{-j2\pi(\frac{t}{T}-\frac{n}{N})(N-1)/2}\frac{1-e^{j2\pi(\frac{t}{T}-\frac{n}{N})N}}{1-e^{j2\pi(\frac{t}{T}-\frac{n}{N})}} \\ &= \sum_{n=0}^{N-1} x(n\Delta t)\frac{\sin[\frac{\pi}{\Delta t}(t-n\Delta t)]}{N\sin[\frac{\pi}{N\Delta t}(t-n\Delta t)]}. \end{aligned}$$

This is the reconstruction formula that can be used to calculate $x(t)$ for any t, based on the signal samples $x(n\Delta t)$ at the instants $n\Delta t$, with $\Delta t < \pi/\Omega_m = 1/(2f_m)$.

In a similar way, the reconstruction formula for an even number of samples N can be obtained.

The sampling theorem reconstruction formula of aperiodic signals follows as a special case as $N \to \infty$, since for a small argument

$$\sin[\frac{\pi}{N\Delta t}(t-n\Delta t)] \to \frac{\pi}{N\Delta t}(t-n\Delta t)$$

and

$$x(t) \to \sum_{n=-\infty}^{\infty} x(n\Delta t)\frac{\sin[\frac{\pi}{\Delta t}(t-n\Delta t)]}{\frac{\pi}{\Delta t}(t-n\Delta t)}.$$

Example 3.8. For a signal $x(t)$ whose period is T, it is known that the signal has components corresponding to the nonzero Fourier series coefficients at the indices $k_1, k_2, \ldots, k_K$. What is the minimum number of signal samples needed to reconstruct the signal? What condition should be satisfied the sampling instants and the frequencies for the reconstruction?

★The signal $x(t)$ can be reconstructed using the Fourier series (1.22). In calculations, a finite number of K nonzero terms will be used,

$$x(t) = \sum_{m=1}^{K} X_{k_m} e^{j2\pi k_m t/T}.$$

Since there are K unknown values X_{k_1}, $X_{k_2}, \ldots, X_{k_K}$ the minimum number of equations to calculate their values is K. The equations are written for K time instants

$$\sum_{m=1}^{K} X_{k_m} e^{j2\pi k_m t_i/T} = x(t_i), \text{ for } i = 1,2,\ldots,K$$

or

$$\begin{aligned}
X_{k_1} e^{j2\pi k_1 t_1/T} + X_{k_2} e^{j2\pi k_2 t_1/T} + \cdots + X_{k_K} e^{j2\pi k_K t_1/T} &= x(t_1) \\
X_{k_1} e^{j2\pi k_1 t_2/T} + X_{k_2} e^{j2\pi k_2 t_2/T} + \cdots + X_{k_K} e^{j2\pi k_K t_2/T} &= x(t_2) \\
&\vdots \\
X_{k_1} e^{j2\pi k_1 t_K/T} + X_{k_2} e^{j2\pi k_2 t_K/T} + \cdots + X_{k_K} e^{j2\pi k_K t_K/T} &= x(t_K).
\end{aligned}$$

A matrix from of this system of equations is

$$\Phi \mathbf{X} = \mathbf{y}, \qquad \mathbf{X} = \Phi^{-1}\mathbf{y}$$

where

$$\mathbf{X} = [X_{k_1} \;\; X_{k_2} \;\; \ldots \;\; X_{k_K}]^T, \qquad \mathbf{y} = [x(t_1) \;\; x(t_2) \;\; \ldots \;\; x(t_K)]^T$$

$$\Phi = \begin{bmatrix}
e^{j2\pi k_1 t_1/T} & e^{j2\pi k_2 t_1/T} & \ldots & e^{j2\pi k_K t_1/T} \\
e^{j2\pi k_1 t_2/T} & e^{j2\pi k_2 t_2/T} & \ldots & e^{j2\pi k_K t_2/T} \\
\vdots & \vdots & \vdots & \vdots \\
e^{j2\pi k_1 t_K/T} & e^{j2\pi k_2 t_K/T} & \ldots & e^{j2\pi k_K t_K/T}
\end{bmatrix}$$

The reconstruction condition is $\det\|\Phi\| \neq 0$ for selected time instants t_i and given frequency indices k_i.

3.7 ANALYSIS OF A SINUSOID USING THE DFT

Analysis and estimation of frequency and amplitude of pure sinusoidal signals is of great importance in many applications.

Consider a simple continuous-time sinusoidal signal

$$x(t) = Ae^{j\Omega_0 t} \tag{3.22}$$

whose Fourier transform is $X(\Omega) = 2\pi A\delta(\Omega - \Omega_0)$. The whole signal energy is concentrated just in one frequency point at $\Omega = \Omega_0$. Obviously, the position of maximum is equal

to the signal frequency. For this operation we will use the notation

$$\Omega_0 = \arg\left\{\max_{-\infty<\Omega<\infty}|X(\Omega)|\right\}. \tag{3.23}$$

Assume that the signal $x(t)$ is sampled with the sampling interval Δt. The discrete-time form of this signal is

$$x(n) = Ae^{j\omega_0 n}\Delta t,$$

where $\omega_0 = \Omega_0\Delta t$. In order to compute the DFT of this signal, we will assume a value of N and calculate

$$X(k) = \sum_{n=0}^{N-1} Ae^{j\omega_0 n}e^{-j2\pi nk/N}\Delta t.$$

In general, the DFT is of the form

$$X(k) = A\sum_{n=0}^{N-1} e^{j\omega_0 n}e^{-j2\pi nk/N}\Delta t = A\frac{1-e^{j\omega_0 N}e^{-j2\pi k}}{1-e^{j\omega_0}e^{-j2\pi k/N}}\Delta t \tag{3.24}$$

$$= Ae^{j((N-1)(\omega_0-2\pi k/N)/2)}\frac{\sin(N(\omega_0-2\pi k/N)/2)}{\sin((\omega_0-2\pi k/N)/2)}\Delta t \tag{3.25}$$

The absolute value of the DFT is given by

$$|X(k)| = |A|\left|\frac{\sin(N(\omega_0-2\pi k/N)/2)}{\sin((\omega_0-2\pi k/N)/2)}\right|\Delta t. \tag{3.26}$$

3.7.1 Leakage Effect

Two cases may appear in the analysis of a sinusoidal signal:

1. The signal frequency is such that

$$\omega_0 = 2\pi k_0/N$$

or $\Omega_0 = 2\pi k_0/(N\Delta t)$, where $0 \le k_0 \le N-1$ is an integer. The sampling interval Δt is such that it is contained an integer number of times within the period $T_0 = 2\pi/\Omega_0$. Then

$$X(k) = A\sum_{n=0}^{N-1} e^{j2\pi k_0 n/N}e^{-j2\pi nk/N}\Delta t = NA\delta(k-k_0)\Delta t. \tag{3.27}$$

Obviously, in this case we can find the signal frequency index from

$$k_0 = \arg\{\max_{0\le k\le N-1}|X(k)|\}. \tag{3.28}$$

Frequency is calculated as $\Omega_0 = 2\pi k_0/(N\Delta t)$ for $0 \le k_0 \le N/2-1$ and $\Omega_0 = 2\pi(k_0-N)/(N\Delta t)$ for $N/2 \le k_0 \le N-1$. This case is illustrated in Fig. 3.14, top row. Noisy signals will be considered later in the book.

The estimate of the signal amplitude is obtained as

$$A = \frac{1}{N\Delta t} X(k_0).$$

2. In reality, the signal period (or Ω_0) is not known in advance (if we knew it, then this analysis would not be needed). So, it is highly unlikely to have the previous case with the frequency on the grid, when $\Omega_0 = 2\pi k_0/(N\Delta t)$ as in Fig. 3.14, top row. More common is the case illustrated in Fig. 3.14, bottom row, when the true signal frequency does not correspond to any DFT sample position. Then, the simple signal of sinusoidal form produces th DFT components at all frequencies since $|X(k)|$ in (3.26) is not zero for any k. This effect that a simple sinusoidal signal produces nonzero DFT values at all frequencies (Fig. 3.14, bottom row) is known as the leakage effect.

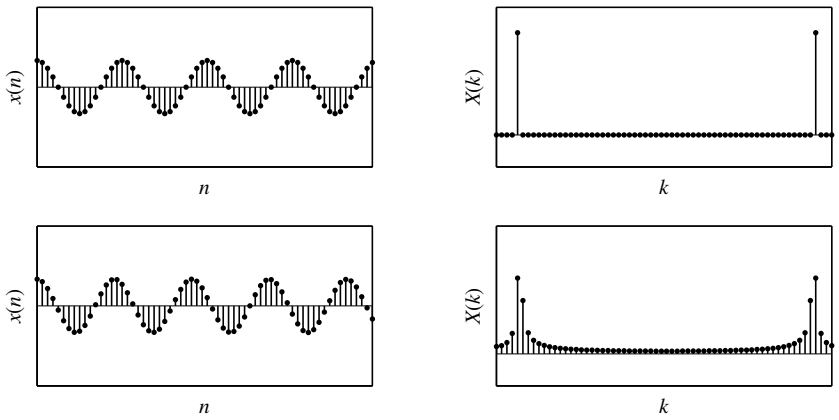


Figure 3.14 Sinusoid $x(n) = \cos(8\pi n/64)$ and its DFT with $N = 64$ (top row) and sinusoid $x(n) = \cos(8.8\pi n/64)$ and its DFT absolute value, with $N = 64$ (bottom row).

The estimation of frequency, based on

$$\hat{k}_0 = \arg\left\{ \max_{0\le k\le N-1} \left| \frac{\sin(N(\omega_0 - 2\pi k/N)/2)}{\sin((\omega_0 - 2\pi k/N)/2)} \right| \right\},$$

will produce an estimation error, defined by

$$e = \Omega_0 - \frac{2\pi}{N\Delta t}\hat{k}_0.$$

The estimation error could be up to a half of the discretization interval, $\Delta\Omega = 2\pi/(N\Delta t)$,

$$-\frac{\pi}{N\Delta t} \le e < \frac{\pi}{N\Delta t} \quad \text{and} \quad \frac{2\pi}{N\Delta t}\hat{k}_0 - \frac{\pi}{N\Delta t} \le \Omega_0 < \frac{2\pi}{N\Delta t}\hat{k}_0 + \frac{\pi}{N\Delta t}. \tag{3.29}$$

Two ways to improve the estimation will be described here.

1. The simplest way to reduce the estimation error is to increase the number of samples and to reduce the discretization interval in frequency $\Delta\Omega = 2\pi/(N\Delta t)$. This could be achieved by appropriate zero-padding in the time domain, before the DFT calculation (corresponding to the interpolation in the frequency domain). This way of error reduction increases the calculation complexity.

2. The other way is based on the window function application in the DFT calculation

$$X(k) = \sum_{n=0}^{N-1} w(n) A e^{j\omega_0 n} e^{-j2\pi nk/N} \Delta t = W\left(e^{j\left(\frac{2\pi k}{N} - \omega_0\right)}\right) \Delta t,$$

where $W(e^{j\omega})$ is the Fourier transform of the window function. Windows, like for example Hann(ing) or Hamming window, smooth the transition and reduce discontinuities at the ending calculation points that cause leakage. A simple realization with, for example, the Hann(ing) window (relation (2.31) and Fig. 2.7)

$$w(n) = \frac{1}{2}\left[1 - \cos(2n\pi/N)\right]\left[u(n) - u(n-N-1)\right].$$

adjusted to the time interval $0 \leq n \leq N-1$, produces

$$\begin{aligned} X_H(k) &= \sum_{n=0}^{N-1} \frac{1}{2}\left[1 - \cos(2n\pi/N)\right] A e^{j\omega_0 n} e^{-j2\pi nk/N} \Delta t \\ &= \frac{A}{2} \sum_{n=0}^{N-1} \left[1 - \frac{1}{2}e^{j2n\pi/N} - \frac{1}{2}e^{-j2n\pi/N}\right] A e^{j\omega_0 n} e^{-j2\pi nk/N} \Delta t \\ &= \frac{1}{2}\left[X_R(k) - \frac{1}{2}X_R(k-1) - \frac{1}{2}X_R(k+1)\right], \end{aligned}$$

where $X_R(k)$ would be the DFT if the rectangular window were used. It is defined by (3.24). The DFT of sinusoids on the grid and outside of the grid, multiplied by a Hann(ing) window, are shown in Fig. 3.15. The leakage effect is reduced. However the DFT is spread over two additional consecutive samples even in the case when the frequency is on the DFT grid, Fig. 3.15(top). In this case the amplitude is estimated as

$$A = \frac{1}{N\Delta t}[X(k_0) + X(k_0+1) + X(k_0-1)].$$

This method is more efficient for the leakage effect reduction than for the improvement in the frequency estimation. However, the idea of using a few neighboring samples in the parameters estimation will be used next to define an approach for accurate frequency estimation.

3.7.2 Displacement

The maximum DFT value and its relation with a few surrounding values of the windowed DFT are used to calculate correction, the *displacement bin*, of the estimated frequency.

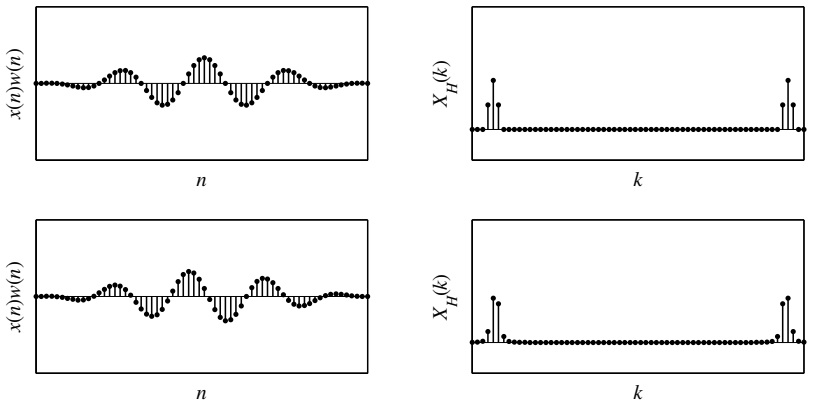


Figure 3.15 Sinusoid $x(n) = \cos(8\pi n/64)$ multiplied by a Hann(ing) window and its DFT with $N = 64$ (top row) and sinusoid $x(n) = \cos(8.8\pi n/64)$ multiplied by a Hann(ing) window and its DFT absolute value, with $N = 64$ (bottom row).

If we apply a window function $w(n)$ in the DFT calculation, we get

$$X(k) = W\left(e^{j(\frac{2\pi k}{N} - \omega_0)}\right)\Delta t.$$

For a given window function it is possible to derive the exact displacement formula for the shift of the maximum position with respect to the detected maxim position. However, instead of deriving an exact formula for every window form, we will present an approach that combines the interpolation and a general fitting polynomial form. It can be used with any window.

We can always interpolate the DFT values $X(k)$ (by appropriate zero-padding of the signal $x(n)$), so that there are several DFT samples within the main lobe. Then, for any symmetric window we can approximate the Fourier transform around the maximum by a quadratic function (in analog domain $X(\Omega) = a\Omega^2 + b\Omega + c$). Since there are three parameters, a, b, and c, in this approximation, we need three Fourier transform values to calculate them. Let us denote the largest sample of the Fourier transform, following from

$$\hat{k}_0 = \arg\left\{\max_{0\le k\le N-1} |X(k)|\right\},$$

by

$$X_0 = |X(\hat{k}_0)|$$

and the two neighboring Fourier transform samples by

$$X_{-1} = |X(\hat{k}_0 - 1)| \quad \text{and} \quad X_1 = |X(\hat{k}_0 + 1)|.$$

By using the Lagrange polynomial interpolation of the second-order, at a point $x = d$, taking the bin index as the independent variable $k_{-1} = -1$, $k_0 = 0$, $k_1 = 1$, with the Fourier

transform values at these points being denoted by X_{-1}, X_0 and X_1, we have the Lagrange second-order polynomial, Fig. 3.16,

$$X(\hat{k}_0+d) = X_{-1}\frac{(d-0)(d-1)}{(-1-0)(-1-1)} + X_0\frac{(d+1)(d-1)}{(0+1)(0-1)} + X_1\frac{(d-0)(d+1)}{(1-0)(1+1)}$$
$$= d^2[-X_0 + X_{-1}/2 + X_1/2)] + d[X_1 - X_{-1}]/2 + X_0. \tag{3.30}$$

This function reaches its maximum at $\partial X(\hat{k}_0+d)/\partial d = 0$, resulting in the displacement bin for the frequency correction

$$d = 0.5\frac{|X(\hat{k}_0+1)| - |X(\hat{k}_0-1)|}{2|X(\hat{k}_0)| - |X(\hat{k}_0+1)| - |X(\hat{k}_0-1)|}, \tag{3.31}$$

with the frequency as in (3.32). The displacement calculation is illustrated in Fig. 3.16. Thus,

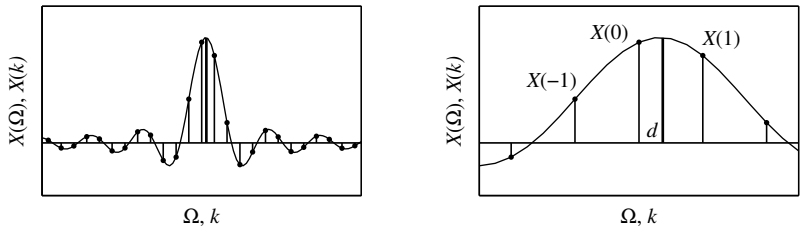


Figure 3.16 Illustration of the displacement bin correction for a true maximum position calculation based on the three neighboring values (full range – left and zoomed graph – right) .

the best frequency estimation is

$$\Omega_0 = \frac{2\pi}{N\Delta t}(\hat{k}_0 + d) \tag{3.32}$$

for $0 \le \hat{k}_0 \le N/2-1$ and $\Omega_0 = \frac{2\pi}{N\Delta t}((\hat{k}_0+d)-N)$ for $N/2 \le \hat{k}_0 \le N-1$.

Example 3.9. The sinusoidal signal $x(t) = A\exp(j\Omega_0 t)$ is sampled with a sampling interval $\Delta t = 1/128$ and $N_0 = 64$ samples are considered. Prior to the DFT calculation, the signal is zero-padded four times, up to $N = 256$. The DFT maximum is detected at the frequency index position $\hat{k}_0 = 95$. The maximum DFT value is $X(95) = 0.9936$. Neighboring DFT values are $X(96) = 0.9432$ and $X(94) = 0.8470$. Calculate the displacement bin d and estimate the value of signal frequency Ω_0.

★The displacement bin value is

$$d = 0.5\frac{0.9432 - 0.8470}{1.9872 - 0.9432 - 0.8470} = 0.2442.$$

The total number of samples in the DFT calculation was $N = 4N_0 = 256$, meaning that the value $\hat{k}_0 = 95$ is within the first half of the samples (corresponding to positive frequency Ω_0). Therefore, we can use (3.32) for the frequency calculation

$$\Omega_0 = \frac{2\pi}{N\Delta t}(\hat{k}_0 + d) = 95.2442\pi.$$

The true signal used in simulation was $x(t) = \exp(j95.25t)/64$, with the estimation error $e = 95.25 - 95.2442 = 0.0058$. If the position of the maximum was only used the estimated frequency would be 95π with an error of $e = 0.25$.

It is possible to derive the exact displacement formula for some specific windows, based on their Fourier transform function. For example, for the Hann(ing) window the exact displacement formula is

$$d_H = \frac{1.5[|X(\hat{k}_0+1)| - |X(\hat{k}_0-1)|]}{|X(\hat{k}_0-1)|\left(1+\frac{|X(\hat{k}_0+1)|}{|X(\hat{k}_0)|}\right) + |X(\hat{k}_0)| + |X(\hat{k}_0+1)|}. \tag{3.33}$$

After the displacement is calculated the signal can also be modulated for the displacement frequency shift in order to produce a signal with the frequency on the frequency grid. This is especially important if we expect that the signal contains much smaller higher-order harmonics that were masked with strong values of the dominant harmonic. If we detected that the k_0th harmonic is dominant and displaced for d then this harmonic should be removed from the signal modulated by the resulting estimated frequency. The DFT of the new signal is used for the analysis of the second largest harmonic and so on.

3.8 DISCRETE COSINE AND SINE TRANSFORMS

The DFT of signal satisfies many desirable properties. Its calculation is simple and efficient using the FFT algorithm. With the DFT calculation the signal periodic extension is assumed and embedded in the discrete transform. However, this periodic extension of signal will, in general, introduce significant signal change (corresponding to *discontinuities* in continuous time) at the period ending points Fig. 3.17 (first and second row). This change (discontinuity) will significantly worsen the DFT coefficients convergence and increase the number of coefficients needed in the signal reconstruction for a given accuracy. In order to reduce influence of this effect and to improve convergence of signal transform coefficients the signal could be extended in an appropriate way.

The discrete cosine transforms (DCT) and the discrete sine transforms (DST) are used to analyze real-valued discrete-time signals, periodically extended to produce even or odd signal forms, respectively. However, this extension is not straightforward for discrete-time signals.

Consider a discrete-time signal of duration N, when $x(n)$ takes nonzero values for $0 \leq n \leq N-1$. If we try with a direct extension (using all signal values) and form a periodic

signal $y(n)$, whose basic period is of duration $2N$, as

$$y(n) = \begin{cases} x(n) & \text{for} \quad 0 \le n \le N-1 \\ x(2N-n-1) & \text{for} \quad N \le n \le 2N-1 \end{cases}$$

the obtained signal is not even, Fig. 3.17(third row). It is obvious that $y(n)$ does not satisfy the condition $y(n) = y(-n) = y(2N-n)$, required for a real-valued DFT. The same holds for an odd extension, Fig. 3.17(fourth row),

$$y(n) = \begin{cases} x(n) & \text{for} \quad 0 \le n \le N-1 \\ -x(2N-n-1) & \text{for} \quad N \le n \le 2N-1 \end{cases}.$$

One of our goals, to have a real-valued transform after the periodic extension of a real-valued signal, is not achieved. However, from Fig. 3.17(third and fourth row) we can see that the signals $y(n)$ are even (or odd) with respect to the vertical line at $n = -1/2$. Thus, if we add zeros between every sample of $y(n)$ and assume that the position which was at $n = -1/2$ in the initial signal is the new coordinate origin, $n = 0$, in the new signal $z(n)$, then these signals will be even and odd, respectively, Fig. 3.17(last two rows). This is just one of possible extensions to make the original discrete-time signal even (or odd). Several forms of the DCT and DST are defined based on other ways of getting an even (odd) signal extension.

The most commonly used form of the DCT is the so called DCT-II or just DCT. If no form of the DCT is referred to in its name, then it is assumed that DCT-II form is used. It will be presented here. Signal periodic extension for this transform corresponds to the already described one in Fig. 3.17. The DCT definition is

$$C(k) = \sum_{n=0}^{N-1} 2x(n) \cos\left(\frac{2\pi(2n+1)}{4N}k\right), \qquad 0 \le k \le N-1.$$

This transform will be derived and explained next. There are two main advantages of this transform over the standard DFT calculation. The DCT coefficients are real-valued for a real-valued signal. This transform can produce a better energy concentration than the DFT. In order to understand why a better energy concentration can be obtained we will compare the DCT to the standard DFT

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi nk/N}, \quad 0 \le k \le N-1$$

calculation procedures for a real-valued signal $x(n)$ with N samples.

In the DCT calculation it is assumed that the signal is extended as an even function, by creating a sequence

$$y(n) = \begin{cases} x(n) & \text{for} \quad 0 \le n \le N-1 \\ x(2N-n-1) & \text{for} \quad N \le n \le 2N-1. \end{cases}$$

This extension eliminates possible signal discontinuities at the period ending points. Thus, in general the Fourier transform of such a signal will converge faster, requiring fewer coefficients

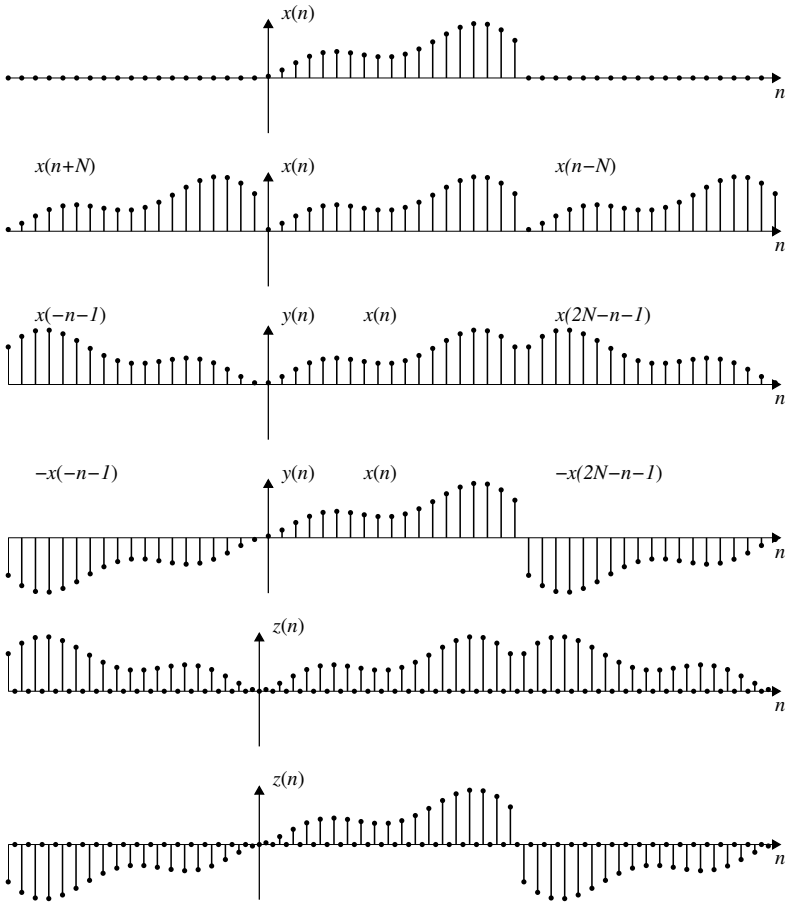


Figure 3.17 Illustration of a signal $x(n)$, its periodic extension corresponding to the DFT, an even and odd discrete-time signal extension corresponding to the DCT and DST of type II.

in the reconstruction. A zero value is then inserted between every pair of samples and an even signal $z(n)$, with period $4N$, is formed

$$z(2n+1) = y(n), \qquad z(2n) = 0.$$

The $4N$-sample DFT of $z(n)$ is calculated

$$X_C(k) = \text{DFT}\{z(n)\} = \sum_{n=0}^{4N-1} z(n) e^{-j2\pi nk/(4N)} = \sum_{n=0}^{2N-1} z(2n+1) e^{-j2\pi(2n+1)k/(4N)}$$
$$= \sum_{n=0}^{2N-1} y(n) e^{-j2\pi(2n+1)k/(4N)} = \sum_{n=0}^{N-1} 2x(n) \cos\left(\frac{2\pi(2n+1)k}{4N}\right) = C(k).$$

Only N terms of the transform are used and the DCT values are obtained.

Since the basis functions are orthogonal the inverse DCT is obtained by multiplying both sides of the DCT by $\cos(\frac{2\pi(2m+1)k}{4N})$ and summing over $0 \leq k \leq N-1$,

$$\sum_{n=0}^{N-1} 2x(n) \sum_{k=0}^{N-1} w_k \cos(\frac{2\pi(2n+1)k}{4N}) \cos(\frac{2\pi(2m+1)k}{4N})$$
$$= \sum_{k=0}^{N-1} w_k C(k) \cos(\frac{2\pi(2m+1)k}{4N}),$$

where $w_0 = 1/2$ and $w_k = 1$ for $k \neq 0$. Since

$$\sum_{k=0}^{N-1} w_k \cos(\frac{2\pi(2n+1)k}{4N}) \cos(\frac{2\pi(2m+1)k}{4N}) = \frac{N}{2}\delta(m-n)$$

we get

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} w_k C(k) \cos(\frac{2\pi(2n+1)k}{4N}). \tag{3.34}$$

A symmetric relation, with the same coefficients in the time and frequency domain, is

$$C(k) = v_k \sum_{n=0}^{N-1} x(n) \cos(\frac{2\pi(2n+1)k}{4N})$$
$$x(n) = \sum_{k=0}^{N-1} v_k C(k) \cos(\frac{2\pi(2n+1)k}{4N}),$$

where $v_0 = \sqrt{1/N}$ and $v_k = \sqrt{2/N}$ for $k \neq 0$.

In a similar way the discrete sine transforms are defined. The most common form is the DST of type II (DST-II), whose definition is

$$S(k) = \sum_{n=0}^{N-1} 2x(n) \sin(\frac{2\pi(2n+1)}{4N}(k+1))$$

for $0 \leq k \leq N-1$. Its relation to the DFT can be established by creating a sequence

$$y(n) = \begin{cases} x(n) & \text{for} \quad 0 \leq n \leq N-1 \\ -x(2N-n-1) & \text{for} \quad N \leq n \leq 2N-1 \end{cases}$$

Zero values are inserted and a signal $z(n)$ is formed as

$$z(2n+1) = y(n)$$
$$z(2n) = 0.$$

Again a $4N$-sample DFT is calculated

$$X_S(k) = \sum_{n=0}^{4N-1} z(n)e^{-j2\pi nk/(4N)} = \sum_{n=0}^{2N-1} y(n)e^{-j2\pi(2n+1)k/(4N)}$$
$$= \text{Im}\left\{ \sum_{n=0}^{N-1} 2jx(n)\sin(\frac{2\pi(2n+1)k}{4N}) \right\} = S(k),$$

with N terms of the transform being used. The DST is the imaginary part of this DFT.

Example 3.10. Consider the signal

$$x(n) = \cos(2\pi(2n+1)/64) + 0.75\cos(7\pi(2n+1)/64).$$

Calculate its DFT with $N = 32$. Plot the periodic extension of the signal $x(n)$. Plot its even extension $y(n)$. Calculate the DFT (the DCT) of such a signal and discuss the results.

★Signal $x(n)$, along with its extended versions and corresponding transforms, is presented in Fig. 3.18. Better energy concentration in the DCT is due to the introduced symmetry in $y(n)$. The

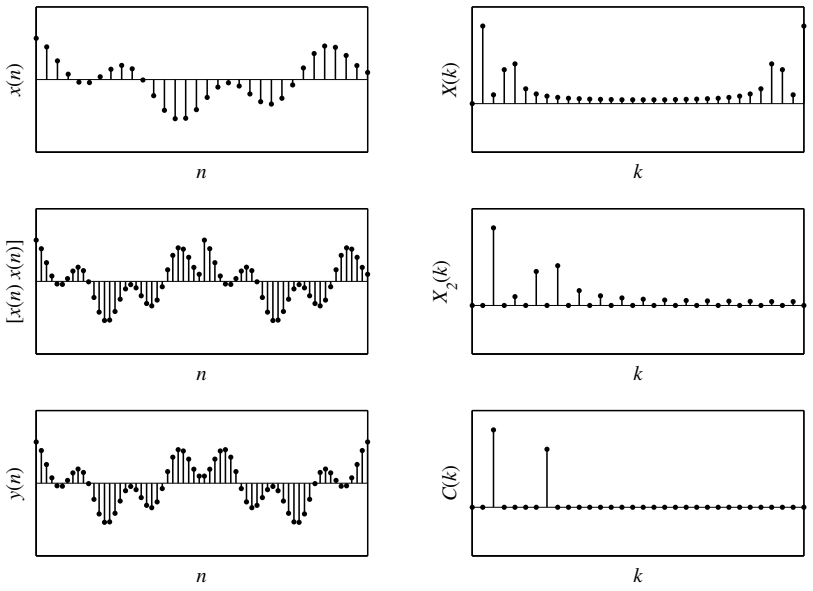


Figure 3.18 Illustration of the DCT calculation.

artificial discontinuity in the DFT, which causes its slow convergence, is eliminated in the DCT.

By using periodic extensions in the cosine transform, the convolution property of the DFT is lost. Thus, this kind of transforms may be used for a signal reconstruction and compression but not in the realization of discrete systems, unless they are properly related to the corresponding DFT values (see Problem 3.10).

Example 3.11. For the signal

$$x(n) = \sin(2\pi(2n+1)/64) - 0.5\cos(7\pi(2n+1)/64).$$

Calculate its DFT with $N = 32$. Plot the periodic extension of this signal. Plot even and odd extensions $y(n)$ of $x(n)$. Calculate the DCT and DST. Comment the results.

★The signal $x(n)$, with its periodic extensions $y_c(n)$ and $y_s(n)$, corresponding to the DFT, DCT, and DST, respectively, is presented in Fig. 3.19(left), as $x(n)$, $[x(n)\ x(n)]$, $y_c(n)$, and $y_s(n)$, respectively. The corresponding transforms are shown in the right panels of this figure. Note that the convergence of the DFT and DCT is similar. Here the DST converges faster, since its extension is "smoother".

3.9 DISCRETE WALSH-HADAMARD AND HAAR TRANSFORMS

Two discrete signal transforms that can be calculated without using multiplications will be presented next. One of them will be used to explain the basic principle of the wavelet transform calculation as well.

Let us consider a two-sample signal $x(n)$, with $N = 2$. The corresponding two-sample DFT is

$$X(k) = \sum_{n=0}^{1} x(n)e^{-j2\pi nk/2} = x(0) + (-1)^k x(1).$$

It can be calculated without using multiplications, $X(0) = x(0) + x(1)$ and $X(1) = x(0) - x(1)$. Now we can show that it is possible to define basis functions for any signal duration in such a way that the multiplications are not used in the signal transformation. These transform values will be denoted by $H(k)$. For two-sample signal case

$$H(0) = x(0) + x(1), \text{ for } k = 0 \text{ and } \quad H(1) = x(0) - x(1), \text{ for } k = 1.$$

The whole frequency interval is represented by a low-frequency value $X(0)$ and a high-frequency value $X(1)$. In a matrix form

$$\begin{bmatrix} H(0) \\ H(1) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \end{bmatrix}. \tag{3.35}$$

The transformation matrix is

$$\mathbf{T}_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}. \tag{3.36}$$

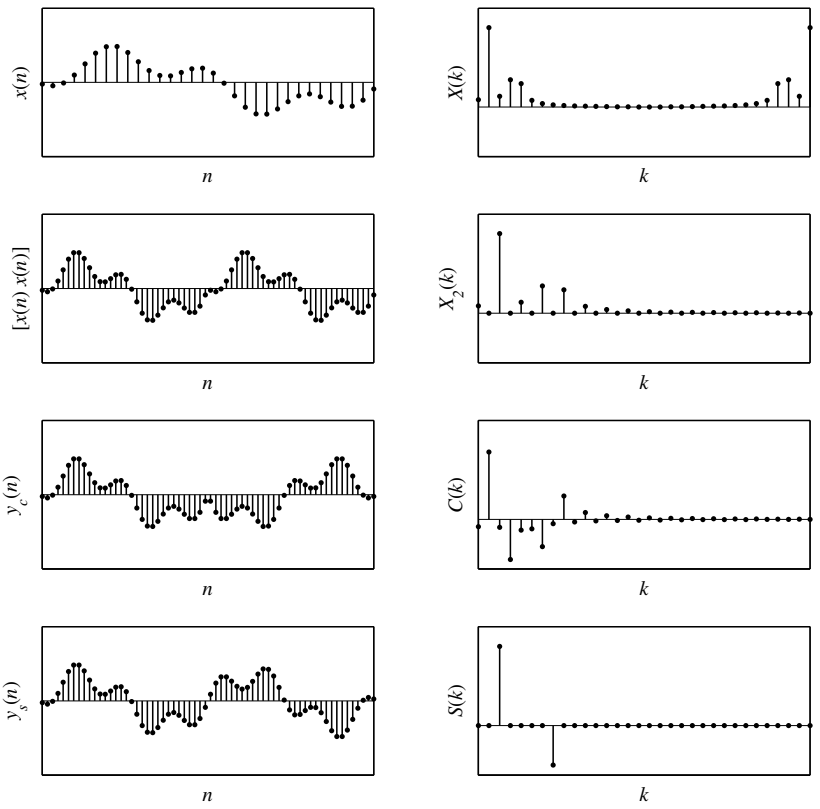


Figure 3.19 Signal and its periodic extensions, corresponding to: the DFT (second row), the cosine transform (third row), and the sine transform (fourth row). Positive frequencies for the DFT are shown.

Example 3.12. For the signal shown in Fig. 3.20 calculate the two-sample DFT for every pair of signal samples

$$H_n(0) = y_L(n) = x(2n) + x(2n+1)$$
$$H_n(1) = y_H(n) = x(2n) - x(2n+1)$$

for $0 \leq n \leq N/2 - 1$. Discuss the results.

★The values of lowpass part, $H_n(0) = y_L(n)$, and the highpass part, $H_n(1) = y_H(n)$, are calculated and are presented in Fig. 3.20. The signal $y_L(n)$ is a low-frequency and smoothed version of the original signal, while the signal $y_H(n)$ contains the details that are lost in the smoothed version $y_L(n)$.

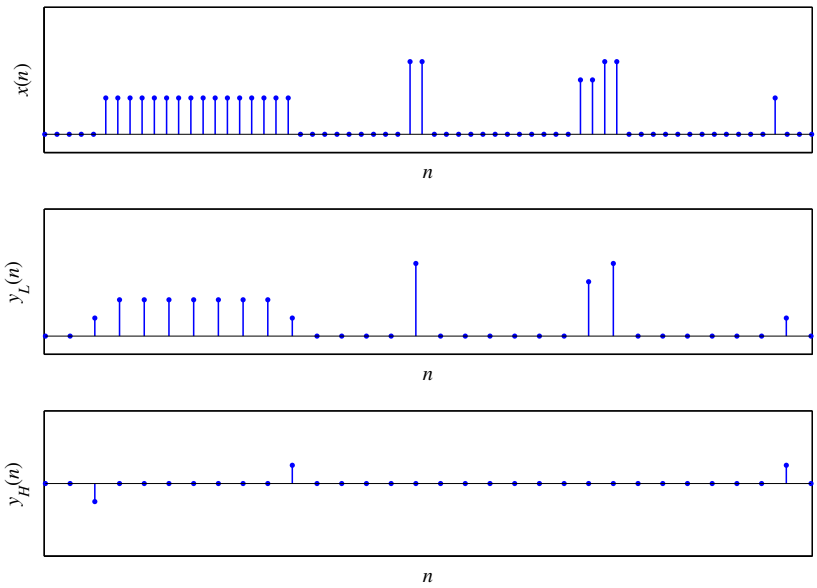


Figure 3.20 Original signal $x(n)$ and its two-sample lowpass part $y_L(n)$ and highpass part $y_H(n)$.

The original signal values may easily be reconstructed from $H_n(0) = y_L(n)$ and $H_n(1) = y_H(n)$ as

$$\begin{bmatrix} x(2n) \\ x(2n+1) \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} H_n(0) \\ H_n(1) \end{bmatrix}$$

for $0 \leq n \leq N/2 - 1$.

In some cases the smoothed version $y_L(n)$, with a half of the samples of the original signal, (3.20), is quite good representative of the original signal, so there is no need to use corrections. Note that for many instants, the correction is zero as well. This result can be used as a basis for the signal compression, when the signal is presented with a reduced set of samples, with no significant distortion.

There are two possibilities to continue and apply the two-point DFT scheme to a signal with N samples:

- The first one consists in splitting further both existing (lowpass and highpass) signals, $y_L(n)$ and $y_H(n)$, into their corresponding lowpass and highpass parts. This scheme leads to the discrete Walsh-Hadamard transform, shown in Fig. 3.21 for the signal $x(n)$ from Fig. 3.20.

- In the second case, the splitting is done for the lowpass part, $y_L(n)$, only, while the highpass correction, $y_H(n)$, is kept unchanged. This scheme leads to the Haar wavelet transform, Fig. 3.22.

These two transforms will be explained in details next.

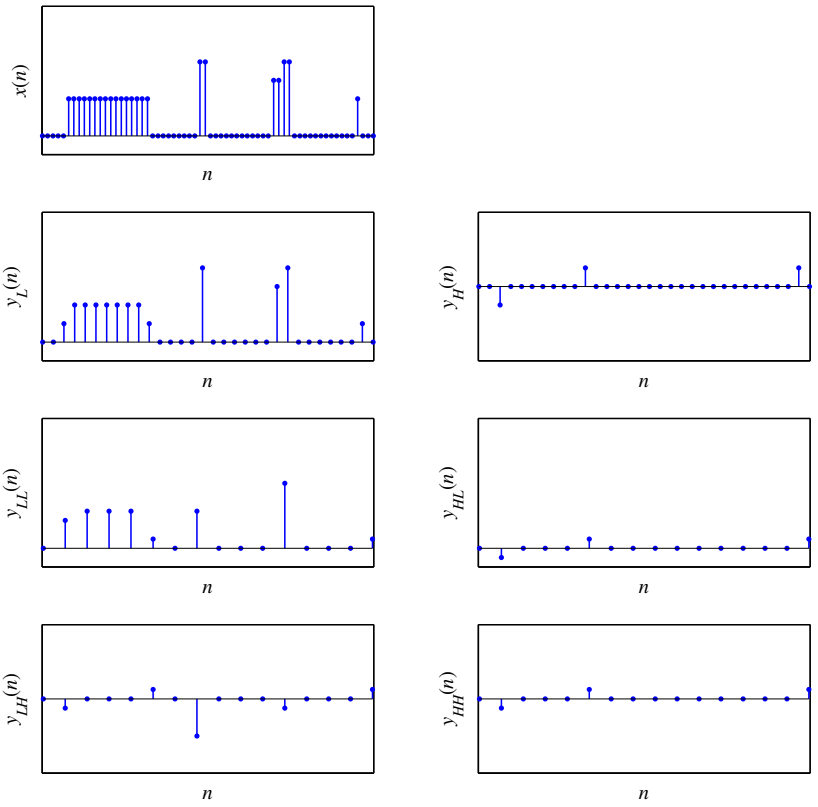


Figure 3.21 Illustration of the procedure leading to the Walsh-Hadamard transform calculation.

3.9.1 Discrete Walsh-Hadamard Transform

Let us continue the idea of splitting both (lowpass and highpass) parts of the signal and define a transformation of a four-sample signal. For this signal form two auxiliary two-sample signals $y_L(n)$ and $y_H(n)$ as

$$y_L(0) = x(0) + x(1),\ y_L(1) = x(2) + x(3) \tag{3.37}$$
$$y_H(0) = x(0) - x(1),\ y_H(1) = x(2) - x(3). \tag{3.38}$$

They represent low-frequency and high-frequency parts of the pairs: $x(0)$, $x(1)$ and $x(2)$, $x(3)$ of two-sample signals. The lowpass part of the auxiliary two-sample lowpass signal $y_L(n)$ is

$$H(0) = y_L(0) + y_L(1) = x(0) + x(1) + x(2) + x(3).$$

The highpass part of the auxiliary two-sample lowpass signal $y_L(n)$ is

$$H(1) = y_L(0) - y_L(1) = x(0) + x(1) - x(2) - x(3).$$

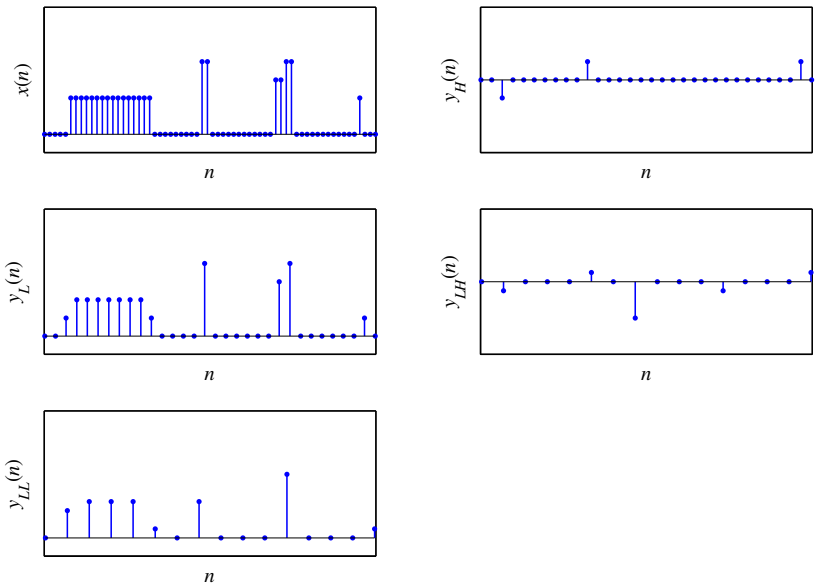


Figure 3.22 Illustration of the procedure leading to the Haar wavelet transform calculation.

Then we calculate the lowpass part of the auxiliary highpass signal as

$$H(3) = y_H(0) + y_H(1) = x(0) - x(1) + x(2) - x(3).$$

Finally the highpass part of the auxiliary highpass signal is

$$H(4) = y_H(0) - y_H(1) = x(0) - x(1) - x(2) + x(3).$$

The transformation matrix, for the case of four-sample transform, is

$$\begin{bmatrix} H(0) \\ H(1) \\ H(2) \\ H(3) \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} y_L(0) \\ y_L(1) \end{bmatrix} \\ \hline \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} y_H(0) \\ y_H(1) \end{bmatrix} \end{bmatrix}. \tag{3.39}$$

By replacing the values of $y_L(n)$ and $y_H(n)$ with signal values $x(n)$, we get the transformation equation

$$\begin{bmatrix} H(0) \\ H(1) \\ H(2) \\ H(3) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \end{bmatrix} = \mathbf{T}_4 \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \end{bmatrix}, \tag{3.40}$$

with the transformation matrix $\mathbf{T}_4$.

The next step would be in grouping the two four-sample transforms into an eight-sample-based analysis. The transformation equation in the case of eight signal samples is

$$\begin{bmatrix} H(0) \\ H(1) \\ H(2) \\ H(3) \\ H(4) \\ H(5) \\ H(6) \\ H(7) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \\ x(4) \\ x(5) \\ x(6) \\ x(7) \end{bmatrix} = \mathbf{T}_8 \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \\ x(4) \\ x(5) \\ x(6) \\ x(7) \end{bmatrix}, \tag{3.41}$$

where the transformation matrix is denoted by $\mathbf{T}_8$. Note that the inverse transform is

$$\mathbf{x} = \mathbf{T}_8^{-1}\mathbf{H} = \frac{1}{8}\mathbf{T}_8^T\mathbf{H}.$$

The four-sample transformation matrix could be written as

$$\mathbf{T}_4 = \begin{bmatrix} \mathbf{T}_2 \otimes \begin{bmatrix} 1 & 1 \end{bmatrix} \\ \mathbf{T}_2 \otimes \begin{bmatrix} 1 & -1 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \mathbf{T}_2 \otimes [\mathbf{T}_2(1,:)] \\ \mathbf{T}_2 \otimes [\mathbf{T}_2(2,:)] \end{bmatrix}$$

where $\otimes$ denotes Kronecker multiplication of two submatrices in $\mathbf{T}_2$ (its rows) with $\mathbf{T}_2$, defined by (3.36). Notation $\mathbf{T}_2(i,:)$ is used for the ith row of $\mathbf{T}_2$. The transformation matrix of order N is obtained by a Kronecker product of $N/2$-order transformation matrix rows and $\mathbf{T}_2$,

$$\mathbf{T}_N = \begin{bmatrix} \mathbf{T}_2 \otimes [\mathbf{T}_{N/2}(1,:)] \\ \mathbf{T}_2 \otimes [\mathbf{T}_{N/2}(2,:)] \\ \cdots \\ \mathbf{T}_2 \otimes [\mathbf{T}_{N/2}(N/2,:)] \end{bmatrix}. \tag{3.42}$$

In this way, although we started from a two-point DFT, in splitting the frequency domain, we did not obtain the Fourier transform of a signal, but a form of the Walsh-Hadamard transform. In ordering the coefficients (matrix rows) in our example, we followed the frequency region order from the Fourier domain (for example, in the four-sample case, low-low, low-high, high-low, and high-high frequency region).

Three ways of ordering transform coefficients in the Walsh-Hadamard transform (ordering of transformation matrix rows) are used. They produce the same result with different orderings of the coefficients and different recursive formulae for constructing the transformation matrices. The presented way of ordering coefficients, as in (3.41), is known as the *Walsh transform* with **dyadic ordering**. It will be used in examples and denoted as the Walsh-Hadamard transform.

The *Hadamard transform* would correspond to the so called **natural ordering** of rows from the transformation matrix $\mathbf{T}_8$,

$$\mathbf{H}_8 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix}$$

It would correspond to $[H(0), H(4), H(2), H(6), H(1), H(5), H(3), H(7)]^T$ order of coefficients in the Walsh transform with dyadic ordering (3.41).

Recursive construction of the Hadamard transform matrix $\mathbf{H_{2N}}$ is easy using the Kronecker product of $\mathbf{T}_2$ defined by (3.36) and $\mathbf{H_N}$,

$$\mathbf{H_{2N}} = \mathbf{T_2} \otimes \mathbf{H_N} = \begin{bmatrix} \mathbf{H_N} & \mathbf{H_N} \\ \mathbf{H_N} & -\mathbf{H_N} \end{bmatrix}.$$

The following order $[H(0), H(1), H(3), H(2), H(6), H(7), H(5), H(4)]^T$ in (3.41) would correspond to a Walsh transform with **sequence ordering**.

Calculation of the Walsh-Hadamard transforms requires only additions. For an N-order transform the number of additions is $(N-1)N$.

3.9.2 Discrete Haar Wavelet Transform

Consider again two pairs of signal samples, $x(0)$,$x(1)$ and $x(2)$,$x(3)$. The high frequency parts of these pairs are calculated as $y_H(n) = x(2n) - x(2n+1)$, for $n = 0,1$. They are used in the Haar transform without any further modification. Since they represent highpass Haar transform coefficients, they will be denoted, by $W(2) = y_H(0) = x(0) - x(1)$ and $W(3) = y_H(1) = x(2) - x(3)$. The lowpass coefficients of these pairs are $y_L(0) = x(0) + x(1)$ and $y_L(1) = x(2) + x(3)$. The highpass and lowpass parts of these signals are calculated as $y_{LH}(0) = [x(0) + x(1)] - [x(2) + x(3)]$ and $y_{LL}(0) = [x(0) + x(1)] + [x(2) + x(3)]$. For a four-sample signal the transformation ends here with $W(1) = y_{LH}(0)$ and $W(0) = y_{LL}(0)$. Note that the order of coefficients is such that the lowest frequency coefficient corresponds to the transform index $k = 0$. Matrix form of the transform for a four-sample signal is

$$\begin{bmatrix} W(0) \\ W(1) \\ W(2) \\ W(3) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \end{bmatrix}.$$

For an eight-sample signal the highpass coefficients would be kept without further modification in every step (scale), while for the lowpass parts of signal their highpass and lowpass parts would be calculated. The transformation matrix in the case of a signal with eight samples is

$$\begin{bmatrix} W(0) \\ W(1) \\ W(2) \\ W(3) \\ W(4) \\ W(5) \\ W(6) \\ W(7) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \\ x(4) \\ x(5) \\ x(6) \\ x(7) \end{bmatrix}. \tag{3.43}$$

This is the Haar transform or the Haar wavelet transform of a signal with eight samples.

The Haar transform is useful in the analysis of signals when we can expect that in a slow-varying signal there are few details.

The Haar wavelet transform is computationally very efficient. The efficiency comes from the fact that the Haar wavelet transform almost does not transform the signal at high frequencies. It leaves it almost as it is, using a very simple two-sample transform. For lower frequencies, the number of operations is increased.

In specific, for the highest $N/2$ coefficients, the Haar transform does only one addition (of two signal values) for every coefficient. For next $N/4$ coefficients the Haar wavelet uses 4 signal values

with 3 additions and so on. The total number of additions for the Haar transform is

$$N_{\text{additions}} = \frac{N}{2}(2-1) + \frac{N}{4}(4-1) + \frac{N}{8}(8-1) + \cdots + \frac{N}{N}(N-1).$$

For N of the form $N = 2^m$ we can write

$$N_{\text{additions}} = N\log_2 N - N\left(\frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots + \frac{1}{2^m}\right)$$

$$= N\log_2 N - N\frac{1}{2}\frac{1-\frac{1}{2^m}}{1-\frac{1}{2}} = N\log_2 N - (N-1) = N[\log_2 N - 1] + 1.$$

This is the same order of additions as in the FFT algorithms.

Example 3.13. Consider the signal

$$x(n) = [2,2,12,-8,2,2,2,2,-3,-3,-3,-3,3,-9,-3,-3].$$

Calculate its Haar and Walsh-Hadamard transform with $N = 16$. Discuss the results.

★Signal $x(n)$ is presented in Fig. 3.23.

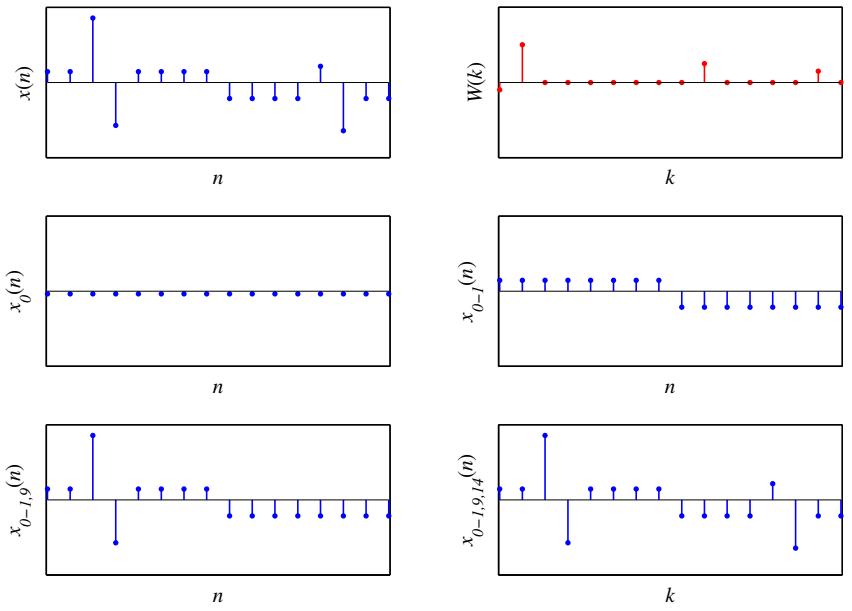


Figure 3.23 Signal $x(n)$ and its discrete Haar transform $H(k)$. Reconstructed signals: using $H(0)$ presented by $x_0(n)$, using two coefficients $H(0)$ and $H(1)$ denoted by $x_{0-1}(n)$, using $H(0)$, $H(1)$, and $H(9)$ denoted by $x_{0-1,9}(n)$, and using $H(0)$, $H(1)$, $H(9)$, and $H(14)$ denoted by $x_{0-1,9,14}(n)$. Vertical axes scales for the signal and transform are different.

In full analogy with (3.43), the Haar transformation matrix of order $N = 16$ is formed. For example, higher coefficients are just two-sample signal transforms,

$$W(k) = x(2(k-8)) - x(1+2(k-8)), \quad k = 8,9,\ldots,15.$$

Although there are some short duration pulses $(x(2), x(3), x(13))$, the Haar transform coefficients $W(2), W(3), \ldots, W(8), W(10), W(11), W(12), W(13), W(15)$ are zero-valued, Fig. 3.23. This is the result of the Haar transform property to decompose the high frequency signal region into short duration (two-sample) basis functions. A short duration pulse is contained in the high frequency part of only one Haar coefficient. This is not the case in the Fourier transform (or Walsh-Hadamard transform) where a single delta pulse will cause that all coefficients are nonzero, Fig. 3.24. Transformation matrix $\mathbf{T_{16}}$ for the Walsh-Hadamard transform is obtained from $\mathbf{T_8}$ using (3.42).

Property that high-frequency coefficients are well localized in the time domain and that they represent a short duration signal components is used in image compression, where adding high frequency coefficients adds details into an image, with important property that one detail in the image corresponds to one (a few) nonzero coefficient. Reconstruction of the signal from the Haar transform, using various number of coefficients, is presented in Fig. 3.23. As explained, it can be considered as "a zooming" a signal toward the details when the higher frequency coefficients are added. Since a half of the coefficients are zero-valued a significant compression ratio can be achieved by storing or transmitting the nonzero coefficients only. This is a basic idea for multiresolution wavelet based image representations and compression.

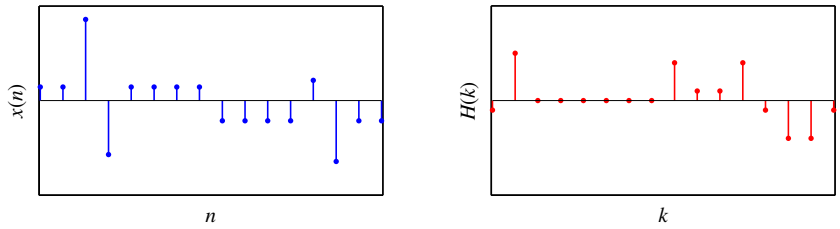


Figure 3.24 Signal $x(n)$ and its Walsh-Hadamard transform $HD(k)$.

Example 3.14. For the signals
(a) $x(n) = [1,-1,1,-1,1,-1,1,-1,1,-1,1,-1,1,-1,1,-1]$ and
(b) $x(n) = [2,0,-2,0,0,-2,0,2,0,2,0,-2,-2,0,2,0]$,
calculate the Haar wavelet transform and the Walsh-Hadamard transform with $N = 16$.

★The Haar wavelet transform and the Walsh-Hadamard transform are shown in Fig. 3.25. We can see that for a signal of long duration, with high frequencies, the number of nonzero coefficients in the Haar wavelet transform is large. Just one such a component in the Walsh-Hadamard transform can require a half of the available coefficients in the Haar wavelet transform, Fig. 3.25(left). In addition, the fact that a much smaller number of coefficients is used for the Walsh-Hadamard

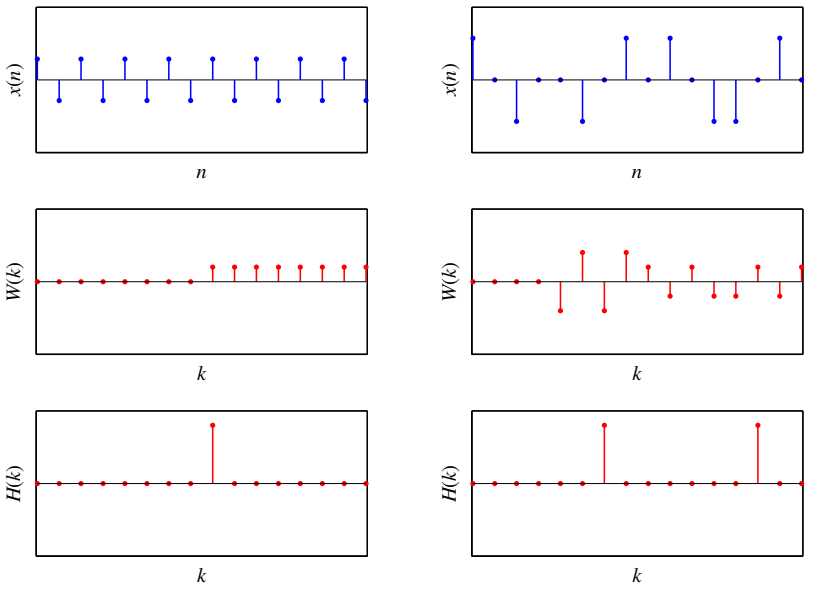


Figure 3.25 The Haar wavelet transform (second row) and the Walsh-Hadamard transform (third row) for high frequency long duration signals (first row). Vertical axes scales for the signal and transform are different.

transform based reconstruction, as compared to a very large number of coefficients in the Haar wavelet transform reconstruction, may annul the Haar transform calculation complexity advantage in this case.

3.10 PROBLEMS

Problem 3.1. Calculate the DFT of the following signals:
(a) $x(n) = \delta(n)$,
(b) $x(n) = \delta(n) + \delta(n-1) - 2j\delta(n-2) + 2j\delta(n-3) + \delta(n-4)$, and
(c) $x(n) = a^n(u(n) - u(n-10))$,
using the smallest possible value of period N.

Problem 3.2. If the signals $g(n)$ and $f(n)$ are real-valued, show that their DFTs, $G(k)$ and $F(k)$, can be obtained from the DFT $Y(k)$ of the signal $y(n)$ defined as $y(n) = g(n) + jh(n)$.

Problem 3.3. Frequency of a continuous time signal is related to the DFT index according to

$$\Omega = \begin{cases} 2\pi k/(N\Delta t) & \text{for } 0 \le k \le N/2 - 1 \\ 2\pi(k-N)/(N\Delta t) & \text{for } N/2 \le k \le N - 1. \end{cases}$$

This mapping is achieved in programs using shift functions. Show that the shift will not be necessary if we use the signal $x(n)(-1)^n$. The DFT values of this signal will be ordered from the one corresponding to the lowest negative frequency, toward the highest positive frequency.

Problem 3.4. If the DFT of a signal $x(n)$, with period N, is $X(k)$ find the DFT of the signals

$$y(n) = \begin{cases} x(n) \text{ for } & n = 2m \\ 0 \text{ for } & n = 2m+1 \end{cases}$$

and

$$z(n) = \begin{cases} 0 \text{ for } & n = 2m \\ x(n) \text{ for } & n = 2m+1. \end{cases}$$

Problem 3.5. Find the convolution of signals $x(n)$ and $h(n)$ whose nonzero values are $x(0) = 1$, $x(1) = -1$ and $h(0) = 2$, $h(1) = -1$, $h(2) = 2$, using their DFTs and the inverse DFT of the resulting product, that is $x(n) *_n h(n) = \text{IDFT}\{\text{DFT}\{x(n)\}\text{DFT}\{y(n)\}\}$.

Problem 3.6. Find the circular convolution of the signals $x(n) = e^{j4\pi n/N} + \sin(2\pi n/N)$ and $h(n) = \cos(4\pi n/N) + e^{j2\pi n/N}$ within the common period for both signals.

Problem 3.7. Find the signal whose DFT is $Y(k) = |X(k)|^2$ and $X(k)$ is the DFT of the signal $x(n) = u(n) - u(n-3)$, calculated with the period $N = 10$.

Problem 3.8. What is the relation between the discrete Hartley transform (DHT) of real-valued signals $x(n)$, defined by

$$H(k) = \sum_{n=0}^{N-1} x(n) \left(\cos \frac{2\pi nk}{N} + \sin \frac{2\pi nk}{N} \right)$$

and the DFT of the same signal? Express the DHT in terms of the DFT and the DFT in terms of the DHT.

Problem 3.9. Show that the DCT of a signal $x(n)$ with N samples, defined by

$$C(k) = \sum_{n=0}^{N-1} 2x(n) \cos(\frac{2\pi k}{2N}(n + \frac{1}{2}))$$

can be calculated using an N-sample DFT of the signal

$$y(n) = \begin{cases} 2x(2n) & \text{for } 0 \le n \le N/2 - 1 \\ 2x(2N - 2n - 1) & \text{for } N/2 \le n \le N - 1 \end{cases}$$

as

$$C(k) = \text{Re}\{e^{-j\frac{\pi k}{2N}} \sum_{n=0}^{N-1} y(n) e^{-j\frac{2\pi k}{N}n}\} = \text{Re}\{e^{-j\frac{\pi k}{2N}} \text{DFT}\{y(n)\}\}.$$

Problem 3.10. A real-valued signal $x(n)$ of a duration shorter than N, defined for $0 \le n \le N-1$, has the Fourier transform $X(k)$. The signal $y(n)$ is formed as

$$y(n) = \begin{cases} 2x(n) \text{ for } 0 \le n \le N-1 \\ 0 \quad \text{for } N \le n \le 2N-1, \end{cases} \tag{3.44}$$

with the DFT denoted by $Y(k)$. The signal $z(n)$ is formed using $y(n)$ as

$$z(2n+1) = y(n)$$
$$z(2n) = 0.$$

(a) What are the real and imaginary parts of $Z(k) = \text{DFT}\{z(n)\}$? How they are related to the DCT and the DST of the signal $x(n)$? (b) The signal $x(n)$ is applied as an input to a linear impulse invariant system with the impulse response $h(n)$ such that $h(n)$ is of the duration shorter than N,

defined within $0 \leq n \leq N-1$, and $x(n) *_n h(n)$ is also within the same interval, $0 \leq n \leq N-1$. The DCT of the output signal is calculated. How the DCT of the output signal is related to the DCT and DST of the input signal $x(n)$?

Problem 3.11. Consider a signal $x(n)$ whose duration is N, with nonzero values within the interval $0 \leq n \leq N-1$. Define the system with the output

$$y_k(n+(N-1)) = \sum_{m=0}^{N-1} x(n+m)e^{-j2\pi mk/N}$$

so that its value $y_k(N-1)$ at the last instant of the signal duration is equal to the DFT of signal, for a given k,

$$y_k(N-1) = \sum_{m=0}^{N-1} x(m)e^{-j2\pi mk/N} = \text{DFT}\{x(n)\} = X(k).$$

Note that the system is causal since $y_k(n)$ uses only $x(n)$ at instant n and previous instants.

Show that the output signal $y_k(n)$ is related to the previous output value $y_k(n-1)$ by the equation

$$y_k(n) = e^{j2\pi k/N} y_k(n-1) + e^{j2\pi k/N}[x(n) - x(n-N)].$$

This equation can be used for a recursive DFT calculation.

Problem 3.12. Show that the discrete Hartley transform (DHT) coefficients of a signal $x(n)$ with an even number of samples N can be calculated, for an even frequency index $k = 2r$, using the DHTs with $N/2$ samples (fast DHT calculation).

Problem 3.13. Find the DFT of the signal $x(n) = \exp(j4\pi\sqrt{3}n/N)$, for $n = 0,1,\ldots,N-1$ with $N = 16$. If the DFT is interpolated four times (signal zero-padded), find the displacement bin, estimate the signal frequency, and compare it with the true frequency value. What is the displacement bin if the general formula is applied without the interpolation?

3.11 EXERCISE

Exercise 3.1. Find the DFT of the signal $x(n) = \delta(n) - \delta(n-3)$ with the assumed periods $N = 4$ and $N = 8$.

Exercise 3.2. Calculate the DFT of the signal $x(n) = \sin(n\pi/4)$ for $0 \leq n < N$ with $N = 8$ and $N = 16$.

Exercise 3.3. For a real-valued signal $x(n)$, the DFT is calculated with $N = 8$ and the following DFT values are known: $X(0) = 1$, $X(2) = 2 - j$, $X(4) = 2$, $X(5) = j$, $X(7) = 3$. Find the remaining DFT values. What are the values of $x(0)$ and $\sum_{n=0}^{7} x(n)$?

Exercise 3.4. Signal $x(n)$ is presented in Fig. 3.26. Find $X(0)$, $X(4)$, and $X(8)$, where $X(k)$ is the DFT of the signal $x(n)$ calculated with the period $N = 16$.

Exercise 3.5. Prove that the DFT value $X(N/2)$ is real-valued for an arbitrary real-valued signal $x(n)$, defined for $0 \leq n < N$, where N is an even integer.

Exercise 3.6. Consider the signal $x(n)$ whose DFT values $X(k)$, calculated with $N = 16$, are presented in Fig. 3.27.

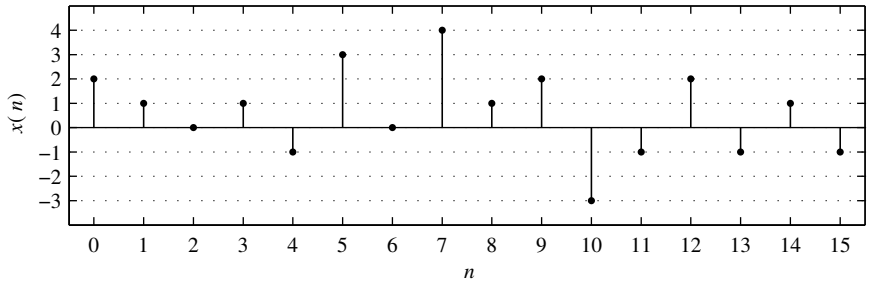


Figure 3.26 Discrete signal $x(n)$ (Exercise 3.4)

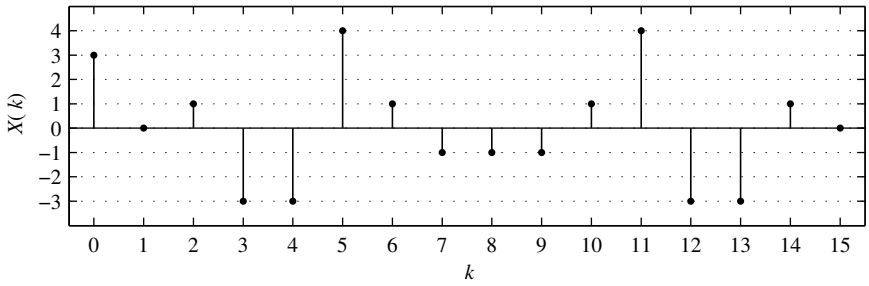


Figure 3.27 DFT of the discrete signal $x(n)$ (Exercise 3.6)

1. Find the DFT of the signal $y_1(n) = x(n) + (-1)^n x(n)$.
2. Find the DFT of the signal $y_2(n) = x(n) - (-1)^n x(n)$.
3. Find the DFT of the signal $y_3(n) = x(n) *_n x(N-n)$, where $*_n$ denotes the circular convolution with the period N.
4. Find the DFT of the signal
$$y_4(n) = \begin{cases} x(n) & \text{for } n \neq 8 \\ 0 & \text{for } n = 8. \end{cases}$$
5. Find the value $x(0)$.
6. Calculate $\sum_{n=0}^{15} x(n)$.
7. Calculate $\sum_{n=0}^{15} |x(n)|^2$.
8. Calculate $\sum_{n=0}^{15} (-1)^n x(n)$.

Exercise 3.7. Prove that if $|x(n)| \leq A$ for $0 \leq n < N$ then $|X(k)| \leq NA$ for any k, where $X(k)$ is the DFT of $x(n)$ calculated with N points.

Exercise 3.8. Prove that if $\sum_{n=0}^{N-1} |x(n)| \leq A$ then $\sum_{k=0}^{N-1} |X(k)| \leq NA$ where $X(k)$ is the DFT of $x(n)$ calculated with N samples.

3.12 SOLUTIONS

Solution 3.1. The DFT assumes that the signals are periodic. In order to calculate the DFT, we have to assume a period of the considered aperiodic signals first. The period N should be greater or equal to the signal duration, so that the signal values do not overlap after their periodic extension. Larger values of N will increase the density of the frequency domain samples, but they will also increase the computation time.

a) For this signal any $N \geq 1$ is acceptable, producing

$$X(k) = 1, \quad k = 0,1,\ldots,N-1,$$

with the assumed period N.

b) For this signal, we may use any $N \geq 5$. With $N = 5$, we get

$$\begin{aligned} X(k) &= \sum_{n=0}^{5-1} x(n)e^{-j2\pi nk/5} = 1 + e^{-j2\pi k/5} - 2je^{-j4\pi k/5} + j2e^{-j6\pi k/5} + e^{-j8\pi k/5} \\ &= 1 + 2\cos(2\pi k/5) - 4\sin(4\pi k/5). \end{aligned}$$

c) For the period $N \geq 10$ the DFT is given by

$$X(k) = \sum_{n=0}^{9} (ae^{-j2\pi k/N})^n = \frac{1 - a^{10}e^{-j20\pi k/N}}{1 - ae^{-j2\pi k/N}}.$$

Solution 3.2. From the signal $y(n) = g(n) + jf(n)$ its real and imaginary parts $g(n)$ and $f(n)$ can be obtained as

$$g(n) = \frac{y(n) + y^*(n)}{2}, \text{ and } f(n) = \frac{y(n) - y^*(n)}{2j}.$$

Since the DFT of $y^*(n)$ is equal to

$$\text{DFT}\{y^*(n)\} = \sum_{n=0}^{N-1} y^*(n)e^{-j2\pi nk/N} = \left(\sum_{n=0}^{N-1} y(n)e^{j2\pi nk/N}\right)^*$$

with $e^{j2\pi nk/N} = e^{j2\pi n(k-N)/N} = e^{-j2\pi n(N-k)/N}$, it follows

$$\text{DFT}\{y^*(n)\} = Y^*(N-k).$$

Then the DFTs of signals $g(n)$ and $f(n)$ are obtained from

$$G(k) = \frac{Y(k) + Y^*(N-k)}{2} \text{ and } F(k) = \frac{Y(k) - Y^*(N-k)}{2j}.$$

Solution 3.3. The DFT of the signal formed as $x(n)(-1)^n$ is given by

$$X_1(k) = \sum_{n=0}^{N-1} x(n)(-1)^n e^{-j2\pi nk/N}.$$

For the range of frequency indices within $0 \le k \le N/2-1$, we can write

$$X_1(k) = \sum_{n=0}^{N-1} x(n)e^{-j\pi n}e^{-j2\pi nk/N} = \sum_{n=0}^{N-1} x(n)e^{-j2\pi n(k+N/2)/N} = X(k+\frac{N}{2}).$$

For $N/2 \le k \le N-1$ holds,

$$X_1(k) = \sum_{n=0}^{N-1} x(n)e^{j\pi n}e^{-j2\pi nk/N} = \sum_{n=0}^{N-1} x(n)e^{-j2\pi n(k-N/2)/N} = X(k-\frac{N}{2}).$$

Solution 3.4. The DFT of the signal $y(n)$ is given by

$$\begin{aligned} Y(k) &= \sum_{n=0}^{N-1} y(n)e^{-j2\pi nk/N} = \frac{1}{2}\sum_{n=0}^{N-1}[x(n)+(-1)^n x(n)]e^{-j2\pi nk/N} \\ &= \frac{1}{2}\sum_{n=0}^{N-1}[x(n)+x(n)e^{-j\pi nN/N}]e^{-j2\pi nk/N} = \frac{1}{2}[X(k)+X(k+\frac{N}{2})] \end{aligned}$$

with $X(k+N/2) = X(k-N/2)$ for $k > N/2$.

For the signal $z(n)$ its DFT can be written in the form

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

It is obvious that the DFT of signal $x(n)$ is equal to the sum of the DFTs of signals $y(n)$ and $z(n)$,

$$Y(k)+Z(k) = X(k).$$

Solution 3.5. For the convolution calculation using the DFTs of signals, the minimum number for the period N is $N = K+L-1 = 4$, where $K = 2$ is the duration of the signal $x(n)$ and $L = 3$ is the duration of the impulse response $h(n)$. With $N = 4$, we get

$$\begin{aligned} X(k) &= 1-e^{-j2\pi k/4} \\ H(k) &= 2-e^{-j2\pi k/4}+2e^{-j4\pi k/4} \\ Y(k) &= X(k)H(k) = 2-3e^{-j2\pi k/4}+3e^{-j4\pi k/4}-2e^{-j6\pi k/4}. \end{aligned}$$

The signal is equal to the inverse DFT of $Y(k) = X(k)H(k)$,

$$y(n) = \text{IDFT}\{Y(k)\} = 2\delta(n)-3\delta(n-1)+3\delta(n-2)-2\delta(n-3).$$

Solution 3.6. The DFT of the circular convolution $y(n)$ of signals $x(n)$ and $h(n)$, $y(n) = x(n)*h(n)$, is equal to the corresponding DFTs: $Y(k) = X(k)H(k)$ with

$$\begin{aligned} X(k) &= \sum_{n=0}^{N-1}[e^{j4\pi n/N}+\frac{1}{2j}e^{j2\pi n/N}-\frac{1}{2j}e^{-j2\pi n/N}]e^{-j2\pi nk/N} \\ &= N\delta(k-2)+\frac{N}{2j}\delta(k-1)-\frac{N}{2j}\delta(k+1) \end{aligned}$$

and

$$H(k) = \sum_{n=0}^{N-1} [\frac{1}{2} e^{j4\pi n/N} + \frac{1}{2} e^{-j4\pi n/N} + e^{j2\pi n/N}] e^{-j2\pi nk/N}$$
$$= \frac{N}{2}\delta(k-2) + \frac{N}{2}\delta(k+2) + N\delta(k-1).$$

The value of $Y(k)$ is

$$Y(k) = \frac{N^2}{2}\delta(k-2) + \frac{N^2}{2j}\delta(k-1).$$

The circular convolution is obtained as the inverse DFT of $Y(k)$

$$y(n) = \frac{N}{2} e^{j4\pi n/N} + \frac{N}{2j} e^{j2\pi n/N}.$$

Solution 3.7. The DFT of signal $y(n)$, equal to $Y(k) = |X(k)|^2$, can be written as $Y(k) = X(k)X^*(k)$. The inverse DFT of this product is equal to the convolution of individual inverse DFTs, that is

$$y(n) = \text{IDFT}\{X(k)\} *_n \text{IDFT}\{X^*(k)\}.$$

Since

$$\text{IDFT}\{X^*(k)\} = \frac{1}{N}\sum_{k=0}^{N-1} X^*(k) e^{j2\pi nk/N} = \left(\frac{1}{N}\sum_{k=0}^{N-1} X(k) e^{-j2\pi nk/N}\right)^*$$
$$= \left(\frac{1}{N}\sum_{k=0}^{N-1} X(k) e^{j2\pi k(N-n)/N}\right)^* = x^*(N-n)$$

we get

$$y(n) = (x(n))_{10} *_n (x^*(10-n))_{10} = (u(n) - u(n-3))_{10} *_n (u(10-n) - u(7-n))_{10}$$
$$= (\delta(n+2) + 2\delta(n+1) + 3\delta(n) + 2\delta(n-1) + \delta(n-2))_{10},$$

where $(x(n))_N$ indicates that the signal $x(n)$ is periodically extended with period N.

Solution 3.8. For a real-valued signals we can write

$$X(k) = \sum_{n=0}^{N-1} [x(n)\cos\frac{2\pi nk}{N} - jx(n)\sin\frac{2\pi nk}{N}]$$
$$X(N-k) = \sum_{n=0}^{N-1} [x(n)\cos\frac{2\pi nk}{N} + jx(n)\sin\frac{2\pi nk}{N}].$$

From the previous equations, we can easily conclude that the following relations hold

$$\sum_{n=0}^{N-1} x(n)\cos\frac{2\pi nk}{N} = \frac{X(k) + X(N-k)}{2} = \frac{H(k) + H(N-k)}{2}$$
$$\sum_{n=0}^{N-1} x(n)\sin\frac{2\pi nk}{N} = \frac{X(N-k) - X(k)}{2j} = \frac{H(k) - H(N-k)}{2}.$$

The DHT can be calculated as a sum of these two terms, that is

$$2H(k) = X(k) + X(N-k) - j[X(N-k) - X(k)].$$

The DFT is obtained using the DHT in the same way as

$$2X(k) = H(k) + H(N-k) - j[H(k) - H(N-k)].$$

Solution 3.9. We can split the summation in the DCT into an even and odd part

$$C(k) = \sum_{n=0}^{N-1} 2x(n)\cos(\frac{2\pi k}{2N}(n+\frac{1}{2})) =$$
$$\sum_{n=0}^{N/2-1} 2x(2n)\cos(\frac{2\pi k}{2N}(2n+\frac{1}{2})) + \sum_{n=0}^{N/2-1} 2x(2n+1)\cos(\frac{2\pi k}{2N}(2n+1+\frac{1}{2})).$$

By reverting the summation index in the second sum, using $n = N/2-1-m$, the summation over m is from $m = N/2-1$, for $n = 0$, down to $m = 0$, for $n = N/2-1$. Then

$$\sum_{n=0}^{N/2-1} 2x(2n+1)\cos(\frac{2\pi k}{2N}(2n+1+\frac{1}{2}))$$
$$= \sum_{m=0}^{N/2-1} 2x(N-2m-1)\cos(\frac{2\pi k}{2N}(N-2m-1+\frac{1}{2})).$$

The summation index in this sum can be shifted for $N/2+m = n$ to get

$$\sum_{m=0}^{N/2-1} 2x(N-2m-1)\cos(\frac{2\pi k}{2N}(N-2m-1+\frac{1}{2}))$$
$$= \sum_{n=N/2}^{N-1} 2x(2N-2n-1)\cos(\frac{2\pi k}{2N}(2N-2n-\frac{1}{2})).$$

Now we can go back to the DCT and replace the second sum, to get

$$C(k) = \sum_{n=0}^{N/2-1} 2x(2n)\cos(\frac{2\pi k}{2N}(2n+\frac{1}{2}))$$
$$+ \sum_{n=N/2}^{N-1} 2x(2N-2n-1)\cos(\frac{2\pi k}{2N}(2n+\frac{1}{2})) = \sum_{n=0}^{N-1} y(n)\cos(\frac{2\pi k}{2N}(2n+\frac{1}{2}))$$

with $\cos(\frac{2\pi k}{2N}(2N-2n-\frac{1}{2})) = \cos(\frac{2\pi k}{2N}(2n+\frac{1}{2}))$ and

$$y(n) = \begin{cases} 2x(2n) & \text{for } 0 \le n \le N/2-1 \\ 2x(2N-2n-1) & \text{for } N/2 \le n \le N-1. \end{cases}$$

We can conclude that the relation between the DFT and DCT is given by

$$C(k) = \mathrm{Re}\{\sum_{n=0}^{N-1} y(n)e^{-j\frac{2\pi k}{2N}(2n+\frac{1}{2})}\} = \mathrm{Re}\{e^{-j\frac{\pi k}{2N}}\mathrm{DFT}\{y(n)\}\}.$$

Solution 3.10. (a) For the signal $z(n)$ we can write

$$\begin{aligned}\mathrm{DFT}\{z(n)\} &= \sum_{n=0}^{4N-1} z(n)e^{-j2\pi nk/(4N)} = \sum_{n=0}^{2N-1} z(2n+1)e^{-j2\pi(2n+1)k/(4N)} \\ &= \sum_{n=0}^{2N-1} y(n)e^{-j2\pi(2n+1)k/(4N)} = \sum_{n=0}^{N-1} 2x(n)e^{-j2\pi(2n+1)k/(4N)}.\end{aligned}$$

The real and imaginary parts of $\mathrm{DFT}\{z(n)\}$ are equal to

$$\begin{aligned}\mathrm{Re}\{\mathrm{DFT}\{z(n)\}\} &= \sum_{n=0}^{N-1} 2x(n)\cos(\frac{2\pi(2n+1)k}{4N}) = C(k) \\ \mathrm{Im}\{\mathrm{DFT}\{z(n)\}\} &= -\sum_{n=0}^{N-1} 2x(n)\sin(\frac{2\pi(2n+1)k}{4N}) = -S(k) \\ \mathrm{DFT}\{z(n)\} &= C(k) - jS(k),\end{aligned}$$

and

$$\begin{aligned}Z(k) = \mathrm{DFT}\{z(n)\} &= e^{-j2\pi k/(4N)} \sum_{n=0}^{N-1} 2x(n)e^{-j2\pi nk/(2N)} \\ Z(k)e^{j\pi k/(2N)} &= Y(k) = 2X(k/2).\end{aligned}$$

Note that $X(k/2)$ is just a notation for $2X(\frac{k}{2}) = Y(k)$, where $Y(k) = \mathrm{DFT}\{y(n)\}$ and $y(n)$ is zero-padded version of $2x(n)$, defined by (3.44).

(b) If the signal $x(n)$ is used as an input to a system then the DCT is calculated for

$$\begin{aligned}x_h(n) &= x(n) *_n h(n) \\ X_h(k) &= X(k)H(k).\end{aligned}$$

It has been assumed that all signals, $x(n)$, $h(n)$, and $x(n) *_n h(n)$, are zero-valued outside the interval $0 \le n \le N-1$ (it means that the duration of $x(n)$ and $h(n)$ should be such that their convolution is within $0 \le n \le N-1$). Then, for the signal $z_h(n)$, related to $x_h(n) = x(n) *_n h(n)$ in the same way as the signal $z(n)$ is related to $x(n)$ in (a), we can write

$$\mathrm{DFT}\{z_h(n)\}e^{j\pi k/(2N)} = 2X_h(\frac{k}{2}) = 2X(\frac{k}{2})H(\frac{k}{2}) = Y(k)H(\frac{k}{2}).$$

Then

$$\begin{aligned}C_h(k) = \mathrm{DCT}\{x_h(n)\} &= \mathrm{Re}\{Y(k)H(\frac{k}{2})e^{-j\pi k/(2N)}\} \\ = \mathrm{Re}\{Y(k)e^{-j\pi k/(4N)}\}\,\mathrm{Re}\{H(\frac{k}{2})\} &- \mathrm{Im}\{Y(k)e^{-j\pi k/(4N)}\}\,\mathrm{Im}\{H(\frac{k}{2})\} \\ &= C(k)\,\mathrm{Re}\{H(\frac{k}{2})\} + S(k)\,\mathrm{Im}\{H(\frac{k}{2})\}.\end{aligned}$$

The system output is given by

$$x(n) *_n h(n) = x_h(n) = \mathrm{IDCT}\{C_h)k)\}$$

, with $\mathrm{IDCT}\{C_h)k)\}$ defined by (3.34). The transform $H(k/2)$ is the DFT of the zero-padded signal $h(n)$ with a factor of 2. Only the first half of the DFT samples are then used in calculation.

Solution 3.11. For the signal $y_k(n)$ we may write

$$y_k(n) = \sum_{m=0}^{N-1} x(n-N+1+m)e^{-j2\pi mk/N}.$$

Now let us shift the summation for one sample

$$y_k(n) = \sum_{m=1}^{N} x(n-N+m)e^{-j2\pi(m-1)k/N} = e^{j\frac{2\pi}{N}k}\sum_{m=1}^{N} x(n-N+m)e^{-j2\pi mk/N}$$
$$= e^{j\frac{2\pi}{N}k}[\sum_{m=0}^{N-1} x(n-N+m)e^{-j2\pi mk/N} - x(n-N)e^{-j2\pi 0k/N} + x(n)e^{-j2\pi Nk/N}]$$
$$= e^{j2\pi k/N}[y_k(n-1) - x(n-N) + x(n)].$$

Within the interval $0 \le n \le N-1$ holds

$$y_k(n) = e^{j2\pi k/N}[y_k(n-1) + x(n)],$$

since $x(n-N) = 0$. This proves the problem statement.

If the signal $x(n)$ continues as a periodic signal after $n = 0$, then

$$x_p(n) = \sum_{l=0}^{\infty} x(n-lN) \tag{3.45}$$

and for $n \ge N$, holds $x_p(n-N) = x_p(n)$ and $y_k(n) = e^{j2\pi k/N}y_k(n-1)$,

$$y_k(n) = \begin{cases} 0 & \text{for } n < 0 \\ y_k(n) = e^{j2\pi k/N}[y_k(n-1) + x(n)] & \text{for } 0 \le n \le N-1 \\ y_k(n) = e^{j2\pi k/N}y_k(n-1) & \text{for } n \ge N \end{cases}$$

for $x_p(n)$ defined by (3.45).

Solution 3.12. For $k = 2r$ the DHT can be written as

$$H(2r) = \sum_{n=0}^{N/2-1} x(n)\left[\cos\tfrac{2\pi rn}{N/2} + \sin\tfrac{2\pi rn}{N/2}\right] + \sum_{n=N/2}^{N-1} x(n)\left[\cos\tfrac{2\pi rn}{N/2} + \sin\tfrac{2\pi rn}{N/2}\right]$$
$$= \sum_{n=0}^{N/2-1} (x(n) + x(n+N/2))\left[\cos\tfrac{2\pi rn}{N/2} + \sin\tfrac{2\pi rn}{N/2}\right].$$

Therefore, the value of $H(2r)$ is equal to

$$H(2r) = \sum_{n=0}^{N/2-1} g(n)\left[\cos\frac{2\pi rn}{N/2} + \sin\frac{2\pi rn}{N/2}\right],$$

where $g(n) = x(n) + x(n+N/2)$. This is the DHT of $g(n)$ with $N/2$ samples.

Note: For odd frequency indices $k = 2r+1$ we can write

$$H(2r+1) = \sum_{n=0}^{N-1} x(n)\left[\cos\frac{2\pi(2r+1)n}{N} + \sin\frac{2\pi(2r+1)n}{N}\right].$$

After some lengthy, but straightforward transformations, we get

$$H(2r+1)=\sum_{n=0}^{N/2-1} f(n)\left[\cos\frac{2\pi nr}{N/2}+\sin\frac{2\pi nr}{N/2}\right],$$

where

$$f(n)=[x(n)-x(n+\frac{N}{2})]\cos\frac{2\pi n}{N}+[x(\frac{N}{2}-n)-x(N-n)]\sin\frac{2\pi n}{N}.$$

This is again the DHT of the signal $f(n)$ with $N/2$ samples. In this way, the DHT of the signal with N samples is split into the two DHTs with $N/2$ samples.

Solution 3.13. The absolute value of the DFT of the signal $x(n)=\exp(j4\pi\sqrt{3}n/N)$, for $n=0,1,\ldots,N-1$ with $N=16$ is defined by

$$|X(k)|=|\sum_{n=0}^{15} e^{j2\pi(2\sqrt{3}-k)n/16}|=\left|\frac{\sin\left(\pi(2\sqrt{3}-k)\right)}{\sin\left(\pi(2\sqrt{3}-k)/16\right)}\right|,\quad\text{with}\tag{3.46}$$

$$|\mathbf{X}|=(1.5799,\ 2.1361,\ 3.5045,\ 10.9192,\ 9.4607,\ 3.3454,$$
$$2.0805,\ 1.5530,\ 1.2781,\ 1.1225,\ 1.0362,\ 0.99781,\ 0.9992,\ 1.0406,\ 1.1310,\ 1.2929),$$

where $|\mathbf{X}|$ is the vector whose elements are the DFT values $|X(k)|$, $k=0,1,\ldots,15$. The maximum DFT absolute value is achieved at $k=3$. This means that the frequency estimation, without displacement bin, would be

$$(2\pi\cdot 3)/16=1.1781,$$

while the true frequency is $(2\pi\cdot 2\sqrt{3})/16=1.3603$. The error is 13.4%.

For the zero-padded signal (interpolated DFT), with a factor of 4,

$$|X(k)|=|\sum_{n=0}^{15} e^{j4\pi\sqrt{3}n/16}e^{-j2\pi nk/64}|=|\sum_{n=0}^{15} e^{j2\pi(8\sqrt{3}-k)n/64}|$$
$$=\left|\frac{\sin\left(\pi(8\sqrt{3}-k)/4\right)}{\sin\left(\pi(8\sqrt{3}-k)/64\right)}\right|.$$

The maximum value is obtained for $k=\left[8\sqrt{3}\right]=14$, where $\left[8\sqrt{3}\right]$ denotes the nearest integer value. The maximum absolute DFT value at $k=14$, along with the absolute values of its neighbors, is

$$|X(14)|=\left|\frac{\sin\left(\pi(8\sqrt{3}-14)/4\right)}{\sin\left(\pi(8\sqrt{3}-14)/64\right)}\right|=15.9662,$$

$$|X(15)|=\left|\frac{\sin\left(\pi(8\sqrt{3}-15)/4\right)}{\sin\left(\pi(8\sqrt{3}-15)/64\right)}\right|=13.9412$$

$$|X(13)|=\left|\frac{\sin\left(\pi(8\sqrt{3}-13)/4\right)}{\sin\left(\pi(8\sqrt{3}-13)/64\right)}\right|=14.8249.$$

The displacement bin is equal to

$$d = 0.5\frac{|X(15)| - |X(13)|}{2|X(14)| - |X(15)| - |X(15)|} = -0.1395.$$

The true frequency index would be $8\sqrt{3} = 13.8564$, with the true frequency $2\pi \cdot 13.8564/64 = 1.3603$. The correct value of frequency index is shifted from the nearest integer $k = 14$ (on the frequency grid) for $14 - 13.8564 = -0.1436$, when the interpolation is done. Thus, the obtained displacement bin value -0.1395 is close to the true shift value -0.1436. The estimated frequency, using the displacement bin, is 1.3608. As compared to the true frequency, the error is 0.03%.

If the displacement formula is applied on the DFT values, without interpolation, we would get $d = 0.3356$, while $2\sqrt{3} = 3.4641$ is displaced from the nearest integer for 0.4641.

Chapter 4

z-Transform

THE Fourier transform of discrete-time 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-time 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

$$X(z) = \sum_{n=-\infty}^{\infty} x(n) z^{-n}, \tag{4.1}$$

where $z = r\exp(j\omega)$ is a complex number. The value of the z-transform along the unit circle, $|z| = 1$ or $z = \exp(j\omega)$, is equal to the Fourier transform $X(e^{j\omega})$ of discrete-time signals.

The z-transform, in general, converges only for some values of the complex argument z. The region 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

$$\begin{aligned} X(z) &= \sum_{n=0}^{\infty} (a^n z^{-n} + b^n z^{-n}) = \sum_{n=0}^{\infty} (a/z)^n + \sum_{n=0}^{\infty} (b/z)^n \\ &= \frac{1}{1 - a/z} + \frac{1}{1 - b/z} = \frac{z}{z-a} + \frac{z}{z-b}. \end{aligned}$$

Infinite geometric series, with progression coefficient (a/z), converges for $|a/z| < 1$. Thus, the region of convergence for the first part of the z-transform is $|z| > |a|$. The other series converges for $|b/z| < 1$, that is, for $|z| > |b|$. The resulting transform is finite if both terms are finite (or do

not cancel out to produce a finite value). Since $|a| < |b|$, the region of convergence for $X(z)$ is $|z| > |b|$, as shown in Fig. 4.1.

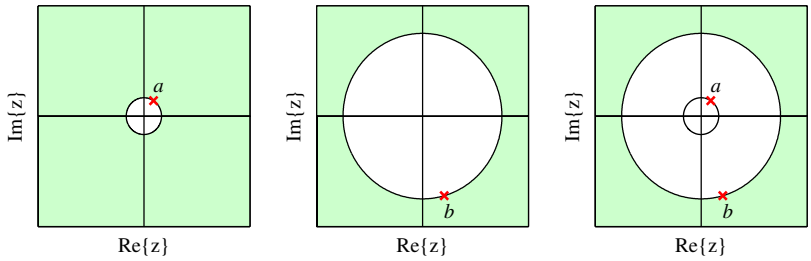


Figure 4.1 Regions of convergence (gray area) for the signal $x(n) = a^n u(n) + b^n u(n)$.

Example 4.2. Consider a discrete-time signal

$$x(n) = a^n u(n-1) - b^n u(-n-1),$$

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

$$X(z) = \sum_{n=1}^{\infty} a^n z^{-n} - \sum_{n=-\infty}^{-1} b^n z^{-n} = \sum_{n=1}^{\infty} a^n z^{-n} - \sum_{n=1}^{\infty} b^{-n} z^n$$
$$= \frac{a/z}{1-a/z} - \frac{z/b}{1-z/b} = \frac{a}{z-a} + \frac{z}{z-b}.$$

Infinite geometric series with progression coefficient (a/z) converges for $|a/z| < 1$. The other series converges for $|z/b| < 1$. 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*. For the signal $b^n u(n)$ the region of convergence was $|b/z| < 1$, while for $-b^n u(-n-1)$ the region of convergence was $|z/b| < 1$.

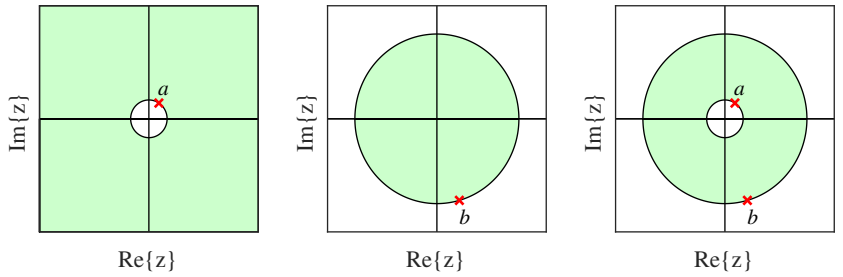


Figure 4.2 Regions of convergence (gray area) for the signal $x(n) = a^n u(n-1) - b^n u(-n-1)$.

4.2 PROPERTIES OF THE z-TRANSFORM

4.2.1 Linearity

The z-transform is linear since

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

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 the transforms.

4.2.2 Time-Shift

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

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

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. Find the z-transform domain form of the equation

$$x(n) - \frac{1}{2}x(n-1) = y(n).$$

★ The z-transform of this equation is obtained using the linearity and the shift property

$$\mathcal{Z}\{x(n) - 0.5x(n-1)\} = \mathcal{Z}\{y(n)\} \quad \text{or}$$
$$X(z) - 0.5X(z)z^{-1} = Y(z).$$

Example 4.4. For a causal signal $x(n) = x(n)u(n)$, find the z-transform of $x(n+n_0)u(n)$, for $n_0 \geq 0$.

★ The signal $x(n+n_0)u(n)$ has the z-transform

$$\mathcal{Z}\{x(n+n_0)u(n)\} = \sum_{n=0}^{\infty} x(n+n_0)z^{-n} = \sum_{n=0}^{\infty} x(n+n_0)z^{-(n+n_0)}z^{n_0}$$
$$= z^{n_0}\left[\sum_{n=0}^{\infty} x(n)z^{-n} - x(0) - x(1)z^{-1} - \cdots - x(n_0-1)z^{-n_0+1}\right]$$
$$= z^{n_0}\left[X(z) - x(0) - x(1)z^{-1} - \cdots - x(n_0-1)z^{-n_0+1}\right].$$

For $n_0 = 1$ follows

$$\mathcal{Z}\{x(n+1)u(n)\} = zX(z) - x(0). \tag{4.2}$$

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 an exponential signal: Modulation

For a signal $x(n)$ multiplied by an exponential signal a^n the z-transform is

$$\mathcal{Z}\{a^n x(n)\} = \sum_{n=-\infty}^{\infty} x(n)a^n z^{-n} = X(\frac{z}{a}),$$

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

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

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

4.2.4 Differentiation

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

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

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

$$\mathcal{Z}\{n(n+1)\ldots(n+N-1)x(n)u(n)\} = (-1)^N z^N \frac{d^N X(z)}{dz^N}.$$

4.2.5 Convolution in time

The z-transform of the convolution of signals $x(n)$ and $h(n)$ is

$$\mathcal{Z}\{x(n) *_n h(n)\} = \mathcal{Z}\{\sum_{m=-\infty}^{\infty} x(m)h(n-m)\}$$
$$= \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} x(m)h(n-m)z^{-n} = \sum_{l=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} x(m)h(l)z^{-m-l} = X(z)H(z),$$

with the region of convergence being at least the intersection of the regions of convergence of $X(z)$ and $H(z)$. 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.

Example 4.5. Find the z-transform of the signal $\sum_{m=-\infty}^{n} x(n)$.

★ This signal can be written as the convolution of signal $x(n$ and $u(n)$, that is

$$\sum_{m=-\infty}^{n} x(n) = x(n) *_n u(n) = \sum_{m=-\infty}^{n} x(m)u(n-m)$$

The z-transform of the convolution of two signal is equal to the product of their corresponding z-transforms,

$$\mathcal{Z}\{x(n) *_n u(n)\} = X(z)\frac{z}{z-1}. \tag{4.3}$$

4.2.6 Initial and Stationary State Signal Value

The initial value of a causal signal may be calculated as

$$x(0) = \lim_{z\to\infty} X(z). \tag{4.4}$$

According to the z-transform definition all terms with z^{-n} vanishes as $z \to \infty$. The term which does not depend on z is obtained as the result of this limit. It is equal to $x(0)$.

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

$$\lim_{n\to\infty} x(n) = \lim_{z\to 1}(z-1)X(z). \tag{4.5}$$

This relation follows from

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

where (4.2) is used. By definition of the z-transform we get

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

Thus,

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

produces the stationary state value (4.5).

4.2.7 Table of the z-transform

Signal $x(n)$	z-transform $X(z)$
$\delta(n)$	1
$u(n)$	$\frac{z}{z-1}$, $\|z\| > \|1\|$
$a^n u(n)$	$\frac{z}{z-a}$, $\|z\| > \|a\|$
$na^{n-1}u(n)$	$\frac{z}{(z-a)^2}$, $\|z\| > \|a\|$
$-a^n u(-n-1)$	$\frac{z}{z-a}$, $\|z\| < \|a\|$
$a^n x(n)$	$X(z/a)$
$a^{\|n\|}$, $\|a\| < 1$	$\frac{z(1-a^2)}{(z-a)(1-az)}$, $\|a\| < \|z\| < \|1/a\|$
$x(n-n_0)$	$z^{-n_0}X(z)$
$nx(n)u(n)$	$-zdX(z)/dz$
$n(n-1)x(n)u(n)$	$z^2d^2X(z)/dz^2$
$\cos(\omega_0 n)u(n)$	$\frac{1-z^{-1}\cos(\omega_0)}{1-2z^{-1}\cos(\omega_0)+z^{-2}}$
$\sin(\omega_0 n)u(n)$	$\frac{z^{-1}\sin(\omega_0)}{1-2z^{-1}\cos(\omega_0)+z^{-2}}$
$\frac{1}{n!}u(n)$	$\exp(1/z)$
$[x(n)u(n)] *_n u(n) = \sum_{m=0}^{n} x(m)$	$\frac{z}{z-1}X(z)$

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 terms z^{-1}, within the region of convergence. After the z-transform is expanded into such a 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

$$\frac{1}{1-q} = 1 + q + q^2 + \cdots = \sum_{n=0}^{\infty} q^n$$

for $|q| < 1$.

Example 4.6. For the z-transform

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

identify possible regions of convergence and find the inverse z-transform for each of them.

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

(1) For the region of convergence $|z| > 1/2$, the term $\frac{1}{2}z^{-1}$ satisfies the condition that $|\frac{1}{2}z^{-1}| < 1$. It can be expanded into geometric series as

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

However, the term $3z$ does not satisfy the condition $|3z| < 1$ for $|z| > 1/2$. This part of $X(z)$ should be modified so that it can also be expanded into geometric series for $|z| > 1/2$. This is achieved if $X(z)$ is rewritten as

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

Now, the second part of $X(z)$ can be expanded using the following geometric series

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

Both of these sums converge for $|z| > 1/2$. The resulting power series expansion of $X(z)$ is

$$\begin{aligned} X(z) &= \sum_{n=0}^{\infty} \frac{1}{2^n} z^{-n} - \frac{1}{3z} \sum_{n=0}^{\infty} \frac{1}{3^n} z^{-n} \\ &= \sum_{n=0}^{\infty} \frac{1}{2^n} z^{-n} - \sum_{n=1}^{\infty} \frac{1}{3^n} z^{-n}. \end{aligned}$$

The inverse z-transform, for this region of convergence $|z| > 1/2$, is

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

(2) For the region of convergence defined by $1/3 < |z| < 1/2$, the z-transform should be written in the form

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

The corresponding geometric series are

$$\frac{1}{1-2z} = \sum_{n=0}^{\infty} (2z)^n = \sum_{n=-\infty}^{0} 2^{-n} z^{-n} \text{ for } |2z| < 1 \text{ or } |z| < \frac{1}{2}$$
$$\frac{1}{1-\frac{1}{3z}} = \sum_{n=0}^{\infty} \left(\frac{1}{3z}\right)^n = \sum_{n=0}^{\infty} \frac{1}{3^n} z^{-n} \text{ for } \left|\frac{1}{3z}\right| < 1 \text{ or } |z| > \frac{1}{3}.$$

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

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

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) Finally, for the region of convergence $|z| < 1/3$ we can write

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

The corresponding geometric series are

$$\frac{1}{1-2z} = \sum_{n=0}^{\infty} (2z)^n = \sum_{n=-\infty}^{0} 2^{-n} z^{-n} \text{ for } |2z| < 1 \text{ or } |z| < \frac{1}{2}$$
$$\frac{1}{1-3z} = \sum_{n=0}^{\infty} (3z)^n = \sum_{n=-\infty}^{0} 3^{-n} z^{-n} \text{ for } |3z| < 1 \text{ or } |z| < \frac{1}{3}.$$

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

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

The inverse z-transform, in this case, is

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

Example 4.7. 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

$$X(z) = e^{a/z} = 1 + (a/z) + \frac{1}{2!}(a/z)^2 + \frac{1}{3!}(a/z)^3 + \dots$$

follows

$$x(n) = \delta(n) + a\delta(n-1) + \frac{1}{2!}a^2\delta(n-2) + \frac{1}{3!}a^3\delta(n-3) + \dots = a^n\frac{1}{n!}u(n).$$

The series converges for any z except $z = 0$.

Example 4.8. 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 $X(z)$ could 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

$$(z^2+1) = A(z-\frac{1}{4}) + B(z-\frac{1}{2})(z-\frac{1}{4}) + C(z-\frac{1}{2})^2. \tag{4.6}$$

For $z = 1/4$ we get $17/16 = C/16$ or $C = 17$. Using the value $z = 1/2$ 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 (4.6) 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 for the 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

$$\frac{1}{(z-a)^{m+1}} = \frac{1}{m!}\frac{d^m}{da^m}\left(\frac{1}{z-a}\right) =$$

$$= \frac{1}{m!}\frac{d^m}{da^m}\left(\sum_{n=1}^{\infty} a^{n-1}z^{-n}\right) = \frac{(n-1)(n-2)..(n-m)}{m!}\sum_{n=1}^{\infty} a^{n-m-1}z^{-n}$$

produces the inverse z-transform

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

$$= \frac{(n-1)(n-2)..(n-m)}{m!}a^{n-m-1}u(n-m-1)$$

$$= \left(\begin{array}{c} n \\ m \end{array}\right)a^{n-m-1}u(n-m-1).$$

4.3.2 Theorem of Residues Based Inversion

In general, the inversion is calculated 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 using the theorem of residues

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

where z_i are the poles of $z^{n-1}X(z)$ within the integration contour C, which is within 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,

$$x(n) = \sum_{z_i}\left\{[z^{n-1}X(z)(z-z_i)]_{|z=z_i}\right\}. \tag{4.7}$$

Example 4.9. For the z-transform

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

find the signal $x(n)$ for $n \geq 0$ if the region of convergence is $|z| > 1/2$.

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

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

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

$$= \frac{\frac{1}{2^{n-1}}4}{\frac{1}{4}} + \frac{\frac{1}{4^{n-1}}\frac{7}{2}}{\frac{-1}{4}} = 16\frac{1}{2^{n-1}} - 14\frac{1}{4^{n-1}}.$$

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

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

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

The value of $x(0)$ can be verified using $x(0) = \lim_{z\to\infty} X(z)$.

The resulting inverse z-transform for $n \geq 0$ is

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

Using the theorem of residuum we can prove that $x(n) = 0$ for $n < 0$ with the region of convergence $|z| > 1/2$.

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

4.4 DISCRETE SYSTEMS AND THE z-TRANSFORM

For a linear time-invariant discrete system described by

$$y(n)=x(n)*_n h(n)=\sum_{m=-\infty}^{\infty}x(m)h(n-m),$$

the z-transform is derived in Section 4.2.5 in the form

$$Y(z)=X(z)H(z).$$

The z-transform of the output signal 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 the linear time invariant systems with the transfer function properties.

The system is stable if

$$\sum_{n=-\infty}^{\infty}|h(n)|<\infty.$$

It means that the z-transform exists at $|z|=1$, that is, 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 $H(z)=h(0)+h(1)z^{-1}+h(2)z^{-2}+\dots$ 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 4.10. For the systems whose transfer functions are

$$H_1(z)=\frac{1}{(z-1/3)(z-3/2)},\ |z|>3/2$$
$$H_2(z)=\frac{1}{z(z-1/3)(z-3/2)},\ 1/3<|z|<3/2$$
$$H_3(z)=\frac{1}{(z-1/3)(z-3/4)},\ |z|>3/4$$

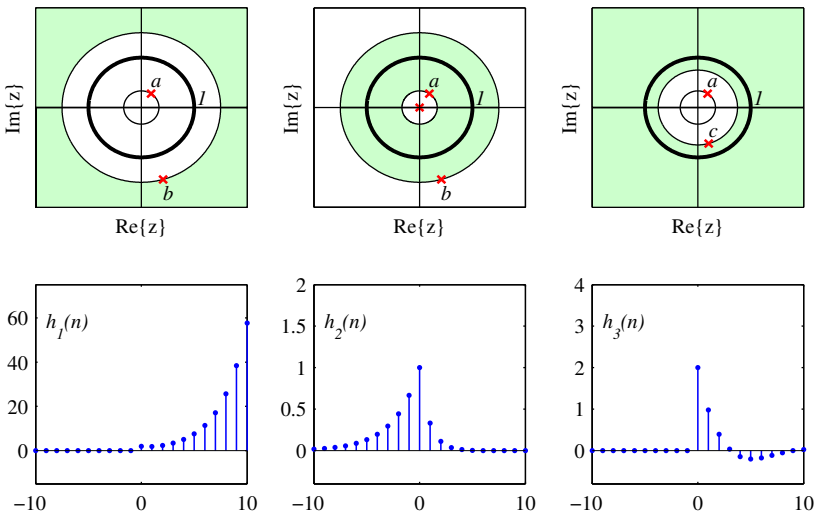


Figure 4.3 Regions of convergence (gray) with corresponding signals. Poles are marked by "x".

plot the regions of convergence and discuss the stability and causality. Find and plot the impulse response for every 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 shown 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

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

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

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

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

Example 4.11. Plot the frequency response of a 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

$$|H(e^{j\omega})| = \frac{|e^{j\omega} - e^{j\pi/3}|}{|e^{j\omega} - 0.95e^{j\pi/3}|} = \frac{\overline{TO_1}}{\overline{TP_1}}$$

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

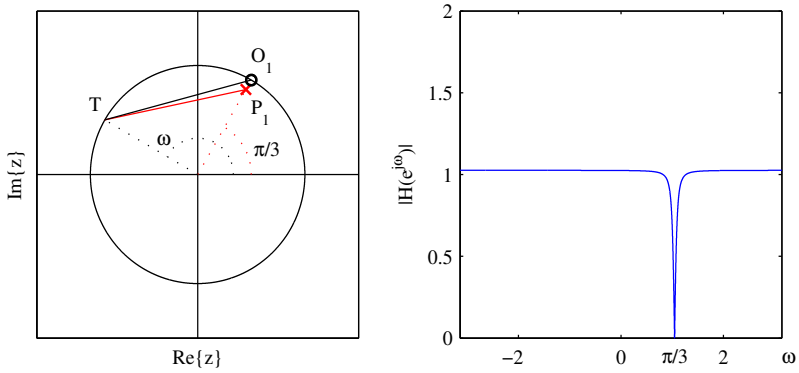


Figure 4.4 Poles and zeros of a first-order notch filter (left). The frequency response of this notch filter (right).

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 $y(n)$, 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

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

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

$$Y(z) = \frac{B_0 + B_1 z^{-1} + \cdots + B_M z^{-M}}{1 + A_1 z^{-1} + \cdots + A_N z^{-N}} X(z).$$

Example 4.12. A causal discrete system is described by the difference equation

$$y(n) - \frac{5}{6}y(n-1) + \frac{1}{6}y(n-2) = x(n). \tag{4.8}$$

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 $Y(z) - \frac{5}{6}z^{-1}Y(z) + \frac{1}{6}z^{-2}Y(z) = X(z)$, producing

$$Y(z) = \frac{1}{1 - \frac{5}{6}z^{-1} + \frac{1}{6}z^{-2}} X(z).$$

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

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

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

$$= 6\frac{1}{2^n} - \frac{8}{3^n} + \frac{3}{4^n}.$$

For $n = 0$ there is no pole at $z = 0$. Thus, the above expressions hold for $n = 0$ as well. The output signal is given by

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

Note: This kind of solution assumes the initial conditions from the system causality and $x(n)$ in the form: $y(0) = x(0) = 1$ and $y(1) - 5y(0)/6 = x(1)$, that is, $y(1) = 13/12$.

Example 4.13. A first-order causal discrete system is described by the following difference equation

$$y(n) + A_1 y(n-1) = B_0 x(n) + B_1 x(n-1). \tag{4.9}$$

Find its impulse response and discuss its behavior in terms of the system coefficient A_1.

★For the impulse response calculation the input signal is defined by $x(n) = \delta(n)$ 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 $z = -A_1$. The 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 $q = A_1 z^{-1} = (A_1/z)$

$$\begin{aligned} Y(z) &= \left(B_0 + B_1 z^{-1}\right)\left(1 - A_1 z^{-1} + A_1^2 z^{-2} - A_1^3 z^{-3} + \cdots + (-A_1 z^{-1})^n + \ldots\right) \\ &= B_0 + B_0 \sum_{n=1}^{\infty} (-A_1)^n z^{-n} + B_1 \sum_{n=1}^{\infty} (-A_1)^{n-1} z^{-n} \\ &= B_0 + (-A_1 B_0 + B_1) \sum_{n=1}^{\infty} (-A_1)^{n-1} z^{-n} \end{aligned}$$

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 is referred to as the *infinite impulse response (IIR) systems or recursive systems*. If the value of coefficient A_1 is zero-valued, that is $A_1 = 0$, then there is no recursion and

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

This is the system with a *finite impulse response (FIR)*. This kind of system produces an output to the signal $x(n)$ in the form

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

and is referred to as the *moving average (MA) system*. Systems without recursion are always stable since a finite sum of the finite signal values is always finite.

Systems that would contain only $x(n)$ and the output signal recursions,

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

are *auto-regressive (AR) systems or all pole systems*. This kind of systems could be unstable, due to the output signal recursion. In our case, the system is obviously unstable if $|A_1| > 1$. Systems defined by (4.9) 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 $q = z/A_1$ as

$$\begin{aligned} Y(z) &= \frac{B_0 + B_1 z^{-1}}{A_1 z^{-1}(z/A_1 + 1)} = \left(\frac{B_0}{A_1} z + \frac{B_1}{A_1}\right) \sum_{n=0}^{\infty} (-A_1^{-1} z)^n \\ &= B_0 \sum_{n=-\infty}^{0} (-A_1)^{n-1} z^{-(n-1)} + \frac{B_1}{A_1} \sum_{n=-\infty}^{0} (-A_1)^n z^{-n} \\ &= B_0 \sum_{n=-\infty}^{-1} (-A_1)^n z^{-n} + \frac{B_1}{A_1} \sum_{n=-\infty}^{0} (-A_1)^n z^{-n} \end{aligned}$$

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.

4.5.2 Solution to Difference Equations in the Time Domain

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

$$y(n) + A_1 y(n-1) + \cdots + A_N y(n-N) = x(n) \tag{4.10}$$

in the time domain will be described next.

The homogeneous part of this difference equation is

$$y(n) + A_1 y(n-1) + \cdots + A_N y(n-N) = 0. \tag{4.11}$$

Solution for the homogeneous equation is of the form

$$y_i(n) = C_i \lambda_i^n,$$

where C_i and λ_i are constants. Replacing $y_i(n)$ into (4.11), the characteristic polynomial equation follows

$$\begin{aligned} &C_i\lambda_i^n + C_i A_1 \lambda_i^{n-1} + \cdots + C_i A_N \lambda_i^{n-N} = 0, \\ \text{or} \qquad &\lambda_i^N + A_1 \lambda_i^{N-1} + \cdots + A_N = 0. \end{aligned}$$

This is a polynomial of the Nth order. In general, it has N solutions λ_i, $i = 1,2,\ldots,N$. All functions $y_i(n) = \lambda_i^n$, $i = 1,2,\ldots,N$ are the solutions of equation (4.11). 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 (4.11). This solution is called homogeneous part of the solution to (4.10).

Next a particular solution $y_p(n)$, corresponding to the form of the input signal $x(n)$, should be found. The solution to equation (4.10) is then

$$y(n) = y_h(n) + y_p(n).$$

The constants $C_i,\ i = 1,2,\ldots,N$ are calculated based on the initial conditions $y(i-1), i = 1,2,\ldots,N$.

Example 4.14. Find the output of a causal discrete system

$$y(n) - \frac{5}{6}y(n-1) + \frac{1}{6}y(n-2) = x(n) \tag{4.12}$$

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 $y(1) = 5$.

★The solution to the homogeneous part of difference equation (4.12)

$$y(n) - \frac{5}{6}y(n-1) + \frac{1}{6}y(n-2) = 0$$

is of the form $y_i(n) = C_i\lambda_i^n$. Its replacement into the equation results in the characteristic polynomial

$$\lambda_i^2 - \frac{5}{6}\lambda_i + \frac{1}{6} = 0,$$

whose roots are $\lambda_1 = 1/2$ 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 $x(n) = (n + 11/6)u(n)$ is a linear function of n, a particular solution is of the form $y_p(n) = An + B$. Replacing $y_p(n)$ into (4.12) 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 to (4.12) is a sum of the homogeneous and the particular solution,

$$y(n) = y_h(n) + y_p(n) = C_1\frac{1}{2^n} + C_2\frac{1}{3^n} + 3n + 1.$$

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 $y(0) = x(0) = 11/6$ 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 4.15. Goertzel algorithm: Show that the discrete-time signal

$$y(n) = e^{j(2\pi k_0 n/N+\varphi)}$$

is the solution to the homogeneous difference equation

$$y(n) - e^{j2\pi k_0/N} y(n-1) = 0. \tag{4.13}$$

Consider a periodic signal $x(n)$ with a period N and its DFT values $X(k)$,

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j2\pi nk/N}. \tag{4.14}$$

If the signal within one of its periods, $0 \le n \le N-1$, is applied as the input to the system described by difference equation (4.13) show that the output signal at $n = N-1$ is equal to the DFT of the signal $x(n)$ at frequency index $k = k_0$, that is

$$y(N-1) = X(k_0).$$

★For the signal $y(n)$ holds

$$\begin{aligned} y(n) &= e^{j(2\pi k_0 n/N+\varphi)} = e^{j(2\pi k_0 (n-1+1)/N+\varphi)} \\ &= e^{j(2\pi k_0/N)} y(n-1). \end{aligned}$$

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 (4.14) in the form

$$x_k(n) = \frac{1}{N} X(k) e^{j2\pi nk/N},$$

for an arbitrary $0 \le k \le N-1$. Then the difference equation (and the z-transform relation) for this input signal reads

$$\begin{aligned} y_k(n) - e^{j2\pi k_0/N} y_k(n-1) &= x_k(n) \\ Y_k(z) &= \frac{\mathcal{Z}\{x_k(n)\}}{1 - e^{j2\pi k_0/N} z^{-1}}. \end{aligned} \tag{4.15}$$

The z-transform of $x_k(n)$, for $0 \leq n \leq N-1$, is

$$\mathcal{Z}\{x_k(n)\} = \mathcal{Z}\{\frac{1}{N}X(k)e^{j2\pi nk/N}\} \tag{4.16}$$
$$= \frac{1}{N}X(k)\sum_{n=0}^{N-1} e^{j2\pi nk/N}z^{-n} = \frac{1}{N}X(k)\frac{1-e^{j2\pi k}z^{-N}}{1-e^{j2\pi k/N}z^{-1}}.$$

The transform $\mathcal{Z}\{x_k(n)\}$, for a given k, has zeros at

$$z_0^N = e^{j2\pi k + j2l\pi}, \ l = 0,1,2,\ldots,N-1$$

or

$$z_0 = e^{j2\pi(k+l)/N}, \ l = 0,1,2,\ldots,N-1.$$

Note that the zero

$$z_0 = e^{j2\pi k/N}, \text{ obtained for } l = 0$$

is canceled with the pole $z_p = e^{j2\pi kn/N}$ in (4.16). Therefore the remaining zeros are at

$$z_0 = e^{j2\pi(k+l)/N}, l = 1,2,\ldots,N-1.$$

The output z-transform, $Y_k(z)$, defined by (4.15), has a pole at

$$z_p = e^{j2\pi k_0/N}.$$

- If $k \neq k_0$ then one of zeros $z_0 = e^{j2\pi(k+l)/N}$, $l = 1,2,\ldots,N-1$ will coincide with the pole $z_p = e^{j2\pi k_0/N}$ and will cancel it out. Thus, for $k \neq k_0$, the function $Y_k(z)$ will not have any pole. Then,

$$y_k(N-1) = \frac{1}{2\pi j}\oint_C z^{N-2}Y_k(z)dz = 0 \tag{4.17}$$

since there are no poles within C, Fig. 4.5.

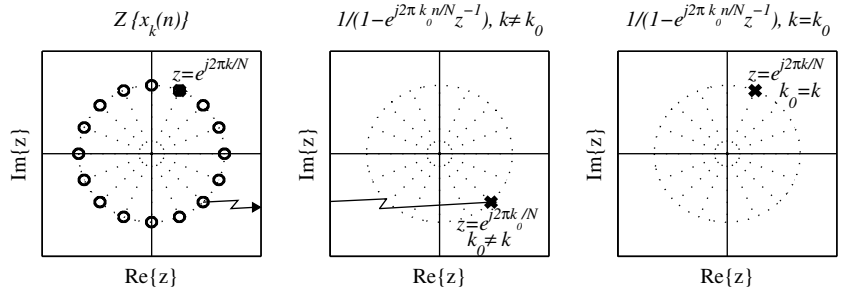


Figure 4.5 Zeros and the pole in $\mathcal{Z}\{x_k(n)\}$ (left), the pole in $1/(1-e^{j2\pi k_0 n/N}z^{-1})$ for $k \neq k_0$ (middle), and the pole in $1/(1-e^{j2\pi k_0 n/N}z^{-1})$ for $k = k_0$ (right). Illustration is for $N = 16$.

- If $k = k_0$, then the pole at $k = k_0$ is already canceled out in $\mathcal{Z}\{x_k(n)\}$ and $z_p = e^{j2\pi k_0/N}$ remains as a pole of $Y(z)$. In this case, the signal value at $n = N-1$ is equal to the residuum of

the function in (4.17) at the pole $z_p = e^{j2\pi k_0/N}$, relation (4.7),

$$\begin{aligned} y_{k_0}(N-1) &= z^{N-2} Y_{k_0}(z)(z - e^{j2\pi k_0/N})\Big|_{z=e^{j2\pi k_0/N}} \\ &= z^{N-1}\frac{1}{N}X(k_0)\frac{1 - e^{j2\pi k_0} z^{-N}}{1 - e^{j2\pi k_0/N} z^{-1}}\Bigg|_{z=e^{j2\pi k_0/N}} \\ &= \frac{1}{N}X(k_0)\lim_{z\to e^{j2\pi k_0/N}} \frac{z^N - e^{j2\pi k_0}}{z - e^{j2\pi k_0/N}} = X(k_0). \end{aligned}$$

Therefore, the output signal of the system, at $n = N-1$, is

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

Note: The difference relation

$$y(n) - e^{j2\pi k_0 n/N} y(n-1) = x(n) \tag{4.18}$$

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

$$\begin{aligned} Y(z) &= \frac{X(z)}{1 - e^{j2\pi k_0 n/N} z^{-1}} \frac{1 - e^{-j2\pi k_0 n/N} z^{-1}}{1 - e^{-j2\pi k_0 n/N} z^{-1}} \\ Y(z) &= \frac{1 - e^{-j2\pi k_0 n/N} z^{-1}}{1 - 2\cos(2\pi k_0 n/N) z^{-1} + z^{-2}} X(z) \end{aligned}$$

In the discrete-time domain the system

$$y(n) - 2\cos(2\pi k_0/N) y(n-1) + y(n-2) = x(n) - e^{-j2\pi k_0 n/N} x(n-1) \tag{4.19}$$

is called the Goertzel algorithm for the DFT calculation at a given single frequency k_0.

It is interesting to note that the computation of (4.19) is more efficient than the computation of (4.18). For the calculation of (4.18), 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 (4.19) we can use linear property and calculate only

$$y_1(n) - 2\cos(2\pi k_0/N) y_1(n-1) + y_1(n-2) = x(n) \tag{4.20}$$

at every instant. It requires a multiplication of complex signal with a real coefficient. It means 2 real multiplications for every instant or $2N$ in total for N instants. The resulting output, at the instant $N-1$, is

$$\begin{aligned} y(N-1) &= T\{x(N-1)\} - e^{-j2\pi k_0 (N-1)/N} T\{x(N-1)\} \\ &= y_1(N) - e^{j2\pi k_0} y_1(N-1). \end{aligned}$$

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 (4.20) 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

The z-transform will be related to the all other transform, presented so far. Consider a continuous-time signal $x(t)$ and its the Laplace transform $X(s)$. If the integral in the Laplace transform is approximated by the corresponding sum, we have

$$X(s) = \int_{-\infty}^{\infty} x(t)e^{-st}dt \cong \sum_{n=-\infty}^{\infty} x(n\Delta t)e^{-sn\Delta t}\Delta t = \sum_{n=-\infty}^{\infty} x(n)e^{-sn\Delta t},$$

with $x(n) = x(n\Delta t)\Delta t$. When this relation is compared to the z-transform definition we can conclude that the Laplace transform of $x(t)$ can be approximated by the z-transform of its samples with

$$z = \exp(s\Delta t),$$

that is,

$$X(s) \leftrightarrow X(z)_{|z=\exp(s\Delta t)}. \tag{4.21}$$

A point in the complex Laplace domain, $s = \sigma + j\Omega$, maps to 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 the unit circle in the z domain, $r < 1$.

In applied mathematics, the transform $X(z)$ at $z = \exp(s\Delta t)$ is called the *starred or star transform*. It can be obtained as the Laplace transform of the sampled signal, denoted in the continuous-time domain as

$$x^*(t) = x(t)\sum_{n=-\infty}^{\infty}\delta(t-n\Delta t).$$

The starred transform is derived by

$$X(s) = \int_{-\infty}^{\infty} x^*(t)e^{-st}dt = \sum_{n=-\infty}^{\infty} x(n\Delta t)e^{-sn\Delta t}\int_{-\infty}^{\infty}\delta(t-n\Delta t)dt = \sum_{n=-\infty}^{\infty} x(n)e^{-sn\Delta t}. \tag{4.22}$$

According to the sampling theorem, for the Laplace transform of discrete-time signal holds $X(s)_{|\sigma=0} = X(j\Omega) = X(j(\Omega + 2k\pi/\Delta t))$.

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 4.16. The Fourier transform of a causal discrete-time signal $x(n)$ is $X(e^{j\omega})$. Write its z-transform in terms of the Fourier transform of the discrete-time signal, that is, write the z-transform values in the complex plane based on their values on the unit circle.

★The signal $x(n)$ can be expressed in term of its Fourier transform as

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

The z-transform of this signal is given by

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

for $|z| > 1$. In this way, the relation between $X(z)$, for any z, and $X(e^{j\omega})$ is established.

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

$$X(k) = X(e^{j\omega})_{|\omega=2\pi k/N} = X(z)_{|z=e^{j2\pi k/N}} = \sum_{n=0}^{N-1} x(n)z^{-n}_{|z=e^{j2\pi k/N}}.$$

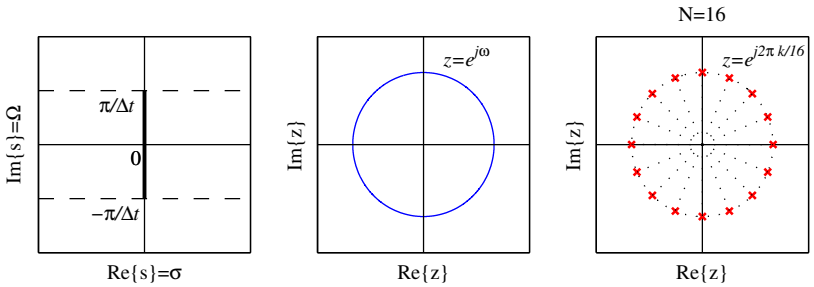


Figure 4.6 Illustration of the z-transform relation with the Laplace transform (left), the Fourier transform of discrete signals (middle), and the DFT (right).

Example 4.17. Consider a discrete-time signal $x(n)$ with N samples different from zero within $0 \le n \le N-1$. 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.

★If the signal has N nonzero samples, then it can be expressed in term of its DFT as

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

Thus, the z-transform of this $x(n)$ can be expressed in terms of the IDFT within $0 \le n \le N-1$, in the form

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

with $X(k) = X(z)$ at $z = \exp(j2\pi k/N)$, $k = 0,1,2,\ldots,N-1$.

In this way, the z-transform $X(z)$ is expressed in terms of its samples on the unit circle at $z = \exp(j2\pi k/N)$.

For a periodic signal $x(n)$, when all periods are included 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) $x(n) = a^{|n|}$,
(c) $x(n) = \frac{1}{2^n}u(n) + \frac{1}{3^n}u(n)$

Problem 4.2. Find the z-transform and the region of convergence for the signals:
(a) $x(n) = \delta(n+1) + \delta(n) + \delta(n-1)$,
(b) $x(n) = \frac{1}{2^n}[u(n) - u(n-10)]$.

Problem 4.3. Using the z-transform 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

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

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 the discrete system is given by

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

Find the impulse response if:

(a) The system is stable,
(b) The region of convergence is $\frac{1}{4} < |z| < \frac{1}{3}$,
(c) The system is anticausal.

Problem 4.9. For the z-transform

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

identify possible regions of convergence. For every case, comment the 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 the following difference equation

$$y(n) - y(n-1) + \frac{5}{16}y(n-2) - \frac{1}{32}y(n-3) = 3x(n) - \frac{5}{4}x(n-1) + \frac{3}{16}x(n-2).$$

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 is characterized by the impulse response

$$h(n) = 1/3^n u(n).$$

Using the z-transform find the output of the system to the input signal $x(n) = u(n) - u(n-6)$.

Problem 4.15. Find the output of the 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

$$x(n+2) + 3x(n+1) + 2x(n) = 0$$

using the z-transform. The initial conditions are $x(0) = 0$ and $x(1) = 1$. The signal $x(n)$ is causal.

Problem 4.17. Solve the difference equation

$$x(n+1) = x(n) + a^n u(n)$$

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

Problem 4.18. Find the output of the causal discrete system

$$y(n) - \frac{\sqrt{2}}{2}y(n-1) + \frac{1}{4}y(n-2) = x(n) \tag{4.23}$$

to the input signal $x(n) = \frac{1}{3^n}u(n)$ by solving the differential equation in the discrete-time domain and 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 mth 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 mth 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 pole-zero geometry plot the amplitude of the frequency response of the system described by the difference equation

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

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

$$x(t) = 2\cos(10\pi t) - \sin(15\pi t) + 0.5e^{j20\pi t}$$

sampled with the sampling interval $\Delta t = 1/60$.

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 $r^{1/N} \cong 1$. Show that this system has the same transfer function as

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

4.8 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) $x(n) = 1/3^{|n|} + 1/2^n u(n)$,
(d) $x(n) = 3^n u(-n) + 2^{-n} u(n)$,
(e) $x(n) = n(1/3)^n u(n)$.
(f) $x(n) = \cos(n\frac{\pi}{2})$.

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) $x(n) = \sum_{k=0}^{n} 3^k$.

Exercise 4.3. Find the inverse z-transform of:
(a) $X(z) = \frac{z^{-8}}{1-z} + 3$, 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 infinite-duration signal $x(n)$.
Find $\sum_{n=-\infty}^{\infty} x(n)$ in this case.

Exercise 4.4. Find the inverse z-transforms of:
(a) $X(z) = \frac{z^{-5}(5z-3)}{(3z-1)(2z-4)}$, 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 of 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) $y(n) = x(n) *_n x(-n)$, where $*_n$ denotes the 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. In each of these cases, 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 the causal discrete system

$$y(n) - \frac{3}{4}y(n-1) + \frac{1}{8}y(n-2) = x(n). \tag{4.24}$$

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

$$y(n) - \frac{5}{6}y(n-1) + \frac{1}{6}y(n-2) = x(n). \tag{4.25}$$

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. This input signal 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 conditions $x(0) = 0$ and $x(1) = 1/2$. The signal $x(n)$ is causal.

Exercise 4.12. Using the basic trigonometric transformations show that the real-valued signal $y(n) = \cos(2\pi k_0 n/N + \varphi)$ is a solution to 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

$$x(n) = \cos(n\pi/3 + \pi/4) + \sin(n\pi/2) + (-1)^n.$$

4.9 SOLUTIONS

Solution 4.1. (a) The z-transform of the signal $x(n) = \delta(n-2)$ is

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

for any $z \neq 0$.

(b) For the signal $x(n) = a^{|n|}$, the z-transform is given by

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

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

(c) In this case, when $x(n) = \frac{1}{2^n} u(n) + \frac{1}{3^n} u(n)$, the z-transform is

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

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

for $|z| > 1/2$ and $|z| > 1/3$. The region of convergence is $|z| > 1/2$.

Solution 4.2. (a) The z-transform of signal $x(n) = \delta(n+1) + \delta(n) + \delta(n-1)$ 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 $z \longrightarrow \infty$.

(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 of $x(n)$ is

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

$$= \frac{z^{-10}}{z^{-1}} \frac{z^{10} - (\frac{1}{2})^{10}}{z - \frac{1}{2}} = \frac{z^{10} - (\frac{1}{2})^{10}}{z^9 (z - \frac{1}{2})}$$

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 $z_{0i} = e^{j2i\pi/10}/2$, 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 $z = 0$.

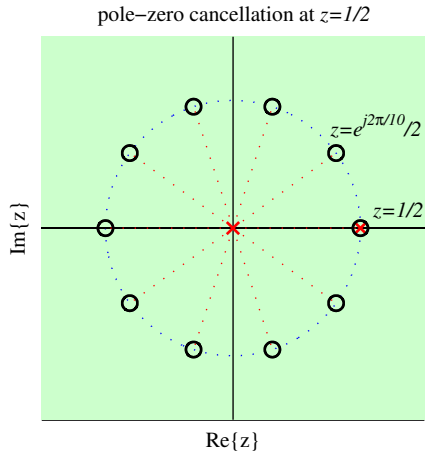


Figure 4.7 Pole-zero cancellation at $z = 1/2$.

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 $z^{-1}X(z) = \mathcal{Z}[x(n-1)]$. This means that

$$x(n) = \frac{a}{n}x(n-1)$$

for $n > 0$. According to the initial value theorem, we the value of $x(0)$ is equal to

$$x(0) = \lim_{z\to\infty} X(z) = 1.$$

The signal samples are obtained in the form

$$x(1) = a, \quad x(2) = \frac{a^2}{2}, \quad x(3) = \frac{a^3}{2\cdot 3}, \ldots$$

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 the signal $x(-n)$ is

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

With the following substitution, $-n = m$, the relation

$$X_1(z) = \sum_{m=-\infty}^{\infty} x(m)z^m = X(1/z),$$

follows. The region of convergence is complementary to the one holding for 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

$$X^*(z^*) = \Big(\sum_{n=-\infty}^{\infty} x(n)(z^*)^{-n}\Big)^* = \sum_{n=-\infty}^{\infty} x^*(n)\left((z^*)^{-n}\right)^*.$$

Since $\left((z^*)^{-n}\right)^* = \left((z^{-n})^*\right)^* = z^{-n}$, we get

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

for the real-valued signal $x(n)$.

Solution 4.5. From the z-transform definition

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

and using the substitution $n + k = m$, the final result,

$$Y(z) = X(z)X(\frac{1}{z}),$$

follows.

Solution 4.6. A direct expansion of the given transform into power series within the region of convergence is used. In order to find the signal $x(n)$, whose z-transform is $X(z) = \frac{1}{2-3z}$, the z-transform should be written in the form of the power series of $X(z)$ in powers of z^{-1}. Since the condition $\left|\frac{3z}{2}\right| < 1$ 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 $\left|\frac{2}{3z}\right| < 1$, that is $|z| > \frac{2}{3}$, corresponds to the given region of convergence. In order to obtain the inverse z-transform, we can write

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

where

$$X_1(z) = \frac{1}{1-\frac{2}{3z}}.$$

The power series expansion of $X_1(z)$ gives

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

It can be concluded that the original z-transform, $X(z)$, can be written as

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

In order to get the z-transform definition form

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n}, \tag{4.26}$$

the last expression will be rewritten as 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 to (4.26) we get the signal $x(n)$ in the form

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

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). The poles of the z-transform $X(z)$ 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 can be written as

$$X(z) = \frac{z+1}{(2z-1)(3z+2)} = \frac{A}{2z-1} + \frac{B}{3z+2}$$
$$A = \frac{3}{7}, \quad B = -\frac{1}{7}.$$

The terms in $X(z)$ should be expressed in such a way that they represent sums of geometric series for the given region of convergence. Based on the solution to the previous problem, we conclude that

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

Now we can write

$$\frac{A}{2z}\frac{1}{1-\frac{1}{2z}} = \frac{A}{2z}\sum_{n=0}^{\infty}\left(\frac{1}{2}\right)^n z^{-n} = \frac{A}{2}\sum_{n=0}^{\infty}\left(\frac{1}{2}\right)^n z^{-n-1}, \quad |z| > \frac{1}{2}$$

and

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

The z-transform, with $m = n+1$, is of 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 to 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

$$H(z) = \frac{3-\frac{5}{6}z^{-1}}{(1-\frac{1}{4}z^{-1})(1-\frac{1}{3}z^{-1})} = \frac{A}{1-\frac{1}{4}z^{-1}} + \frac{B}{1-\frac{1}{3}z^{-1}},$$

with $A = 1$ and $B = 2$.

(a) The region of convergence must contain $|z| = 1$, for a stable system. This region of convergence 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 $|z| > \frac{1}{4}$. This term 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,

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

The 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

$$\frac{1}{1-\frac{1}{4}z^{-1}}=-\frac{4z}{1-4z}=-\sum_{n=1}^{\infty}(4z)^n\underset{m=-n}{=}-\sum_{m=-\infty}^{-1}(4z)^{-m}\text{ with }|z|<\frac{1}{4}.$$

The signal corresponding to this term is $-4^{-n}u(-n-1)$. The impulse response of the anticausal discrete system, with the given transfer function $H(z)$, is

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

Solution 4.9. The transfer function $H(z)$ can be written as

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

with the 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 $|z_2|=|z_3|=1/2$, the 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 the input signal

$$x(n)=2\cos(n\pi/2)=1+\cos(n\pi)=1+(-1)^n$$

is

$$y(n)=H(e^{j\omega})_{|\omega=0}+H(e^{j\omega})_{|\omega=\pi}(-1)^n=H(z)_{|z=1}+H(z)_{|z=-1}(-1)^n=-0.8681+0.0945(-1)^n.$$

Solution 4.10. The transfer function $H(z)$ can be written as

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

The multiplication of both sides of the last expression 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 A, B, and C follow from the system of equations

$$A+B=0,\quad -2B+C=1,\text{ and }\quad -2C=2$$

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 of a causal system is $|z| > 2$. The inverse z-transform of $H(z)$ for causal system is the system impulse response equal to

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

This system is not stable.

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

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

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 value 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

$$\begin{aligned} H(z) &= \frac{3-\frac{5}{4}z^{-1}+\frac{3}{16}z^{-2}}{1-z^{-1}+\frac{5}{16}z^{-2}-\frac{1}{32}z^{-3}} = \frac{3-\frac{5}{4}z^{-1}+\frac{3}{16}z^{-2}}{(1-\frac{1}{2}z^{-1}+\frac{1}{16}z^{-2})(1-\frac{1}{2}z^{-1})} \\ &= -\frac{1}{1-\frac{1}{4}z^{-1}} - \frac{1}{\left(1-\frac{1}{4}z^{-1}\right)^2} + \frac{5}{(1-\frac{1}{2}z^{-1})}. \end{aligned}$$

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

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

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

the inverse z-transform is

$$h(n) = -\frac{1}{4^n}u(n) - (n+1)\frac{1}{4^n}u(n) + 5\frac{1}{2^n}u(n).$$

Solution 4.13. The transfer function of the system defined by difference equation

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

is

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

The z-transform of the input signal $x(n) = 1/4^n u(n)$ is equal to

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

with the region of convergence $|z| > 1/4$. The z-transform of the output signal 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 transform is a finite duration output signal, given by

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

Solution 4.14. The system transfer function is given by

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

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

$$Y(z) = \frac{1 - z^{-6}}{(1 - z^{-1})(1 - 1/3z^{-1})} = Y_1(z) - Y_1(z)z^{-6}$$

with

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

Its inverse z-transform is

$$y_1(n) = \left[\frac{3}{2} - \frac{1}{2}\left(\frac{1}{3}\right)^n\right] u(n).$$

The system output is obtained in the form

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

Solution 4.15. The transfer function follows from

$$Y(z)(1 - \frac{11}{6}z^{-1} + \frac{1}{2}z^{-2}) = X(z)(2 - \frac{3}{2}z^{-1})$$

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 $|z| > 3/2$. This means that the system is not stable, Fig. 4.8.

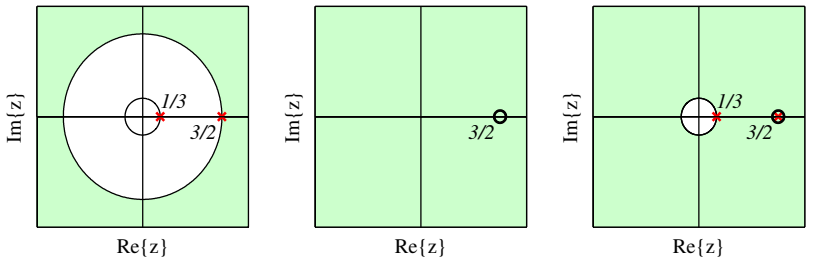


Figure 4.8 Poles and zeros of the system (left), the input signal z-transform (middle), and the z-transform of the output signal (right).

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

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

The output signal transform does not have a pole at $z = 3/2$, since this pole is canceled out. The output signal is

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

Solution 4.16. The z-transform of the signal $x(n+2)$ is

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

while for the signal $x(n+1)$, the z-transform is

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

The z-transform domain form the difference equation $x(n+2) + 3x(n+1) + 2x(n) = 0$ 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 this 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 to the homogeneous part of the equation

$$y(n) - \frac{\sqrt{2}}{2}y(n-1) + \frac{1}{4}y(n-2) = 0 \tag{4.27}$$

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 homogeneous solution to the difference equation is

$$y_h(n) = C_1(\frac{\sqrt{2}}{4} + j\frac{\sqrt{2}}{4})^n + C_2(\frac{\sqrt{2}}{4} - j\frac{\sqrt{2}}{4})^n = C_1\frac{1}{2^n}e^{jn\pi/4} + C_2\frac{1}{2^n}e^{-jn\pi/4}.$$

A particular solution is assumed in the form of the input signal $x(n) = \frac{1}{3^n}u(n)$, that is

$$y_p(n) = A\frac{1}{3^n}u(n).$$

The constant A is obtained by replacing this signal into (4.23)

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

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

Its value is $A = 0.886$. The general solution to the considered difference equation is equal to the sum of the homogeneous solution and the particular solution

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

Since the system is causal with $y(n) = 0$ for $n < 0$, the constants C_1 and C_2 can be obtained from the initial conditions 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

$$y(1) = \frac{\sqrt{2}}{2}y(0) + x(1) = \frac{\sqrt{2}}{2} + \frac{1}{3}$$

. With this initial conditions, we get

$$C_1 + C_2 + 0.886 = 1 \tag{4.28}$$
$$C_1(\frac{\sqrt{2}}{2} + j\frac{\sqrt{2}}{2})/2 + C_2(\frac{\sqrt{2}}{2} - j\frac{\sqrt{2}}{2})/2 + 0.886\frac{1}{3} = \frac{\sqrt{2}}{2} + \frac{1}{3},$$

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

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

or

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

Using the residual value based inversion of the z-transform, we can get the signal in the form

$$y(n) = \sum_{z_{1,2,3} = \frac{\sqrt{2}}{4} \pm j\frac{\sqrt{2}}{4}, 1/3} \left\{ [z^{n-1}Y(z)(z - z_i)]_{|z=z_i} \right\}.$$

With the residual value definition for the simple poles z_1, z_2, and z_3 we get

$$y(n) = z^{n+2}\left.\frac{1}{(z - \frac{\sqrt{2}-j\sqrt{2}}{4})(z - \frac{1}{3})}\right|_{\frac{\sqrt{2}+j\sqrt{2}}{4}} + z^{n+2}\left.\frac{1}{(z - \frac{\sqrt{2}+j\sqrt{2}}{4})(z - \frac{1}{3})}\right|_{z=\frac{\sqrt{2}-j\sqrt{2}}{4}}$$
$$+ z^{n+2}\left.\frac{1}{(z - \frac{\sqrt{2}+j\sqrt{2}}{4})(z - \frac{\sqrt{2}-j\sqrt{2}}{4})}\right|_{z=1/3}$$
$$= \frac{1}{\frac{j\sqrt{2}}{2}}\left(\frac{\sqrt{2}+j\sqrt{2}}{4}\right)^{n+2}\frac{1}{\frac{\sqrt{2}+j\sqrt{2}}{4} - \frac{1}{3}} - \frac{1}{\frac{j\sqrt{2}}{2}}\left(\frac{\sqrt{2}-j\sqrt{2}}{4}\right)^{n+2}\frac{1}{\frac{\sqrt{2}-j\sqrt{2}}{4} - \frac{1}{3}} + \frac{1}{3^{n+2}}\frac{1}{(\frac{1}{9} - \frac{1}{3}\frac{\sqrt{2}}{2} + \frac{1}{4})}$$
$$= \frac{1}{2^{n+2}}e^{j(n+2)\pi/4}\frac{-j\sqrt{2}}{\frac{\sqrt{2}+j\sqrt{2}}{4} - \frac{1}{3}} + \frac{1}{2^{n+2}}e^{-j(n+2)\pi/4}\frac{j\sqrt{2}}{\frac{\sqrt{2}-j\sqrt{2}}{4} - \frac{1}{3}} + 0.886\frac{1}{3^n}$$
$$= \frac{1}{2^n}e^{jn\pi/4}\frac{\sqrt{2}}{\sqrt{2} + j\sqrt{2} - \frac{4}{3}} + \frac{1}{2^n}e^{-jn\pi/4}\frac{\sqrt{2}}{\sqrt{2} - j\sqrt{2} - \frac{4}{3}} + 0.886\frac{1}{3^n}$$
$$= 2 \times 0.9984\frac{1}{2^n}\cos(n\pi/4 - 1.5137) + 0.886\frac{1}{3^n},$$

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

$$\begin{aligned}\nabla^2 x(n) &= \nabla[\nabla x(n)] = \nabla[x(n) - x(n-1)] = \nabla x(n) - \nabla x(n-1) \\ &= x(n) - 2x(n-1) + x(n-2).\end{aligned}$$

Its z-transform is

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

In the same way we get

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

The z-transform of the first forward difference is

$$\begin{aligned}\mathcal{Z}[\Delta x(n)] &= \mathcal{Z}[x(n+1) - x(n)] = zX(z) - zx(0) - X(z) \\ &= (z-1)X(z) - zx(0).\end{aligned}$$

The second forward difference is

$$Z[\Delta^2 x(n)] = x(n+2) - 2x(n+1) + x(n),$$

with the z-transform

$$Z[\Delta^2 x(n)] = (z-1)^2 X(z) - z(z-1)x(0) - z\Delta x(0).$$

The z-transform of the mth 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 can be written in the form

$$\begin{aligned}H(z) &= \frac{1 - \sqrt{2}z^{-1} + z^{-2}}{1 - r\sqrt{2}z^{-1} + r^2 z^{-2}} = \frac{[1 - (\frac{\sqrt{2}}{2} + j\frac{\sqrt{2}}{2})z^{-1}][1 - (\frac{\sqrt{2}}{2} - j\frac{\sqrt{2}}{2})z^{-1}]}{[1 - r(\frac{\sqrt{2}}{2} + j\frac{\sqrt{2}}{2})z^{-1}][1 - r(\frac{\sqrt{2}}{2} - j\frac{\sqrt{2}}{2})z^{-1}]} \\ &= \frac{[z - (\frac{\sqrt{2}}{2} + j\frac{\sqrt{2}}{2})][z - (\frac{\sqrt{2}}{2} - j\frac{\sqrt{2}}{2})]}{[z - r(\frac{\sqrt{2}}{2} + j\frac{\sqrt{2}}{2})][z - r(\frac{\sqrt{2}}{2} - j\frac{\sqrt{2}}{2})]}.\end{aligned}$$

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

The amplitude of the frequency response is

$$|H(e^{j\omega})| = \left|\frac{B_0}{A_0}\right| \frac{\overline{TO_1}\,\overline{TO_2}}{\overline{TP_1}\,\overline{TP_2}} = \frac{\overline{TO_1}\,\overline{TO_2}}{\overline{TP_1}\,\overline{TP_2}}.$$

The values of $\overline{TP_1}$ and $\overline{TO_1}$, and $\overline{TP_2}$ and $\overline{TO_2}$, are almost the same for any ω, except $\omega = \pm\pi/4$, where the distance to the corresponding zeros of the transfer function is 0, while the

distance to the corresponding pole is small but finite. Based on this analysis, the amplitude of the frequency response is shown in Fig. 4.9.

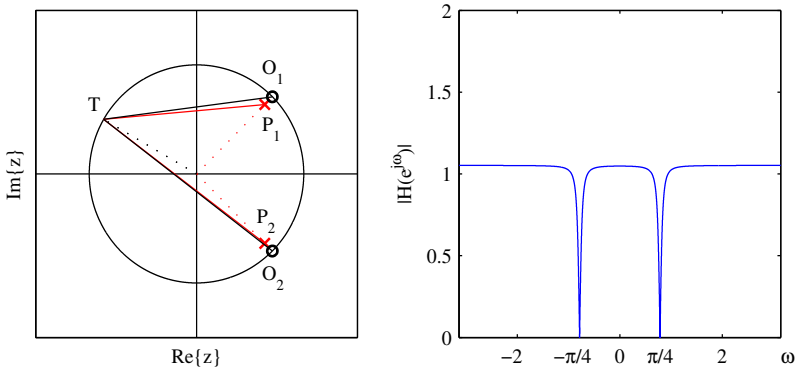


Figure 4.9 Location of zeros and poles for a second-order system.

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 the components at $\omega = \pm\pi/4$. The output discrete-time signal is

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

The corresponding continuous-time form of the output signal is

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

Solution 4.21. The zeros of the system are

$$z_o^{-N} = 1 = e^{-j2\pi m}$$
$$z_{om} = e^{j2\pi m/N},\ m = 0,1,\ldots,N-1.$$

Similarly, the poles are equal to $z_{mp} = r^{1/N}e^{j2\pi m/N}$, $m = 0,1,\ldots,N-1$. 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 $r^{1/N} \cong 1$, we get

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

The same conclusions hold 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 in the form

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

Chapter 5

From Continuous to Discrete Systems

TRANSFORMATION of continuous-time systems into corresponding discrete-time systems is of high importance. Some discrete-time systems are designed and realized in order to replace or perform as equivalents of the continuous-time systems. It is quite common to design a continuous-time system with desired properties, since the designing procedures in this domain are simpler and well developed. In the next step the obtained continuous-time system is then transformed into the corresponding discrete-time system.

Consider an Nth order linear continuous-time system described by a differential equation with constant coefficients

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

The Laplace transform domain equation for this system is

$$[a_N s^N + \cdots + a_1 s + a_0]Y(s) = [b_M s^M + \cdots + b_1 s + b_0]X(s), \tag{5.1}$$

assuming the zero-valued initial conditions.

The topic of this chapter is to find a corresponding discrete-time system, described by the difference equation

$$A_0 y(n) + A_1 y(n-1) + \cdots + A_N y(n-N) = B_0 x(n) + B_1 x(n-1) + \cdots + B_M x(n-M).$$

The z-transform domain form of this system is

$$[A_0 + A_1 z^{-1} + \cdots + A_N z^{-N}]Y(z) = [B_0 + B_1 z^{-1} + \cdots + B_M z^{-M}]X(z). \tag{5.2}$$

There are several approaches to establish the relation between the continuous-time system in (5.1) and the discrete-time system in (5.1), represented by their corresponding impulse responses or transfer functions.

5.1 IMPULSE INVARIANCE METHOD

A natural approach to transform a continuous-time system into the corresponding discrete-time system is based on the relation between the impulse responses of these two systems. Assume that the impulse response of the continuous-time system is $h_c(t)$. The impulse response $h(n)$ of the corresponding discrete-time system, according to this approach, is equal to the samples of $h_c(t)$,

$$h(n) = h_c(n\Delta t)\Delta t.$$

Obviously this relation can be used only if the sampling theorem is satisfied for the sampling interval

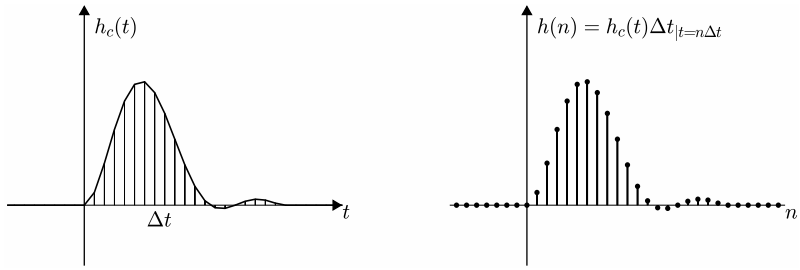


Figure 5.1 Sampling of the impulse response for the impulse invariance method.

Δt. It means that the frequency response of the continuous-time system must satisfy the sampling theorem condition

$$H(\Omega) = \mathrm{FT}\{h_c(t)\} = 0 \quad \text{for} \quad |\Omega| > \Omega_m$$

and $\Delta t < \pi/\Omega_m$. Otherwise the discrete-time version will not correspond to the continuous-time version of the frequency response. Here, the frequency response of the discrete-time system is related to a periodically extended form of the continuous-time system frequency response $H(\Omega)$ as

$$\sum_{k=-\infty}^{\infty} H(\Omega + 2k\pi/\Delta t) = H(e^{j\omega}), \qquad \Omega = \omega/\Delta t.$$

The transfer function of the continuous-time system in (5.1) may be decomposed using the partial fractions as

$$H(s) = \frac{b_M s^M + \cdots + b_1 s + b_0}{a_N s^N + \cdots + a_1 s + a_0} = \frac{k_1}{s - s_1} + \frac{k_2}{s - s_2} + \cdots + \frac{k_N}{s - s_N}, \tag{5.3}$$

where only simple poles, $s_1, s_2, \ldots, s_N$, of the transfer function are used. The case of multiple poles will be discussed later. The inverse Laplace transform of a causal system, described by the previous transfer function, is

$$h_c(t) = k_1 e^{s_1 t} u(t) + k_2 e^{s_2 t} u(t) + \cdots + k_N e^{s_N t} u(t).$$

The impulse response of the corresponding discrete-time system is equal to the the samples of $h_c(t)$,

$$h(n) = h_c(n\Delta t)\Delta t = [k_1 \Delta t e^{s_1 n\Delta t} u(n) + k_2 \Delta t e^{s_2 n\Delta t} u(n) + \cdots + k_N \Delta t e^{s_N n\Delta t} u(n)],$$

since $u(n\Delta t) = u(n)$. The z-transform of the impulse response $h(n)$ of the discrete-time system is

$$H(z) = \frac{k_1 \Delta t}{1 - e^{s_1 \Delta t} z^{-1}} + \frac{k_2 \Delta t}{1 - e^{s_2 \Delta t} z^{-1}} + \cdots + \frac{k_N \Delta t}{1 - e^{s_N \Delta t} z^{-1}}. \tag{5.4}$$

Comparing (5.3) to (5.4) it can be concluded that the terms in the transfer functions are transformed from the continuous-time to the discrete-time case as

$$\frac{k_i}{s - s_i} \to \frac{k_i \Delta t}{1 - e^{s_i \Delta t} z^{-1}}. \tag{5.5}$$

If a multiple pole s_i of the $(m+1)$th order exists in the continuous-time system transfer function, then this term can be written as

$$\frac{k_i}{(s-s_i)^{m+1}} = \frac{1}{m!}\frac{d^m}{ds_i^m}\frac{k_i}{s-s_i}.$$

The term in the discrete-time system, corresponding to this continuous-time system term, is

$$\frac{1}{m!}\frac{d^m}{ds_i^m}\left\{\frac{k_i}{s-s_i}\right\} \to \frac{1}{m!}\frac{d^m}{ds_i^m}\left\{\frac{k_i\Delta t}{1-e^{s_i\Delta t}z^{-1}}\right\}. \tag{5.6}$$

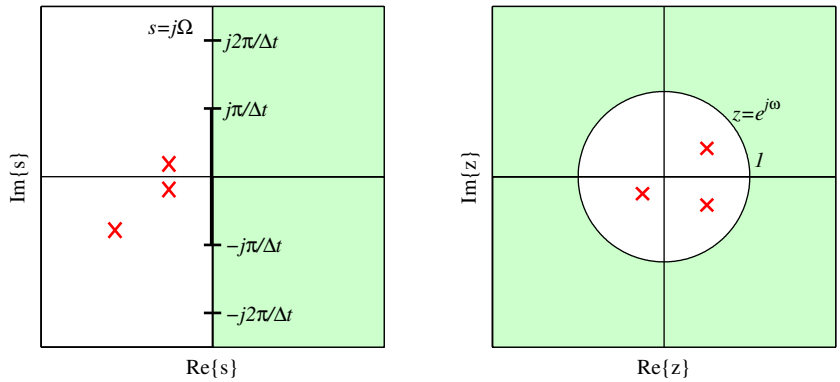


Figure 5.2 Illustration of the impulse invariance method mapping.

In the impulse invariance method, the poles are mapped according to

$$s_i \to e^{s_i\Delta t}.$$

This mapping relation does not hold for zeros, Fig. 5.2.

Impulse response with an initial instant discontinuity. In the case when the continuous-time impulse response $h_c(t)$ has a discontinuity at $t=0$, that is, when

$$h_c(t)|_{t=-0} \neq h_c(t)|_{t=+0},$$

then the previous forms assume that the discrete-time impulse response $h(0) = h_c(t)|_{t=+0}$. Recall that the theory of Fourier transforms in this case states that the inverse Fourier transform $\mathrm{IFT}\{H(j\Omega)\} = h_c(t)$ where the signal $h_c(t)$ is continuous and

$$\mathrm{IFT}\{H(j\Omega)\} = \left(h_c(t)|_{t=-0} + h_c(t)|_{t=+0}\right)/2$$

at the discontinuity points (in this case at $t=0$). The special case of discontinuity at $t=0$ can be easily detected for a causal system by mapping $H(s)$ onto $H(z)$ and by checking is the following relation satisfied

$$0 = h_c(t)|_{t=-0} = h_c(t)|_{t=+0} = h(n)|_{n=0} = \lim_{z\to\infty} H(z).$$

If $\lim_{z\to\infty} H(z) \neq 0$ then a discontinuity exists and we should use

$$h(0) = \lim_{z\to\infty} H(z)/2,$$

since $h_c(t)|_{t=-0} = 0$ and $h_c(t)|_{t=+0}\,\Delta t = \lim_{z\to\infty} H(z)$. The resulting frequency response is

$$H(z) - \lim_{z\to\infty} H(z)/2.$$

Example 5.1. A continuous-time system has a transfer function of the form

$$H(s) = \frac{s+\frac{3}{2}}{s^2+\frac{3}{2}s+\frac{1}{2}}. \tag{5.7}$$

What is the corresponding discrete-time system according to the impulse invariance method with $\Delta t = 1$?

★The partial fraction decomposition of the transfer function is

$$H(s) = \frac{s+\frac{3}{2}}{(s+1)(s+\frac{1}{2})} = \frac{k_1}{s+1} + \frac{k_2}{s+\frac{1}{2}}$$

with

$$k_1 = H(s)(s+1)|_{s=-1} = -1,$$
$$k_2 = H(s)(s+\frac{1}{2})\Big|_{s=-1/2} = 2.$$

Thus, we get

$$H(s) = \frac{-1}{s+1} + \frac{2}{s+\frac{1}{2}}.$$

According to (5.5) the discrete-time system is

$$H(z) = \frac{-1}{1-e^{-1}z^{-1}} + \frac{2}{1-e^{-1/2}z^{-1}}.$$

Since $\lim_{z\to\infty} H(z) = 1$, obviously there is a discontinuity in the impulse response and the resulting transfer function should be corrected as

$$H(z) = \frac{-1}{1-e^{-1}z^{-1}} + \frac{2}{1-e^{-1/2}z^{-1}} - 1/2.$$

The impulse response and the frequency response of the discrete-time systems with uncorrected and corrected discontinuity effect are presented in Fig. 5.3.

Example 5.2. A continuous-time system has a transfer function of the form

$$H(s) = \frac{(1-3s/2)}{(6s^2+5s+1)(s+1)^2}.$$

What is the corresponding discrete-time system according to the impulse invariance method with $\Delta t = 1$?

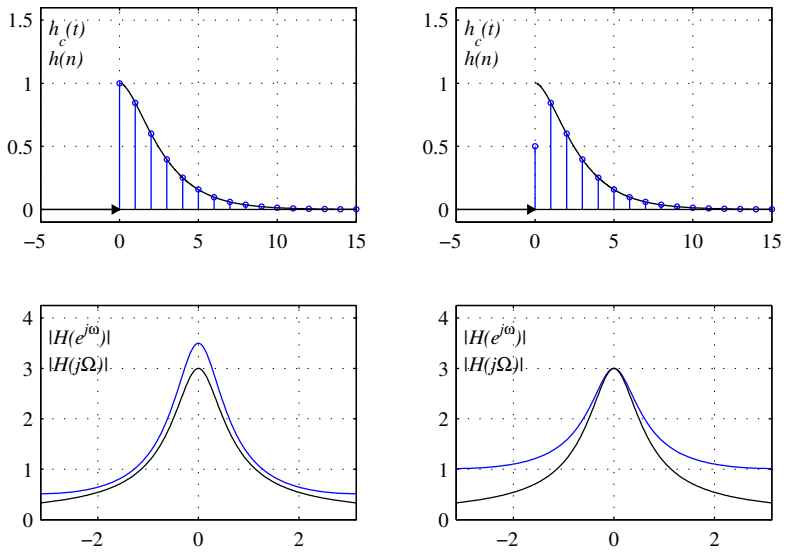


Figure 5.3 Impulse responses of systems in continuous and discrete-time domains (top). Amplitude of the frequency response of systems in continuous and discrete-time domains (bottom). System without discontinuity correction (left) and system with discontinuity correction (right).

★The transfer function $H(s)$ should expanded using partial fractions as

$$H(s) = \frac{1-3s/2}{6(s+\frac{1}{2})(s+\frac{1}{3})(s+1)^2} = \frac{k_1}{s+\frac{1}{2}} + \frac{k_2}{s+\frac{1}{3}} + \frac{k_3}{(s+1)^2} + \frac{k_4}{s+1}$$

with

$$k_1 = H(s)(s+1/2)|_{s=-1/2} = -7, \quad k_2 = 27/8, \quad \text{and} \quad k_3 = H(s)(s+1)^2\Big|_{s=-1} = 5/4.$$

The coefficient k_4 follows, for example, from

$$H(0) = 1 = 2k_1 + 3k_2 + k_3 + k_4,$$

as $k_4 = 29/8$. Thus, we get

$$H(s) = \frac{-7}{s+\frac{1}{2}} + \frac{27/8}{s+\frac{1}{3}} + \frac{5/4}{(s+1)^2} + \frac{29/8}{s+1}.$$

According to (5.5) and (5.6) the discrete-time system is

$$\begin{aligned} H(z) &= \frac{-7}{1-e^{-1/2}z^{-1}} + \frac{27/8}{1-e^{-1/3}z^{-1}} + \frac{d}{ds_i}\Big\{\frac{5/4}{1-e^{s_i}z^{-1}}\Big\}\Big|_{s_i=-1} + \frac{29/8}{1-e^{-1}z^{-1}} \\ &= \frac{-7z}{z-e^{-1/2}} + \frac{27z/8}{z-e^{-1/3}} + \frac{5e^{-1}z/4}{(z-e^{-1})^2} + \frac{29z/8}{z-e^{-1}}. \end{aligned}$$

Since $h(0) = \lim_{z\to\infty} H(z) = 0$ there no need to consider possible impulse response correction due to discontinuity.

When the transfer function is rearranged into the form

$$H(z) = -\frac{0.0341z(z-1.9894)(z+0.3259)}{(z-0.7165)(z-0.6065)(z-0.3679)^2}$$

we can easily see that the poles are mapped according to $s_{pi} \to e^{s_{pi}\Delta t}$, Fig. 5.4, while there is no direct correspondence among zeros of the transfer functions. The impulse responses of the continuous-time system and the discrete-time system are shown in Fig. 5.5.

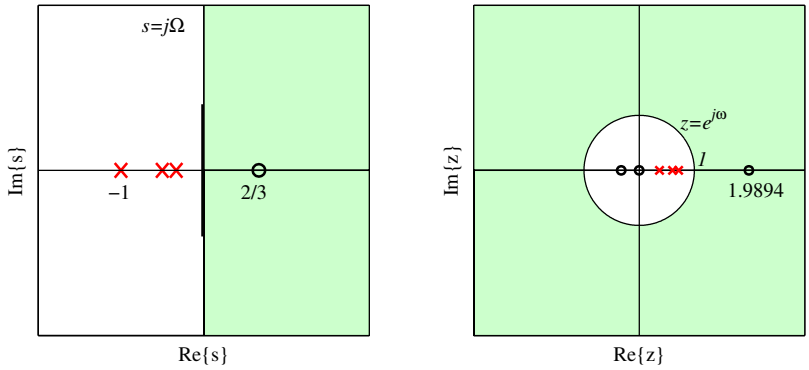


Figure 5.4 Pole-zero locations in the s-domain and the z-domain using the impulse invariance method.

5.2 MATCHED z-TRANSFORM METHOD

The matched z-transform method is based on a discrete-time approximation of the Laplace transform derived in the previous chapter as the starred transform (4.22)

$$X(s) = \int_{-\infty}^{\infty} x(t)e^{-st}dt \cong \sum_{n=-\infty}^{\infty} x(n)e^{-sn\Delta t} = X(z)|_{z=e^{s\Delta t}}.$$

This approximation leads to a relation between the Laplace domain and the z-domain in the form of

$$z = e^{s\Delta t}.$$

If we use this relation to map all zeros and poles of a continuous-time system transfer function

$$H(s) = \frac{b_M s^M + \cdots + b_1 s + b_0}{a_N s^N + \cdots + a_1 s + a_0} = \frac{b_M(s-s_{01})(s-s_{02})\ldots(s-s_{0M})}{a_N(s-s_{p1})(s-s_{p2})\ldots(s-s_{pN})}$$

into the corresponding z-plane locations, Fig. 5.6,

$$z_{0i} = e^{s_{0i}\Delta t}$$
$$z_{pi} = e^{s_{pi}\Delta t},$$

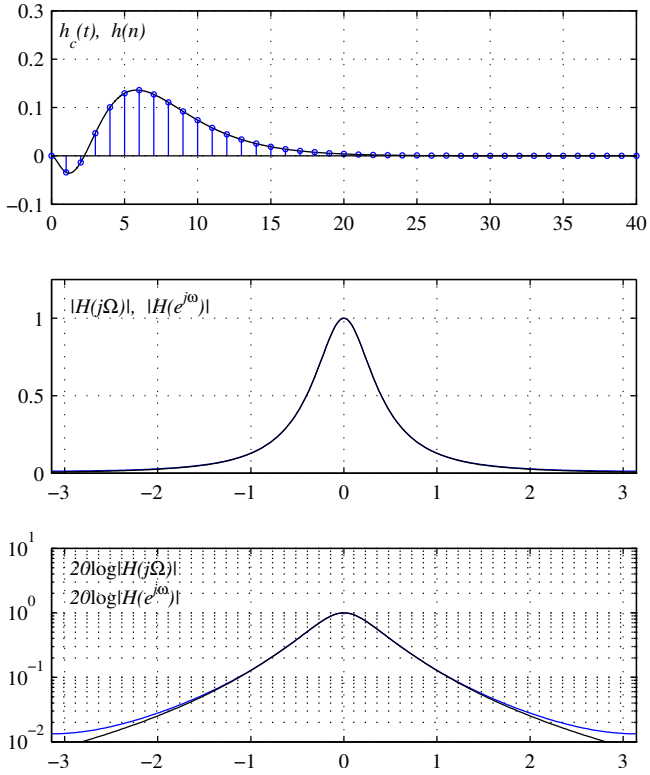


Figure 5.5 Impulse responses of systems in continuous and discrete-time domains (top). Amplitude of the frequency response of systems in continuous and discrete-time domains (middle). Amplitude of the frequency response of systems in continuous and discrete-time domains in logarithmic scale (bottom).

the matched z-transform method of the system follows. The discrete-time system transfer function is

$$H(z) = C\frac{(z - e^{s_{01}\Delta t})(z - e^{s_{02}\Delta t})\dots(z - e^{s_{0M}\Delta t})}{(z - e^{s_{p1}\Delta t})(z - e^{s_{p2}\Delta t})\dots(z - e^{s_{pN}\Delta t})}.$$

The constant C follows from the amplitude condition. For example, it can be calculated from $H(s)_{|s=0} = H(z)_{|z=1}$.

Example 5.3. For the continuous-time system with the transfer function

$$H(s) = \frac{1-s}{8s^2 + 6s + 1}$$

find the corresponding discrete-time system according to the matched z-transform method and $\Delta t = 1$?

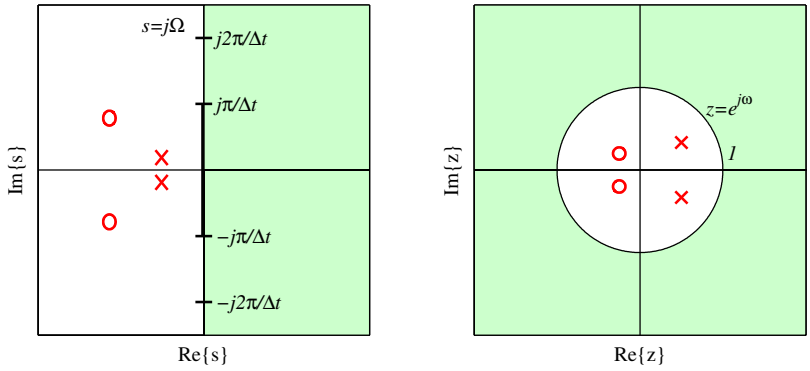


Figure 5.6 Illustration of the zeros and poles mapping in the matched $z-$transform method.

★The transfer function of the discrete-time system is obtained from

$$H(s) = \frac{1-s}{8(s+\frac{1}{2})(s+\frac{1}{4})},$$

using the matched method mapping, $z_{0i} = e^{s_{0i}\Delta t}$ and $z_{pi} = e^{s_{pi}\Delta t}$, as

$$H(z) = k\frac{z-e}{8(z-e^{-1/2})(z-e^{-1/4})}.$$

Since $H(s)|_{s=0} = 1$, if we want that $H(z)|_{z=e^{j0}} = 1$ then $k = -1/2.4678 = -0.4052$.

5.3 DIFFERENTIATION AND INTEGRATION

The first-order backward difference is a common method to approximate the first-order derivative of a continuous-time signal

$$y(t) = \frac{dx(t)}{dt}$$
$$y(n\Delta t) \cong \frac{x(n\Delta t) - x((n-1)\Delta t)}{\Delta t}.$$

The Laplace transform domain of the continuous-time first derivative is

$$Y(s) = sX(s). \tag{5.8}$$

In the discrete-time domain, with $y(n) = y(n\Delta t)\Delta t$ and $x(n) = x(n\Delta t)\Delta t$, the approximation of this derivative results in the first-order linear difference equation

$$y(n) = \frac{x(n) - x(n-1)}{\Delta t}.$$

In the z-transform domain this equation is of the form

$$Y(z) = \frac{1 - z^{-1}}{\Delta t} X(z). \tag{5.9}$$

Based on (5.8) and (5.9) we can conclude that the mapping of the corresponding differentiation operators from the continuous-time to the discrete-time domain is

$$s = \frac{1 - z^{-1}}{\Delta t}. \tag{5.10}$$

With a normalized discretization step $\Delta t = 1$ this mapping is of the form

$$s = 1 - z^{-1}.$$

The same result could be obtained by considering a rectangular rule approximation of a continuous-time integral, at an instant $t = n\Delta t$,

$$y(n\Delta t) = \int_{-\infty}^{n\Delta t} x(t)dt \cong \int_{-\infty}^{n\Delta t - \Delta t} x(t)dt + x(n\Delta t)\Delta t.$$

The value of this integral can be approximated as

$$y(n\Delta t) \cong y(n\Delta t - \Delta t) + x(n\Delta t)\Delta t.$$

In the discrete-time domain, this relation takes the form

$$y(n) = y(n-1) + x(n)\Delta t.$$

The Laplace and the z-transform domain forms of the previous integral equations are

$$Y(s) = \frac{1}{s} X(s)$$
$$Y(z) = \frac{\Delta t}{1 - z^{-1}} X(z).$$

The same mapping of the z-plane to the s-plane as in (5.10) follows.

Consider the imaginary axis from the s-plane (the Fourier transform line). According to (5.10) the mapping, with $\Delta t = 1$, is defined by

$$1 - s \to z^{-1}. \tag{5.11}$$

Now we will consider the region which corresponds to the imaginary axis and the left semi-plane of the s-domain (containing poles of a stable system), Fig. 5.7(left). The aim is to find the corresponding region in the z-domain.

If we start from the s-domain and the region in Fig. 5.7(left), the first mapping is to reverse the s-domain to $-s$ and shift it for $+1$, as

$$1 - s \to p.$$

The corresponding domain, after this mapping, is shown in Fig. 5.7(middle).

The next step is to map the region from p-domain into the z-domain, according to (5.11), as

$$p \to z^{-1}.$$

By denoting $\mathrm{Re}\{z\} = x$ and $\mathrm{Im}\{z\} = y$ we get that the line $\mathrm{Re}\{p\} = 1$ in the $p-$domain, corresponding to the imaginary axis in the s-plane, is transformed into the z-domain according to

$$\mathrm{Re}\{p\} = \mathrm{Re}\{\frac{1}{z}\}$$

$$1 = \mathrm{Re}\{\frac{1}{x+jy}\}$$

$$1 = \mathrm{Re}\{\frac{1}{x+jy}\frac{x-jy}{x-jy}\}$$

resulting in

$$1 = \frac{x}{x^2+y^2}$$

or in

$$(x-\frac{1}{2})^2 + y^2 = \left(\frac{1}{2}\right)^2. \tag{5.12}$$

Therefore, the imaginary axis in the s-plane is mapped onto a circle defined by (5.12), Fig. 5.7(right) in the z-plane. From the mapping relation $1-s \to z^{-1}$ it is easy to conclude that the origin $s = 0 + j0$ maps into $z = 1$ and that $s = 0 \pm j\infty$ maps into $z = \pm 0$, according to $1/(1-s) \to z$.

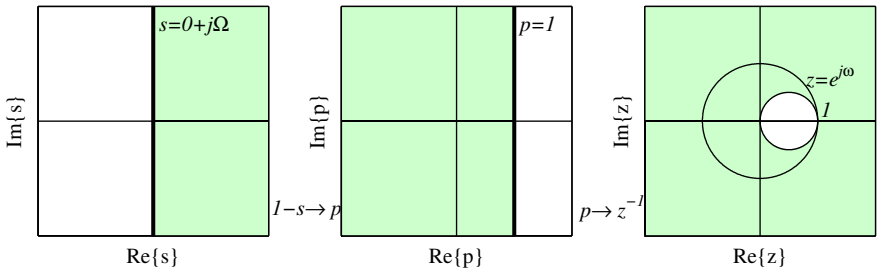


Figure 5.7 Illustration of the differentiation based mapping of the left $s-$semi-plane with the imaginary axis (left), translated and reversed $p-$domain (middle), and the $z-$domain (right).

Mapping of the imaginary axis into z-domain can also be analyzed from

$$\sigma + j\Omega \to \frac{1-(re^{j\omega})^{-1}}{\Delta t} = \frac{1-r^{-1}\cos\omega}{\Delta t} + j\frac{r^{-1}}{\Delta t}\sin\omega.$$

For $\sigma = 0$, follows

$$1 - r^{-1}\cos\omega = 0 \quad \text{or} \quad r = \cos\omega, \tag{5.13}$$

with

$$\Omega = \frac{r^{-1}}{\Delta t}\sin\omega = \frac{\tan\omega}{\Delta t}.$$

Obviously, $\omega = 0$ maps to $\Omega = 0$ (with $\Omega \cong \omega/\Delta t$ for small ω), and $\omega = \pm\pi/2$ maps into $\Omega \to \pm\infty$. Thus, the whole imaginary axis maps onto $-\pi/2 \le \omega \le \pi/2$. These values of ω could be used within the basic period. Relation (5.13), with $-\pi/2 \le \omega \le \pi/2$, is a circle defined by (5.12) if we replace $r = \sqrt{x^2+y^2}$ and $\cos\omega = x/\sqrt{x^2+y^2}$ with $\sigma < 0$ (semi-plane with negative real values) being mapped into $r < \cos\omega$ (interior of unit circle).

Example 5.4. A continuous-time system is described by the differential equation

$$y''(t) + \frac{3}{4}y'(t) + \frac{1}{8}y(t) = x(t),$$

with the zero initial conditions and the transfer function

$$H(s) = \frac{1}{s^2 + \frac{3}{4}s + \frac{1}{8}}.$$

What is the corresponding transfer function of the discrete-time system using the first-order backward difference approximation with $\Delta t = 1/2$? What is the solution to the differential equation for $x(t) = u(t)$. Compare it with the solution to difference equation $y(n)$ with $\Delta t = 1/8$.

★The discrete-time system transfer function is obtained using $s = \left(1 - z^{-1}\right)/\Delta t$ in $H(s)$ as

$$\begin{aligned} H(z) &= \frac{1}{\left(\frac{1-z^{-1}}{\Delta t}\right)^2 + \frac{3}{4}\frac{1-z^{-1}}{\Delta t} + \frac{1}{8}} \\ &= \frac{(\Delta t)^2}{1 + \frac{3}{4}\Delta t + \frac{1}{8}(\Delta t)^2 - [2 + \frac{3}{4}\Delta t]z^{-1} + z^{-2}} \end{aligned}$$

with

$$\begin{aligned} y(n) &= B_0 x(n) + A_1 y(n-1) + A_2 y(n-2) \\ B_0 &= \frac{(\Delta t)^2}{1 + \frac{3}{4}\Delta t + \frac{1}{8}(\Delta t)^2} = 0.1778 \\ A_1 &= \frac{[2 + \frac{3}{4}\Delta t]}{1 + \frac{3}{4}\Delta t + \frac{1}{8}(\Delta t)^2} = 1.6889 \\ A_2 &= -\frac{1}{1 + \frac{3}{4}\Delta t + \frac{1}{8}(\Delta t)^2} = -0.7111, \end{aligned}$$

where $\Delta t = 1/2$. For $x(t) = u(t)$, the continuous-time output signal is obtained from

$$\begin{aligned} Y(s) = H(s)X(s) &= \frac{1}{s(s^2 + \frac{3}{4}s + \frac{1}{8})} \\ &= \frac{8}{s} + \frac{8}{s + \frac{1}{2}} - \frac{16}{s + \frac{1}{4}} \end{aligned}$$

as

$$y(t) = [8 + 8e^{-t/2} - 16e^{-t/4}]u(t).$$

The results of the difference equation for $y(n)$ are compared with the exact solution $y(t)$ in Fig. 5.8. The agreement is high. It could be additionally improved by reducing the sampling interval, for example, to $\Delta t = 1/8$.

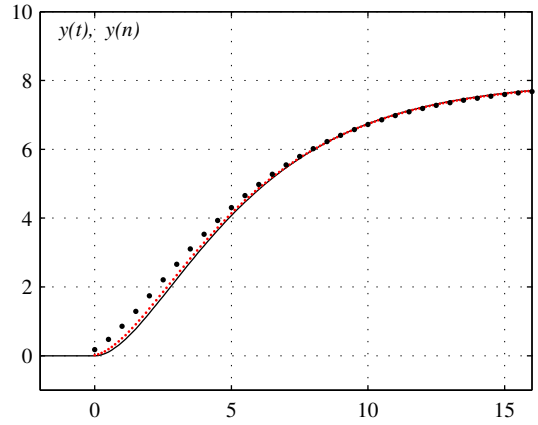


Figure 5.8 The exact solution to the differential equation for $y(t)$, in solid line, and the discrete-time system output $y(n)$, in large dots for $\Delta t = 1/2$ and in small dots for $\Delta t = 1/8$.

5.4 BILINEAR TRANSFORM

In the case of a differentiator based mapping, the imaginary axis in the $s-$domain, corresponding to the Fourier transform values, has been mapped onto a circle with radius $1/2$ and the center at $z = 1/2$ in the $z-$domain, as shown in Fig. 5.7. This mapping does not correspond to the Fourier transform of discrete-time signals position in the $z-$plane, which is along the circle line $|z| = 1$. A transformation that will map the imaginary axis from the $s-$domain onto the unit circle in the $z-$domain is presented next.

Consider numerical integration in the case of the first-order system (for example, the charge on a capacitor), using the trapezoid rule

$$y(n\Delta t) = \int_{-\infty}^{n\Delta t} x(t)dt \cong \int_{-\infty}^{n\Delta t-\Delta t} x(t)dt + \frac{x(n\Delta t) + x((n-1)\Delta t)}{2}\Delta t$$

$$y(n) = y(n-1) + \frac{x(n) + x(n-1)}{2}\Delta t.$$

In the Laplace and the z-transform domain, these relations have the following forms

$$Y(s) = \frac{1}{s}X(s)$$

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

The mapping from the $s-$domain to the $z-$domain is defined here by

$$s \to \frac{2}{\Delta t}\frac{1-z^{-1}}{1+z^{-1}}. \tag{5.14}$$

In complex analysis this mapping is known as the *bilinear transform*.

Now we can repeat the transformation of the continuous-time system, $H(s)$, from Example 5.1 to get the discrete-time system, $H(z)$, by replacing s with $2(1-z^{-1})/(1+z^{-1})$ in (5.7).

Within the derivatives framework, the bilinear transform can be understood as the following derivative approximation. Consider the first-order backward derivative approximation

$$y(n) = x(n) - x(n-1).$$

The same signal samples can used for the first-order forward derivative approximation

$$y(n-1) = x(n) - x(n-1).$$

If we assume that the difference $x(n) - x(n-1)$ fits better to the mean of $y(n)$ and $y(n-1)$ than to any single one of them, then the derivative approximation using the difference equation

$$\frac{y(n) + y(n-1)}{2} = x(n) - x(n-1),$$

produces the bilinear transform.

In order to prove that *the unit circle in the $z-$domain maps onto the imaginary axis in the $s-$domain* we may simply replace $z = e^{j\omega}$ into (5.14) and obtain

$$2\frac{1 - e^{-j\omega}}{1 + e^{-j\omega}} = 2\frac{e^{j\omega/2} - e^{-j\omega/2}}{e^{j\omega/2} + e^{-j\omega/2}} = 2j\tan(\frac{\omega}{2}) \to s\Delta t.$$

For $s = \sigma + j\Omega$, follows

$$\sigma = 0$$
$$\Omega = \frac{2}{\Delta t}\tan(\frac{\omega}{2}).$$

Therefore, the unit circle $z = e^{j\omega}$ maps onto the imaginary axis $\sigma = 0$. The frequency points $\omega = 0$ and $\omega = \pm\pi$ map into $\Omega = 0$ and $\Omega \to \pm\infty$, respectively.

The linearity of the frequency mapping $\Omega \to \omega$ is lost. It holds for small values of ω only

$$\Omega = \frac{2}{\Delta t}\tan(\frac{\omega}{2}) \cong \frac{\omega}{\Delta t}, \text{ for } |\omega| \ll 1.$$

From

$$z = \frac{1 + \frac{s\Delta t}{2}}{1 - \frac{s\Delta t}{2}} \quad \text{and} \quad |z| = \frac{\sqrt{(1 + \frac{\sigma\Delta t}{2})^2 + (\frac{\Omega\Delta t}{2})^2}}{\sqrt{(1 - \frac{\sigma\Delta t}{2})^2 + (\frac{\Omega\Delta t}{2})^2}}$$

it may easily be concluded that $\sigma < 0$ maps into $|z| < 1$, since $1 + \frac{\sigma\Delta t}{2} < 1 - \frac{\sigma\Delta t}{2}$ for $\sigma < 0$.

The bilinear transform mapping can be derived using a series of complex plane mappings. Since

$$z = \frac{1 + \frac{s\Delta t}{2}}{1 - \frac{s\Delta t}{2}} = \frac{2}{1 - \frac{s\Delta t}{2}} - 1,$$

we can write

$$1 - \frac{s\Delta t}{2} \to p_1, \quad \frac{1}{p_1} \to p_2, \quad \text{and} \quad 2p_2 - 1 \to z.$$

This series of mappings from the s-domain to the z-domain is illustrated in Fig. 5.9, with $\Delta t = 1$. The fact that $\frac{1}{p_1} \to p_2$ maps the line $\mathrm{Re}\{p_1\} = 1$ onto the circle $(x - \frac{1}{2})^2 + y^2 = \left(\frac{1}{2}\right)^2$ in p_2-domain is proven in the previous section.

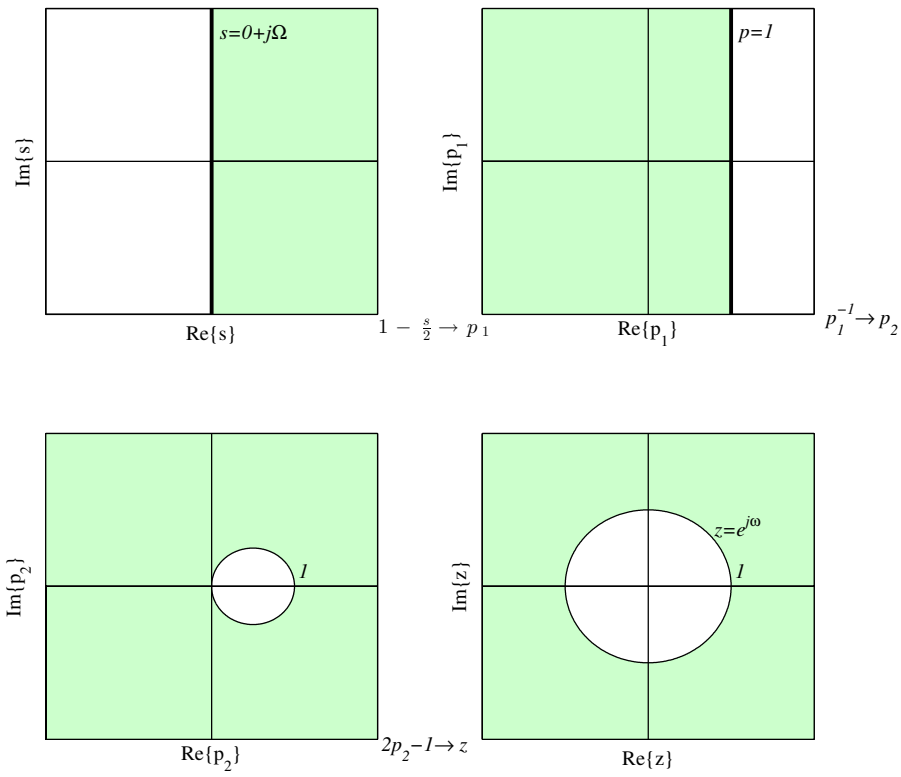


Figure 5.9 Bilinear mapping illustration trough a series of elementary complex plane mappings.

Since the bilinear transform introduces a nonlinear transformation of the frequency axis from the continuous-time domain to the discrete-time domain, $\Omega = \frac{2}{\Delta t}\tan(\frac{\omega}{2})$, this nonlinearity must be compensated during the system design. Usually it is done by pre-modifying the desired important frequency values Ω_c from the analog domain using $\Omega_d = \frac{2}{\Delta t}\tan(\frac{\omega_c}{2})$, and $\omega_c = \Omega_c \Delta t$. The new continuous-time domain frequencies Ω_d will be returned back to the desired values ω_c and $\Omega_c = \omega_c/\Delta t$ after the bilinear transformation.

Example 5.5. A continuous-time system

$$H(s) = \frac{2Q\Omega_1}{s^2 + 2\Omega_1 Qs + \Omega_1^2 + Q^2} + \frac{2Q\Omega_2}{s^2 + 2\Omega_2 Qs + \Omega_2^2 + Q^2}$$

is designed to pass the signal

$$x(t) = A_1\cos(\Omega_1 t + \varphi_1) + A_2\cos(\Omega_2 t + \varphi_2).$$

and to stop all other possible signal components. The parameters are $Q = 0.01$, $\Omega_1 = \pi/4$, and $\Omega_2 = 3\pi/5$. The signal is sampled with $\Delta t = 1$ and the discrete-time signal $x(n)$ is formed. Using

the bilinear transform, design the discrete-time system that corresponds to the continuous-time system with the transfer function $H(s)$.

★For the beginning just use the bilinear transform relation

$$s \to 2\frac{1-z^{-1}}{1+z^{-1}} \tag{5.15}$$

and map $H(s)$ to $H_B(z)$ without any pre-modification. The result is presented in the first two subplots of Fig. 5.10. The discrete frequencies are shifted since the bilinear transform (5.15) made a nonlinear frequency mapping from the continuous-time to discrete-time domain, according to

$$\Omega = 2\tan(\frac{\omega}{2}).$$

Thus, obviously, the system $H_B(z)$ is not a system that will filter the corresponding frequencies in $x(n)$ in the same way as $H(s)$ filters $x(t)$.

In order to correct the shift introduced by the bilinear transform mapping, the continuous-time system should be pre-modified as

$$H_d(s) = \frac{2Q\Omega_{1d}}{s^2 + 2\Omega_{1d}Qs + \Omega_{1d}^2 + Q^2} + \frac{2Q\Omega_{2d}}{s^2 + 2\Omega_{2d}Qs + \Omega_{2d}^2 + Q^2}$$

with

$$\Omega_{1d} = \frac{2}{\Delta t}\tan(\frac{\Omega_1 \Delta t}{2}) = 0.8284 = 0.2637\pi$$
$$\Omega_{2d} = \frac{2}{\Delta t}\tan(\frac{\Omega_2 \Delta t}{2}) = 2.7528 = 0.8762\pi.$$

We see that the shift of $\Omega_1 = 0.25\pi$ to $\Omega_{1d} = 0.2637\pi$ is small since the bilinear transform frequency mapping for small frequency values is almost linear. However, for $\Omega_2 = 0.6\pi$, the shift to $\Omega_{2d} = 0.8762\pi$ is significant due to a high nonlinearity of mapping in that region. The modified system $H_d(s)$ is presented in subplot 3 of Fig. 5.10. Next, using the bilinear transform mapping $s \to 2\frac{1-z^{-1}}{1+z^{-1}}$ the modified frequencies will map to the desired ones $\omega_1 = \Omega_1 \Delta t$ and $\omega_2 = \Omega_2 \Delta t$. The obtained discrete-time system transfer function is of the form

$$H(z) = \frac{2Q\Omega_{1d}}{\left(2\frac{1-z^{-1}}{1+z^{-1}}\right)^2 + 4\Omega_{1d}Q\frac{1-z^{-1}}{1+z^{-1}} + \Omega_{1d}^2 + Q^2} + \frac{2Q\Omega_{2d}}{\left(2\frac{1-z^{-1}}{1+z^{-1}}\right)^2 + 4\Omega_{2d}Q\frac{1-z^{-1}}{1+z^{-1}} + \Omega_{2d}^2 + Q^2}.$$

For the given values of Q, Ω_{21}, and Ω_{2d}, we get

$$H(z) = \frac{0.016569}{\left(2\frac{1-z^{-1}}{1+z^{-1}}\right)^2 + 0.0331375\frac{1-z^{-1}}{1+z^{-1}} + 0.68641} + \frac{0.0551}{\left(2\frac{1-z^{-1}}{1+z^{-1}}\right)^2 + 0.1101\frac{1-z^{-1}}{1+z^{-1}} + 7.5778}.$$

When the expression for $H(z)$ is appropriately rearranged, its final form is given by

$$H(z) = \frac{0.016569(1+z^{-1})^2}{4.65327z^{-2} - 6.6272z^{-1} + 4.7195} + \frac{0.0551(1+z^{-1})^2}{11.4677z^{-2} + 7.1556z^{-1} + 11.6879} =$$
$$\frac{0.003567(1+z^{-1})^2}{(z^{-1} - 1.0071e^{j0.25\pi})(z^{-1} - 1.0071e^{-j0.25\pi})} + \frac{0.0048(1+z^{-1})^2}{(z^{-1} - 1.0096e^{j0.6\pi})(z^{-1} - 1.0096e^{-j0.6\pi})}.$$

The frequency response of this system is shown in panel 4 of Fig. 5.10. This is the desired discrete-time system corresponding to the continuous-time system in panel 1 of this figure. In calculations the coefficients are rounded to four decimal places.

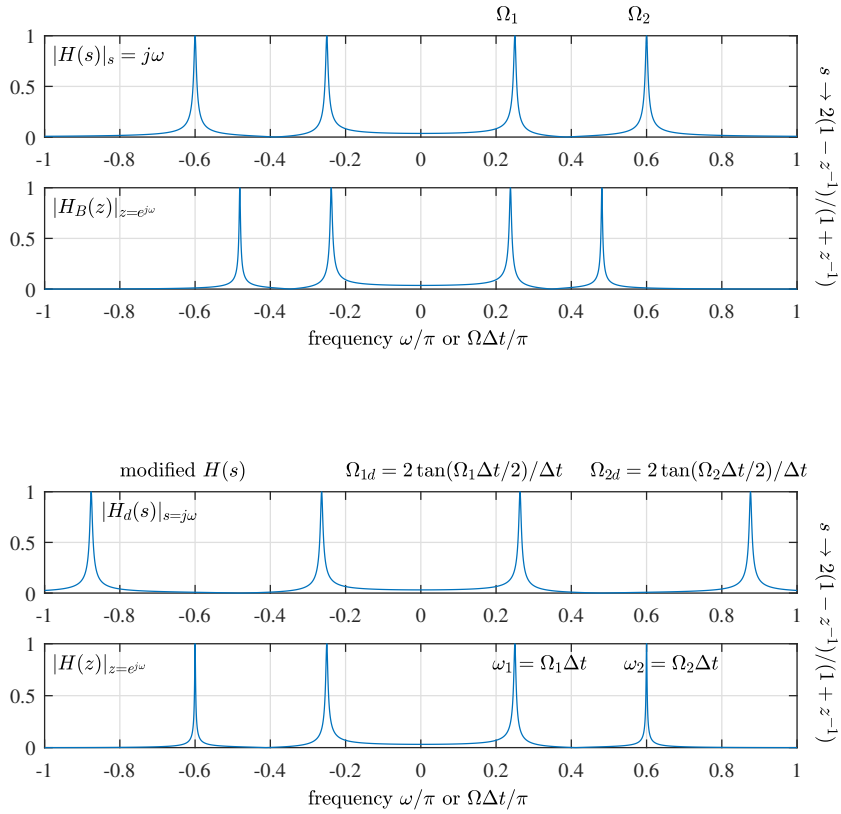


Figure 5.10 Amplitude of the continuous-time system with the transfer function $H(s)$ and the amplitude of the transfer function $H_B(z)$ of the discrete-time system obtained by the bilinear transform (first two panels). A premodified system to take into account the nonlinearity of the frequency mapping in the bilinear transform, $H_d(s)$, and the amplitude of the transfer function $H(z)$ of the discrete-time system obtained by the bilinear transform of $H_d(s)$ (last two panels).

Comparison of the mapping methods presented in this section is summarized in the next table.

Method	Fourier transform $H(s)\|_{s=j\Omega} \to H(z)\|_{z=e^{j\omega}}$	Sampling theorem condition
Impulse Invariance	Yes, $\Omega = \omega/\Delta t$	Yes
Matched z-transform	No	No
First-oder difference	No	No
Bilinear transform	Yes, $\Omega = \frac{\tan(\omega/2)}{\Delta t/2}$	No

5.5 DISCRETE FILTERS DESIGN

The digital filter design will be explained here. The lowpass filter is assumed as the basic filter form, while the other filters (highpass and bandpass) are designed by modifying the system that corresponds to the discrete-time lowpass filter. In the examples, the lowpass Butherworth filters will be used.

5.5.1 Lowpass filters

An ideal discrete lowpass filter is defined by the frequency response

$$H(e^{j\omega}) = \begin{cases} 1 & \text{for } |\omega| < \omega_c \\ 0 & \text{for } \omega_c < |\omega| < \pi \end{cases} .$$

The frequency response is periodic in ω with period 2π.

The implementation of an ideal lowpass filter in the DFT domain is obvious, by multiplying all DFT coefficients corresponding to frequency $\omega_c < |\omega| < \pi$ by zero. For on-line implementations in the discrete-time domain, the ideal filter should be approximated by a corresponding transfer function form that can be implemented, since the impulse response of the ideal filter

$$h(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{j\omega n} d\omega = 2\frac{\sin(\omega_c n)}{n}$$

is noncausal.

There are several methods to approximate the ideal lowpass filter frequency response. One of them is the Butterworth approximation. Some of commonly used approximations are Chebyshev and elliptic forms as well.

A lowpass filter of the Butterworth type is shown in Fig. 5.11, along with the ideal one.

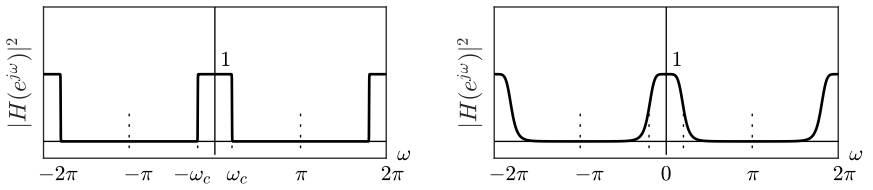


Figure 5.11 Lowpass filter frequency response: ideal case (left) and Butterworth type (right).

Example 5.6. Implement the Butterworth discrete-time filter of the order $N = 4$ with a critical frequency corresponding to the continuous-time domain filter with the critical frequency $f_c = 4[kHz]$ and the sampling interval $\Delta t = 31.25[\mu\text{sec}]$, using:
(a) The impulse invariance method and
(b) the bilinear transform.

★The discrete-time frequency is $\omega_c = \Omega_c \Delta t = 2\pi f_c \Delta t = \pi/4$.

The poles of the fourth-order Butterworth filter in the continuous-time domain (Chapter I, Subsection 1.6) are

$$s_0 = \Omega_c\left[\cos(\frac{\pi}{2}+\frac{\pi}{8})+j\sin(\frac{\pi}{2}+\frac{\pi}{8})\right] = \Omega_c(-0.3827+j0.9239)$$
$$s_1 = \Omega_c\left[\cos(\frac{\pi}{2}+\frac{3\pi}{8})+j\sin(\frac{\pi}{2}+\frac{3\pi}{8})\right] = \Omega_c(-0.9239+j0.3827)$$
$$s_2 = \Omega_c\left[\cos(\frac{\pi}{2}+\frac{5\pi}{8})+j\sin(\frac{\pi}{2}+\frac{5\pi}{8})\right] = \Omega_c(-0.9239-j0.3827)$$
$$s_3 = \Omega_c\left[\cos(\frac{\pi}{2}+\frac{7\pi}{8})+j\sin(\frac{\pi}{2}+\frac{7\pi}{8})\right] = \Omega_c(-0.3827-j0.9239).$$

The transfer function of the filter in the Laplace domain is

$$H(s) = \frac{\Omega_c^4}{(s^2+0.7654\Omega_c s+\Omega_c^2)(s^2+1.8478\Omega_c s+\Omega_c^2)}. \tag{5.16}$$

(a) For the impulse invariance method the transfer function (5.16) should be expanded into partial fractions,

$$H(s) = \frac{k_0}{s-s_0}+\frac{k_1}{s-s_1}+\frac{k_2}{s-s_2}+\frac{k_3}{s-s_3},$$

with the constants k_i calculated based on $k_i = H(s)(s-s_i)_{|s=s_i}$ as

$$k_0 = (-0.3628+j0.1503)/\Delta t,$$
$$k_1 = (0.3628-j0.8758)/\Delta t,$$
$$k_2 = (0.3628+j0.8758)/\Delta t,$$
$$k_3 = (-0.3628-j0.1503)/\Delta t.$$

Using the impulse invariance method we get the transfer function of the Butterworth filter

$$\begin{aligned} H(z) &= \frac{k_0\Delta t}{1-e^{s_0\Delta t}z^{-1}}+\frac{k_1\Delta t}{1-e^{s_1\Delta t}z^{-1}}+\frac{k_2\Delta t}{1-e^{s_2\Delta t}z^{-1}}+\frac{k_3\Delta t}{1-e^{s_3\Delta t}z^{-1}} \\ &= \frac{-0.3628+j0.1503}{1-e^{\omega_c(-0.3827+j0.9239)}z^{-1}}+\frac{0.3628-j0.8758}{1-e^{\omega_c(-0.9239+j0.3827)}z^{-1}} \\ &+\frac{0.3628+j0.8758}{1-e^{\omega_c(-0.9239-j0.3827)}z^{-1}}+\frac{-0.3628-j0.1503}{1-e^{\omega_c(-0.3827-j0.9239)}z^{-1}}. \end{aligned}$$

It can be seen that the discrete-time filter is a function of ω_c. Thus, for a given continuous domain frequencies and the sampling interval Δt, it is possible to calculate the corresponding discrete-time frequency $\omega_c = \Omega_c\Delta t$ and to use this frequency in the filter design with the normalized $\Delta t = 1$. Using the value of the critical frequency, $\omega_c = \pi/4$, we get

$$\begin{aligned} H(z) &= \frac{-0.3628+j0.1503}{1-(0.5539+j0.4913)z^{-1}}+\frac{-0.3628-j0.1503}{1-(0.5539-j0.4913)z^{-1}} \\ &+\frac{0.3628-j0.8758}{1-(0.4623+j0.1433)z^{-1}}+\frac{0.3628+j0.8758}{1-(0.4623-j0.1433)z^{-1}}. \end{aligned}$$

A system form with the real-valued coefficients is obtained by grouping the complex-conjugate terms,

$$H(z) = \frac{-0.7256+0.2542z^{-1}}{1-1.1078z^{-1}+0.5482z^{-2}}+\frac{0.7256-0.084z^{-1}}{1-0.9246z^{-1}+0.2343z^{-2}}.$$

(b) For the bilinear transform, the critical frequency ω_c has to be pre-modified according to

$$\Omega_d = \frac{2}{\Delta t}\tan(\frac{\omega_c}{2}) = \frac{0.8284}{\Delta t}.$$

Then, the frequency Ω_d is used for the design in (5.16), instead of Ω_c. The frequency Ω_d will be transformed back to $\Omega_c = \omega_c/\Delta t$, after the bilinear transform is used. Using the substitutions $s \to \frac{2}{\Delta t}\frac{1-z^{-1}}{1+z^{-1}}$ and

$$\omega_d = \Omega_d \Delta t = 0.8284$$

in (5.16), the filter transfer function follows as

$$\begin{aligned}
H(z) &= \frac{{\omega_d}^4}{[4(\frac{1-z^{-1}}{1+z^{-1}})^2 + 2\omega_d 0.7654\frac{1-z^{-1}}{1+z^{-1}} + \omega_d^2][4(\frac{1-z^{-1}}{1+z^{-1}})^2 + 2\omega_d 1.8478\frac{1-z^{-1}}{1+z^{-1}} + \omega_d^2]} \\
&= \frac{0.4710}{[4(\frac{1-z^{-1}}{1+z^{-1}})^2 + 1.2626\frac{1-z^{-1}}{1+z^{-1}} + 0.6863][4(\frac{1-z^{-1}}{1+z^{-1}})^2 + 3.0481\frac{1-z^{-1}}{1+z^{-1}} + 0.6863]} \\
&= \frac{0.4710\left(1+z^{-1}\right)^4}{\left(3.4237z^{-2} - 6.6274z^{-1} + 5.9484\right)\left(1.6382z^{-2} - 6.6274z^{-1} + 7.7704\right)} \\
&= \frac{0.084\left(1+z^{-1}\right)^4}{\left(z^{-2} - 1.9357z^{-1} + 1.7343\right)\left(z^{-2} - 4.0455z^{-1} + 4.7433\right)} \\
&= \frac{0.084z^{-4} + 0.336z^{-3} + 0.504z^{-2} + 0.336z^{-1} + 0.084}{z^{-4} - 5.9810z^{-3} + 14.3z^{-2} - 16.1977z^{-1} + 8.2263}.
\end{aligned}$$

The transfer function (amplitude and phase) of the continuous-time filter and the discrete-time filters, obtained using the impulse invariance method and the bilinear transform, are presented in Fig. 5.12, within one period in frequency. The agreement between the amplitude and the phase functions is high. The difference equation describing this Butterworth filter is

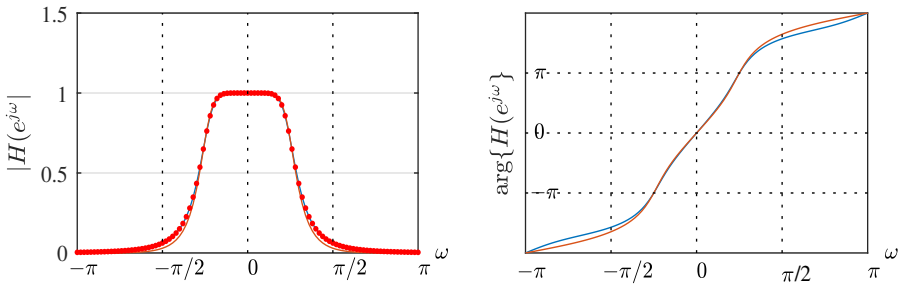


Figure 5.12 Amplitude and phase of the fourth-order Butterworth filter frequency response, obtained using the impulse invariance method and bilinear transform.

$$\begin{aligned}
y(n) = 1.969y(n-1) - 1.7383y(n-2) + 0.7271y(n-3) - 0.1216y(n-4) \\
+0.0102x(n) + 0.0408x(n-1) + 0.0613x(n-2) \\
+0.0408x(n-3) + 0.0102x(n-4).
\end{aligned}$$

In calculations, the coefficients are normalized by 8.2263 and rounded to four decimal places. Rounding may cause small quantization errors (that will be discussed within the next chapter).

Example 5.7. Design a continuous-time lowpass filter whose parameters are:
- passband frequency $\Omega_p = 2\pi f_p$, $f_p = 3$ kHz,
- stopband frequency $\Omega_s = 2\pi f_s$, $f_s = 6$ kHz,
- maximum attenuation in the passband $a_p = -2$ dB, and
- minimum attenuation in the stopband $a_s = -15.5$ dB.

Find the corresponding discrete-time filter using the bilinear transform and $\Delta t = 50[\mu\,\text{sec}]$.

★The maximum attenuation in the passband and the minimum attenuation in the stopband are

$$a_p = 20\log(A_p)$$
$$A_p = 10^{a_p/20} = 0.7943$$
$$A_s = 10^{a_s/20} = 0.1679.$$

The relations for the filter order N and the critical frequency Ω_c are (Chapter I, Subsection 1.6)

$$\frac{1}{1+\left(\frac{\Omega_p}{\Omega_c}\right)^{2N}} \geq A_p^2, \qquad \frac{1}{1+\left(\frac{\Omega_s}{\Omega_c}\right)^{2N}} \leq A_s^2. \tag{5.17}$$

Using the equality in both of these relations, the value of N follows from

$$N = \frac{1}{2}\frac{\ln \frac{\frac{1}{A_p^2}-1}{\frac{1}{A_s^2}-1}}{\ln \frac{\Omega_p}{\Omega_s}} = 2.9407. \tag{5.18}$$

The first greater integer is assumed for the filter order,

$$N = 3.$$

We can use any of the relations in (5.17) with the equality sign in order to calculate Ω_c. For the first relation, the value of Ω_c will be such that $\left|H(j\Omega_p)\right|^2 = A_p^2$ is satisfied. Then,

$$\Omega_c = \frac{\Omega_p}{\sqrt[2N]{\frac{1}{A_p^2}-1}} = 2\pi \times 3.2805 \text{ kHz},$$
$$\omega_c = \Omega_c \Delta t = 1.0306.$$

The poles of the Butterworth filter in the continuous-time domain are

$$s_k = \Omega_c e^{j(2\pi k+\pi)/6 + j\pi/2}, \quad k = 0,1,2$$
$$s_0 = 2\pi \times 3.2805\left[\cos\frac{2\pi}{3} + j\sin\frac{2\pi}{3}\right] \times 10^3$$
$$s_1 = -2\pi \times 3.2805 \times 10^3$$
$$s_2 = 2\pi \times 3.2805\left[\cos\frac{2\pi}{3} - j\sin\frac{2\pi}{3}\right] \times 10^3.$$

The transfer function of the Butterworth filter is

$$H(s) = \frac{(2\pi 3.2805 \times 10^3)^3}{(s + 2\pi 3.2805 \times 10^3)(s^2 + 2\pi 3.2805 \times 10^3 s + (2\pi 3.2805 \times 10^3)^2)}.$$

In the discrete-time filter design using the bilinear transform we should not use this transfer function. For the bilinear transform we have to pre-modify the frequencies Ω_p and Ω_s so that they are returned back to the specified values when the bilinear transform is applied. These frequencies are

$$\Omega_{dp} = \frac{2}{\Delta t}\tan(\frac{\Omega_p \Delta t}{2}) = 2\pi \times 3.2437 \text{ kHz}$$
$$\Omega_{ds} = \frac{2}{\Delta t}\tan(\frac{\Omega_s \Delta t}{2}) = 2\pi \times 8.7623 \text{ kHz}$$

with $N = 2.0512$, which follows from (5.18). Assuming $N = 3$, we get $\Omega_{dc} = 2\pi \times 3.5470$ kHz with $\omega_{dc} = 1.1143$. The modified transfer function in the continuous-time domain is

$$H_d(s) = \frac{(2\pi 3.5470 \times 10^3)^3}{(s + 2\pi 3.5470 \times 10^3)(s^2 + 2\pi 3.5470 \times 10^3 s + (2\pi 3.5470 \times 10^3)^2)}.$$

The discrete-time Butterworth filter transfer function $H(z)$ follows when the substitution

$$s = \frac{2}{\Delta t}\frac{1 - z^{-1}}{1 + z^{-1}}$$

is performed. This filter is of the form

$$\begin{aligned} H(z) &= \frac{1.1143^3}{(2\frac{1-z^{-1}}{1+z^{-1}} + 1.1143)(\left(2\frac{1-z^{-1}}{1+z^{-1}}\right)^2 + 1.1143\frac{1-z^{-1}}{1+z^{-1}} + 1.1143^2)} \\ &= \frac{0.0595(1 + z^{-1})^3}{1 - 1.0229z^{-1} + 0.6133z^{-2} - 0.1147z^{-3}}. \end{aligned}$$

The corresponding difference equation of this filter is

$$\begin{aligned} y(n) &= 1.0229y(n-1) - 0.6133y(n-2) + 0.1147y(n-3) \\ &\quad + 0.0595x(n) + 0.1784x(n-1) + 0.1784x(n-2) + 0.0595x(n-3). \end{aligned}$$

Example 5.8. The continuous-time signal

$$x(t) = 8\cos(\frac{22\pi}{3}t) + 4\sin(\pi t) + 4\cos(\frac{8\pi}{3}t + \frac{\pi}{4})$$

is sampled with $\Delta t = 1/4$. The discrete-time signal is passed through the ideal lowpass filter with the frequency $\omega_c = \pi/3$. Find the output signal. What is the corresponding continuous-time output signal?

★The discrete-time signal is

$$x(n) = 2\cos(\frac{11\pi}{6}n) + \sin(\frac{\pi}{4}n) + \cos(\frac{2\pi}{3}n + \frac{\pi}{4}).$$

Its Fourier transform is given by

$$\begin{aligned} X(e^{j\omega}) = {} & 2\pi \sum_{k=-\infty}^{\infty} [\delta(\omega - \frac{11\pi}{6} + 2k\pi) + \delta(\omega + \frac{11\pi}{6} + 2k\pi)] \\ & + \frac{\pi}{j} \sum_{k=-\infty}^{\infty} [\delta(\omega - \frac{\pi}{4} + 2k\pi) - \delta(\omega + \frac{\pi}{4} + 2k\pi)] \\ & + \pi \sum_{k=-\infty}^{\infty} [\delta(\omega - \frac{2\pi}{3} + 2k\pi)e^{j\pi/4} + \delta(\omega + \frac{2\pi}{3} + 2k\pi)e^{-j\pi/4}]. \end{aligned}$$

The Fourier transform value, within the basic period $-\pi \leq \omega \leq \pi$, is is

$$\begin{aligned} X(e^{j\omega}) = {} & 2\pi[\delta(\omega - \frac{11\pi}{6} + 2\pi) + \delta(\omega + \frac{11\pi}{6} - 2\pi)] \\ & + \frac{\pi}{j}[\delta(\omega - \frac{\pi}{4}) - \delta(\omega + \frac{\pi}{4})] + \pi[\delta(\omega - \frac{2\pi}{3})e^{j\pi/4} + \delta(\omega + \frac{2\pi}{3})e^{-j\pi/4}]. \end{aligned}$$

In addition to the last two components that have frequencies corresponding to the analog signal there is the first component

$$2\pi[\delta(\omega - \frac{11\pi}{6} + \frac{12\pi}{6}) + \delta(\omega + \frac{11\pi}{6} - \frac{12\pi}{6})]$$

that corresponds to

$$x_1(n) = 2\cos(\frac{\pi}{6}n).$$

The lowpass filter output is

$$y(n) = 2\cos(\frac{\pi}{6}n) + \sin(\frac{\pi}{4}n).$$

It corresponds to the continuous-time signal

$$y(t) = 8\cos(\frac{\pi}{6}t) + 4\sin(\pi t).$$

One component at the frequency $\omega = 2\pi/3 > \pi/3$ is filtered out. The component at $\omega = \pi/4$ is unchanged. One more component has appeared at the frequency $\omega = \pi/6$ due to the periodic extension of the Fourier transform of the discrete-time signal.

In general, a signal component $x(t) = \exp(j\Omega_0 t)$, $\Omega_0 < 0$, with a sampling interval Δt such that

$$K\pi \leq \Omega_0 \Delta t < (K+1)\pi$$

will, after sampling, result into a component within the basic period of the Fourier transform of the discrete-time signal, corresponding to the continuous signal at $\exp(j(\Omega_0 t - \frac{K}{\Delta t}\pi t)$. This effect is known as aliasing. The most obvious visual effect is when a wheel rotating with $f_0 = 25$ [Hz], $\Omega_0 = 50\pi$, is sampled in a video sequence at $\Delta t = 1/50$ [sec]. Then $\Omega_0 \Delta t = \pi$ corresponds to $\exp(j(\Omega_0 t - 50\pi t)) = e^{j0}$, that is, the wheel looks as it were static (nonmoving) object.

5.5.2 Highpass Filters

Highpass filters can be obtained by transforming the corresponding continuous-time filters into the discrete-time domain. For example, if a lowpass filter $H(s)$, with cutoff frequency Ω_c, is transformed using $H_H(s) = H(1/s)$, then the resulting filter $H_H(s)$ is of the highpass type, with the cutoff frequency $1/\Omega_c$.

In the discrete-time domain, a highpass filter frequency response, $H_H(e^{j\omega})$, is obtained by shifting the corresponding lowpass filter response, $H(e^{j\omega})$, for π in frequency, Fig. 5.13, that is

$$H_H(e^{j\omega}) = H(e^{j(\omega-\pi)}).$$

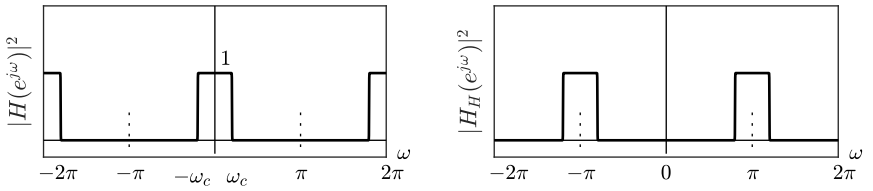


Figure 5.13 Ideal highpass filter, $H_H(e^{j\omega})$, as a shifted version of the ideal lowpass filter, $H(e^{j\omega})$.

The shift in frequency corresponds to the modulation of the impulse response,

$$h_H(n) = e^{j\pi n}h(n) = (-1)^n h(n).$$

Thus, if we have an impulse response $h(n)$ of a lowpass filter, the corresponding highpass filter impulse response, $h_H(n)$, is obtained by multiplying the impulse response values $h(n)$ by $(-1)^n$. The output of the highpass filter to any input signal $x(n)$ is given by

$$y(n) = x(n) *_n h_H(n) = \sum_{m=-\infty}^{\infty} x(m)(-1)^{n-m}h(n-m)$$

$$= (-1)^n \sum_{m=-\infty}^{\infty} (-1)^m x(m)h(n-m) = (-1)^n \Big(\Big[(-1)^n x(n) \Big] *_n h(n) \Big) \tag{5.19}$$

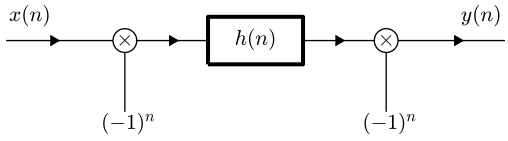


Figure 5.14 Highpass filter realization using the corresponding lowpass filter.

This relation means that the lowpass filter can be implemented using the scheme shown in Fig. 5.14.

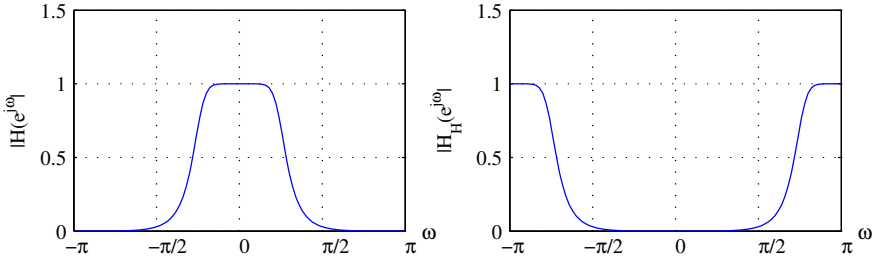


Figure 5.15 Amplitude of the frequency response of a lowpass Butterworth filter (left) and the filter obtained from the lowpass Butterworth filter when z is replaced by $-z$ (right).

Example 5.9. For the lowpass Butterworth discrete-time filter

$$H(z) = \frac{0.1236\left(1+z^{-1}\right)^4}{\left(z^{-2}-1.9389z^{-1}+1.7420\right)\left(z^{-2}-4.0790z^{-1}+4.7686\right)}$$

from Fig. 5.15 plot the frequency response if z is replaced by $-z$.

★The impulse response is obtained by changing the sign for every other sample in $h(n)$. In the z-transform definition that means using $(-z)^{-n}$ instead of z^{-n}. The frequency response of

$$H_H(z) = \frac{0.1236\left(1-z^{-1}\right)^4}{\left(z^{-2}+1.9389z^{-1}+1.7420\right)\left(z^{-2}+4.0790z^{-1}+4.7686\right)}$$

is shown in Fig. 5.15.

5.5.3 Bandpass Filters

A bandpass filter is obtained from the corresponding lowpass filter by shifting its frequency response for ω_0 and $-\omega_0$, as shown in Fig. 5.16. The frequency response of the bandpass filter is

$$H_B(e^{j\omega}) = H(e^{j(\omega-\omega_0)}) + H(e^{j(\omega+\omega_0)}).$$

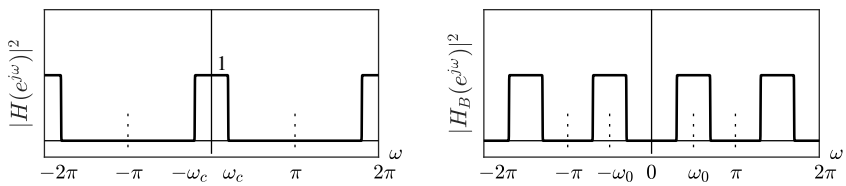


Figure 5.16 Bandpass filter as a shifted version of the lowpass filter.

In the discrete-time domain these frequency shifts correspond to

$$h_B(n) = e^{j\omega_0 n}h(n) + e^{-j\omega_0 n}h(n) = 2\cos(\omega_0 n)h(n).$$

In general, for an input signal $x(n)$, the output of a bandpass filter is

$$\begin{aligned}
y(n) = h_B(n) *_n x(n) &= \sum_{m=-\infty}^{\infty} h_B(m)x(n-m) = 2\sum_{m=-\infty}^{\infty} \cos(\omega_0 m)h(m)x(n-m) \\
&= 2\sum_{m=-\infty}^{\infty} \cos(\omega_0 n + \omega_0 m - \omega_0 n)h(m)x(n-m) \\
&= 2\sum_{m=-\infty}^{\infty} [\cos(\omega_0 n)\cos(\omega_0 m - \omega_0 n) - \sin(\omega_0 n)\sin(\omega_0 m - \omega_0 n)]h(m)x(n-m) \\
&= 2\cos(\omega_0 n)\sum_{m=-\infty}^{\infty} \cos(\omega_0(n-m))x(n-m)h(m) \\
&\quad + 2\sin(\omega_0 n)\sum_{m=-\infty}^{\infty} \sin(\omega_0(n-m))x(n-m)h(m).
\end{aligned}$$

The last relation indicates that we may write the output of a bandpass filter as a function of the lowpass impulse response in the form

$$y(n) = 2\cos(\omega_0 n)\{[\cos(\omega_0 n)x(n)] * h(n)\} + 2\sin(\omega_0 n)\{[\sin(\omega_0 n)x(n)] * h(n)\}.$$

This relation leads to the realization of the bandpass filter using the corresponding lowpass filter, as shown in Fig. 5.17.

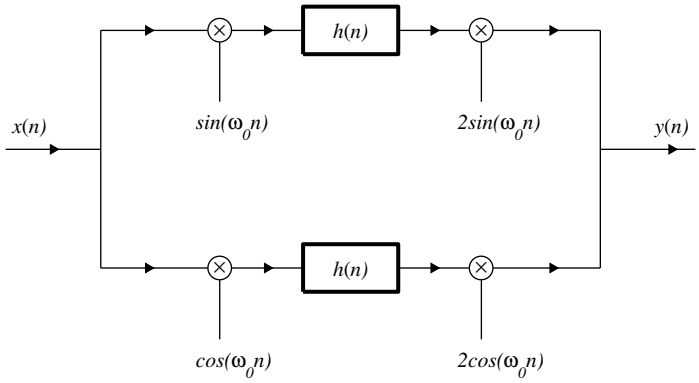


Figure 5.17 Bandpass system realization using the corresponding lowpass filter and signal modulation.

5.5.4 Allpass Systems - System Stabilization

A system (filter) with unit (constant) amplitude of the frequency response is defined by

$$H_A(z) = \frac{z^{-1} - ae^{-j\theta}}{1 - ae^{j\theta}z^{-1}} = \frac{1 - zae^{-j\theta}}{z - ae^{j\theta}} = \frac{z - \frac{1}{a}e^{j\theta}}{1 - \frac{1}{a}e^{-j\theta}z}e^{-j2\theta},$$

where $0 < a < 1$ is real-valued and θ is an arbitrary phase. For this system

$$\left|H_A(e^{j\omega})\right| = 1.$$

To prove this statement consider

$$\left|H_A(e^{j\omega})\right| = \left|\frac{e^{-j\omega} - ae^{-j\theta}}{1 - ae^{j\theta}e^{-j\omega}}\right| = \left|\frac{e^{j(\theta-\omega)} - a}{1 - ae^{j\theta}e^{-j\omega}}\right|$$
$$= \sqrt{\frac{(\cos(\theta-\omega) - a)^2 + \sin^2(\theta-\omega)}{(1 - a\cos(\theta-\omega))^2 + a^2\sin^2(\theta-\omega)}} = \sqrt{\frac{a^2 - 2a\cos(\theta-\omega) + 1}{1 - 2a\cos(\theta-\omega) + a^2}} = 1.$$

Example 5.10. Given the system

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

This system cannot be causal and stable since there is a pole at $z = 2$. Define an allpass system to be connected to $H(z)$ in cascade such that the resulting system is causal and stable, with the same amplitude of the frequency response as $H(z)$.

★When an allpass system, $H_A(z)$, is added in cascade with the given system, $H(z)$, the overall system transfer function, $H_s(z)$ is of the form

$$H_s(z) = H(z)H_A(z) = \frac{z+2}{(z-\frac{1}{2})(z-\frac{1}{3})(z-2)}\frac{z-\frac{1}{a}e^{j\theta}}{1-\frac{1}{a}e^{-j\theta}z}e^{-j2\theta}.$$

The values of a and θ are chosen is such a way that the undesirable pole at $z = 2$ is canceled out, that is, $a = 1/2$ and $\theta = 0$. With these values of a and θ we get

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

This system has the same amplitude of the frequency response as the initial system $H(z)$, since

$$\left|H_s(e^{j\omega})\right| = \left|H(e^{j\omega})H_A(e^{j\omega})\right| = \left|H(e^{j\omega})\right|\left|H_A(e^{j\omega})\right| = \left|H(e^{j\omega})\right|.$$

The allpass system can be generalized to the form

$$H_A(z) = \frac{z^{-1} - a_1e^{-j\theta_1}z^{-1}}{1 - a_1e^{j\theta_1}z^{-1}}\frac{z^{-1} - a_2e^{-j\theta_2}z^{-1}}{1 - a_2e^{j\theta_2}z^{-1}}\cdots\frac{z^{-1} - a_Ne^{-j\theta_N}z^{-1}}{1 - a_Ne^{j\theta_N}z^{-1}},$$

where $0 < a_i < 1$ and θ_i, $i = 1,2,\ldots,N$ are arbitrary real-valued constants and phases, respectively. The resulting frequency response amplitude is

$$\left|H_A(e^{j\omega})\right| = 1.$$

This system can be used for multiple poles cancellation and phase correction.

5.5.5 Inverse and Minimum Phase Systems

An inverse system to the system $H(z)$ is defined by

$$H_i(z) = \frac{1}{H(z)}.$$

It is obvious that

$$H(z)H_i(z) = 1$$
$$h(n) * h_i(n) = \delta(n).$$

The inverse system can be used to reverse the signal distortion. For example, assume that the Fourier transform of a signal $x(n)$ is distorted during the transmission by a transfer function $H(z)$, that is, the received signal z-transform is $R(z) = H(z)X(z)$. In this case, the distortion can be compensated by processing the received signal using the inverse system. The output signal is obtained as

$$Y(z) = \frac{1}{H(z)}R(z) = X(z).$$

The system $H_i(z) = 1/H(z)$ should be stable as well. It means that the poles of the inverse system should be within the unit circle. The poles of the inverse system are equal to the zeros of $H(z)$.

The system $H(z)$ whose both poles and zeros are within the unit circle is called a minimum phase system.

Example 5.11. (a) Which of these two systems

$$H_1(z) = \frac{z^2 + z - \frac{5}{16}}{z^2 + z + \frac{3}{16}}$$
$$H_2(z) = \frac{z^2 - z + \frac{3}{16}}{z^2 + z + \frac{3}{16}}$$

is a minimum phase system?

(b) If the amplitude of the Fourier transform of the discrete-time received signal is distorted as $R(z) = H_1(z)X(z)$, where $H_1(z)$ is defined in (a), what is the stable and causal system $H_D(z)$ that will produce $\left|Y(e^{j\omega})\right| = \left|X(e^{j\omega})\right|$ at its output if the input is the received signal $r(n)$?

★ a) The systems can be written as

$$H_1(z) = \frac{(z - \frac{1}{4})(z + \frac{5}{4})}{(z + \frac{1}{4})(z + \frac{3}{4})}$$
$$H_2(z) = \frac{(z - \frac{1}{4})(z - \frac{3}{4})}{(z + \frac{1}{4})(z + \frac{3}{4})}$$

The first system is causal and stable for the region of convergence $|z| > 3/4$. However one of its zeros is at $|z| = 5/4 > 1$ and the system is not a minimum phase system, since its causal inverse form is not stable. The second system is causal and stable. The same holds for its inverse, since

all poles of the inverse system are within $|z| < 1$. Thus, the system $H_2(z)$ is a minimum phase system.

(b) In this case

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

An inverse system to $H_1(z)$ cannot be used since it will not be stable. However, the inverse system can be stabilized with an allpass system $H_A(z)$ so that the amplitude is not changed

$$Y(z) = R(z)\frac{1}{H_1(z)}H_A(z) = H_1(z)X(z)\frac{1}{H_1(z)}H_A(z)$$

where

$$H_A(z) = \frac{z + \frac{5}{4}}{1 + \frac{5}{4}z}$$

and

$$H_D(z) = \frac{1}{H_1(z)}H_A(z) = \frac{(z + \frac{1}{4})(z + \frac{3}{4})}{(z - \frac{1}{4})(z + \frac{5}{4})}\frac{(z + \frac{5}{4})}{(1 + \frac{5}{4}z)} = \frac{(z + \frac{1}{4})(z + \frac{3}{4})}{(z - \frac{1}{4})(1 + \frac{5}{4}z)}$$

This system is stable and causal and will produce $\left|Y(e^{j\omega})\right| = \left|X(e^{j\omega})\right|$.

If a system is the minimum phase system (with all poles and zeros within $|z| < 1$), then this system has a minimum group delay out of all systems with the same amplitude of the frequency response. Thus, any nonminimum phase system will have a more negative phase compared to the minimum phase system. The negative part of the phase is called the phase-lag function. The name minimum phase system comes from the minimum phase-lag function.

In order to prove this statement consider a system $H(z)$ with the same amplitude of the frequency response as a nonminimum phase system $H_{\min}(z)$. Its frequency response can, therefore, be written as

$$H(z) = H_{\min}(z)H_A(z) = H_{\min}(z)\frac{z^{-1} - ae^{-j\theta}}{1 - ae^{j\theta}z^{-1}}$$

Here, we assumed the first-order allpass system, without any loss of generality, since the same proof can be used for any number of allpass systems that multiply $H_{\min}(z)$. Since $0 < a < 1$ and the system $H_{\min}(z)$ is stable the system $H(z)$ has a zero at $|z| = 1/a > 1$.

The phases of the systems in the previous equation are related as

$$\arg\{H(e^{j\omega})\} = \arg\{H_{\min}(e^{j\omega})\} + \arg\{H_A(e^{j\omega})\}.$$

The phase of the allpass system is

$$\arg\{H_A(e^{j\omega})\} = \arg\{\frac{e^{-j\omega} - ae^{-j\theta}}{1 - ae^{j\theta}e^{-j\omega}}\} = \arg\{e^{-j\omega}\frac{1 - ae^{-j\theta}e^{j\omega}}{1 - ae^{j\theta}e^{-j\omega}}\}$$

$$= -\omega + \arg\{1 - ae^{-j\theta}e^{j\omega}\} - \arg\{1 - ae^{j\theta}e^{-j\omega}\} = -\omega - 2\arctan\frac{a\sin(\omega - \theta)}{1 - a\cos(\omega - \theta)}.$$

Its derivative (group delay) is

$$\tau_{gA}(\omega) = -\frac{d(\arg\{H_A(e^{j\omega})\})}{d\omega} = 1 + 2\frac{a\cos(\omega-\theta) - a^2}{1 - 2a\cos(\omega-\theta) + a^2}$$
$$= \frac{1-a^2}{1-2a\cos(\omega-\theta)+a^2} = \frac{1-a^2}{\left|1-ae^{j(\omega-\theta)}\right|^2}.$$

Since $a < 1$, the group delay $\tau_{gA}(\omega)$ is always positive and

$$\tau_g(\omega) = \tau_{g\min}(\omega) + \tau_{gA}(\omega)$$
$$\tau_g(\omega) \geq \tau_{g\min}(\omega),$$

with $\tau_g(\omega)$ and $\tau_{g\min}(\omega)$ being the phase derivatives (group delays) of systems $H(z)$ and $H_{\min}(z)$, respectively.

The phase behavior of all pass system is

$$\arg\{H_A(e^{j0})\} = \arg\{\frac{1-ae^{-j\theta}}{1-ae^{j\theta}}\} = 0 \tag{5.20}$$

$$\arg\{H_A(e^{j\omega})\} = -\int_0^{\omega} \tau_g(\omega)d\omega \leq 0 \tag{5.21}$$

$$\text{since } \tau_g(\omega) > 0 \text{ for } 0 \leq \omega < \pi.$$

We can conclude that the minimum phase systems satisfy the following conditions.

1. A minimum phase system is system of minimum group delay out of the systems with the same amplitude of frequency response. A system containing one or more allpass parts with uncompensated zeros outside of the unit circle will have larger delay than the system which does not contain zeros outside the unit circle.

2. The phase of a minimum phase system will be lower than the phase of any other system with the same amplitude of frequency response since, according to (5.21),

$$\arg\{H(e^{j\omega}) = \arg\{H_{\min}(e^{j\omega})\} + \arg\{H_A(e^{j\omega})\} \leq \arg\{H_{\min}(e^{j\omega})\}.$$

This proves the fact that the phase of any system $\arg\{H(e^{j\omega})$ is always lower than the phase of minimum phase system $\arg\{H_{\min}(e^{j\omega})\}$, having the same amplitude of the frequency response.

3. Since the group delay is minimum we can conclude that

$$\sum_{m=0}^{n} |h_{\min}(m)|^2 \geq \sum_{m=0}^{n} |h(m)|^2.$$

This relation may be proven in a similar way like the minimum phase property, by considering the outputs of a minimum phase system and a system $H(z) = H_{\min}(z)H_A(z)$.

Example 5.12. A system has absolute squared amplitude of the frequency response equal to

$$\left|H(e^{j\omega})\right|^2 = \frac{\left(2\cos(\omega) + \frac{5}{2}\right)^2}{(12\cos(\omega)+13)(24\cos(\omega)+25)}$$

Find the corresponding minimum phase system.

★ For the system we can write

$$\left|H(e^{j\omega})\right|^2 = H(e^{j\omega})H^*(e^{j\omega}) = H(e^{j\omega})H(e^{-j\omega})$$

In the $z-$domain the system with this amplitude of the frequency response (with real-valued coefficients) satisfies

$$H(z)H^*(\frac{1}{z^*})\Big|_{z=e^{j\omega}} = H(z)H(\frac{1}{z})\Big|_{z=e^{j\omega}} = \left|H(e^{j\omega})\right|^2 = H(e^{j\omega})H(e^{-j\omega}).$$

In this sense

$$\left|H(e^{j\omega})\right|^2 = \frac{\left(e^{j\omega}+e^{-j\omega}+\frac{5}{2}\right)^2}{(6e^{j\omega}+6e^{-j\omega}+13)(12e^{j\omega}+12e^{-j\omega}+25)}$$

and

$$H(z)H(\frac{1}{z}) = \frac{\left(z+\frac{5}{2}+z^{-1}\right)^2}{(6z+13+6z^{-1})(12z+25+12z^{-1})}$$
$$= \frac{\left(z^2+\frac{5}{2}z+1\right)^2}{(6z^2+13z+6)(12z^2+25z+12)}$$
$$= \frac{1}{72}\frac{(z+2)^2(z+\frac{1}{2})^2}{(z+\frac{2}{3})(z+\frac{3}{2})(z+\frac{3}{4})(z+\frac{4}{3})} = \frac{1}{72}\frac{(\frac{1}{z}+\frac{1}{2})^2(z+\frac{1}{2})^2}{(z+\frac{2}{3})(\frac{1}{z}+\frac{2}{3})(z+\frac{3}{4})(\frac{1}{z}+\frac{3}{4})}.$$

The minimum phase system, with the desired amplitude of the frequency response, is a part of $H(z)H^*(\frac{1}{z^*}) = H(z)H(\frac{1}{z})$ with the zeros and poles inside the unit circle

$$H(z) = \frac{\sqrt{2}}{12}\frac{(z+\frac{1}{2})^2}{(z+\frac{2}{3})(z+\frac{3}{4})}.$$

The other poles and zeros then belong to $H^*(1/z^*)$.

5.6 PROBLEMS

Problem 5.1. The transfer function of an RLC circuit is given by

$$H(s) = \frac{\frac{1}{LC}}{s^2 + s\frac{R}{L} + \frac{1}{LC}}$$

with $R/L = 8$ and $1/(LC) = 25$. Find the difference equation describing the corresponding discrete-time system obtained by the impulse invariance method. What is the impulse response of the discrete-time system. Use the sampling interval $\Delta t = 1$.

Problem 5.2. Could the method of impulse invariance be used to map the system

$$H(s) = \frac{s^2 - 3s + 3}{s^2 + 3s + 3}$$

to the discrete-time domain. What is the corresponding discrete-time system obtained by the bilinear transform with $\Delta t = 1$?

Problem 5.3. A continuous-time system is described by the differential equation

$$y''(t) + \frac{3}{2}y'(t) + \frac{1}{2}y(t) = x(t)$$

with the zero initial conditions. What is the corresponding transfer function of the discrete-time system using the first-order backward difference approximation with $\Delta t = 1/10$? Write the difference equation of the system whose output approximates the output of the continuous-time system.

Problem 5.4. Transfer function of a continuous-time system is

$$H(s) = -\frac{2s}{s^2 + 2s + 2}.$$

What is the corresponding discrete-time system using the invariance impulse method and the bilinear transform for system mapping with the sampling interval $\Delta t = 1$?

Problem 5.5. A continuous-time system is described by the transfer function of the form

$$H(s) = \frac{(1 + 4s)}{(s + 1/2)(s + 1)^3}.$$

What is the corresponding discrete-time system according to:

(a) the impulse invariance method,
(b) the bilinear transform,
(c) the matched z-transform?
Use $\Delta t = 1$.

Problem 5.6. The continuous-time system

$$H(s) = \frac{2Q\Omega_1}{s^2 + 2\Omega_1 Qs + \Omega_1^2 + Q^2}$$

is designed to pass the signal

$$x(t) = A_1 \cos(\Omega_1 t + \varphi_1)$$

and to stop all other possible signal components. The parameters are $Q = 0.01$, $\Omega_1 = \pi/2$. The signal is sampled with $\Delta t = 1$ and a discrete-time signal $x(n)$ is formed. Using the bilinear transform design the discrete system that corresponds to the continuous-time system with the transfer function $H(s)$.

Problem 5.7. (a) By using the bilinear transform find the transfer function of the second-order Butterworth filter with $f_{ac} = 4kHz$. The sampling interval is $\Delta t = 50\mu$sec.
(b) Translate the discrete-time transfer function to obtain a highpass filter. Find its corresponding critical frequency in the continuous-time domain.

Problem 5.8. Design a discrete-time lowpass Butterworth filter for the sampling frequency $1/\Delta t = 10$ kHz. The passband should be from 0 to 1 kHz, the maximum attenuation in the passband should be 3 dB ($a_p \geq -3$dB) and the attenuation should be more than 10 dB ($a_s < -10$dB) for frequencies above 2 kHz.

Problem 5.9. Using the impulse invariance method design a Butterworth filter with the passband frequency $\omega_p = 0.1\pi$ and the stopband frequency $\omega_s = 0.3\pi$ in the discrete-time domain. The maximum attenuation in the passband region should be less than 2dB, and the minimum attenuation in the stopband should be 20dB.

Problem 5.10. A highpass filter can be obtained from a lowpass using $H_H(s) = H(1/s)$. With the bilinear transform with $\Delta t = 2$ we can transform the continuous-time domain function into discrete domain using the relation $s = (z-1)/(z+1)$. If we have a design of a lowpass filter how to change its coefficients in order to get a highpass filter.

Problem 5.11. For filtering of a continuous-time signal, a discrete-time filter is used. Find the corresponding continuous-time filter frequencies if the discrete-time filter is: a) lowpass with $\omega_p = 0.15\pi$, b) bandpass within $0.2\pi \leq \omega \leq 0.25\pi$, c) highpass with $\omega_p = 0.35$. Consider the cases when $\Delta t = 0.001s$ and $\Delta t = 0.1s$.

What should be the frequencies to design these systems in the continuous-time domain if the impulse invariance method is used and what are the design frequencies if the bilinear transform is used?

Problem 5.12. A transfer function of the first-order lowpass system is

$$H(z) = \frac{1-\alpha}{1-\alpha z^{-1}}.$$

Find the corresponding bandpass system transfer function with frequency shifts for $\pm\omega_c$.

Problem 5.13. Using an appropriate allpass system find the stable systems with the same amplitude of the frequency response as the systems:

(a)
$$H_1(z) = \frac{2-3z^{-1}+2z^{-2}}{1-4z^{-1}+4z^{-2}}$$

(b)
$$H_2(z) = \frac{z}{(4-z)(1/3-z)}.$$

Problem 5.14. The z-transform

$$R(z) = \frac{(z-\frac{1}{4})(z^{-1}-\frac{1}{4})(z+\frac{1}{2})(z^{-1}+\frac{1}{2})}{(z+\frac{4}{5})(z^{-1}+\frac{4}{5})(z-\frac{3}{7})(z^{-1}-\frac{3}{7})}$$

can can be written as

$$R(z) = H(z)H^*(\frac{1}{z^*}).$$

Find $H(z)$ for the minimum phase system.

Problem 5.15. A signal $x(n)$ has passed trough a media whose influence can be described by the transfer function

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

Signal $r(n)$ is obtained. Find a causal and stable system to process $r(n)$ in order to obtain the output signal $y(n)$ such that $\left|Y(e^{j\omega})\right| = \left|X(e^{j\omega})\right|$.

5.7 EXERCISE

Exercise 5.1. The transfer function of the continuous-time system is

$$H(s) = \frac{(s+2)}{4s^2+s+1}.$$

What is the corresponding discrete-time system obtained with $\Delta t = 1$ by using the impulse invariance method and the bilinear transform.

Exercise 5.2. A continuous system is described by the differential equation

$$y''(t) + 6y'(t) - y(t) = x(t) + \frac{1}{2}x'(t)$$

with the zero initial conditions. What is the corresponding transfer function of the discrete system obtained using the first-order backward difference approximation with $\Delta t = 1$?

Exercise 5.3. (a) The continuous-time system

$$H(s) = \frac{2Q\Omega_0}{s^2 + 2\Omega_0 Qs + \Omega_0^2 + Q^2}$$

with $Q = 0.01$ is designed to pass the signal

$$x(t) = A\cos(\Omega_0 t + \varphi)$$

for $\Omega_0 = 3\pi/4$ and to stop all other possible signal components.

The signal is sampled with $\Delta t = 1$ and a discrete-time signal $x(n)$ is formed. Using the bilinear transform, design a discrete-time system that corresponds to the continuous-time system with transfer function $H(s)$.

(b) What is the output $r(n)$ of the obtained discrete-time system to the samples $y(n)$ of the analog signal

$$y(t) = 1 + 2\sin(250\pi t) - \cos(2750\pi t) + 2\sin(750\pi t)$$

sampled with the sampling interval $\Delta t = 10^{-3}$ s. What would be the corresponding continuous-time output signal after an ideal D/A converter.

Exercise 5.4. (a) By using the bilinear transform find the transfer function of the third-order Butterworth filter with the cutoff frequency $f_c = 3.4$ kHz. The sampling step is $\Delta t = 40$ μsec.

(b) Translate the discrete transfer function to obtain a bandpass system with the corresponding central frequency $f_0 = 12.5$ kHz in the continuous-time domain.

Exercise 5.5. Design a continuous0time lowpass filter whose parameters are:
- passband frequency $\Omega_p = 2\pi f_p$, $f_p = 3.5$ kHz,
- stopband frequency $\Omega_s = 2\pi f_s$, $f_s = 6$ kHz,
- maximum attenuation in passband $a_p = 2$ dB, and
- minimum attenuation in the stopband $a_s = 16$ dB.

Find the corresponding discrete-time filter using:
(a) the impulse invariance method and
(b) the bilinear transform,

with $\Delta t = 0.05 \times 10^{-3}$ sec.

(c) Write the corresponding highpass filter transfer functions, obtained by a frequency shift in the discrete domain for π, for both cases.

Exercise 5.6. Using an allpass system find a stable and causal system with the same amplitude of the frequency response as the systems:

$$H_1(z) = \frac{2 - 5z^{-1} + 2z^{-2}}{1 - 4z^{-1} + z^{-2}},$$

$$H_2(z) = \frac{z - 1}{(2 - z)(1/4 - z)}.$$

Exercise 5.7. The z-transform

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

can can be written as

$$R(z) = H(z)H^*(\frac{1}{z^*}).$$

Find $H(z)$ for the minimum phase system. If $h(n)$ is the impulse response of $H(z)$ and $h_1(n)$ is the impulse response of

$$H_1(z) = H(z)\frac{z^{-1} - a_1 e^{-j\theta_1}}{1 - a_1 e^{j\theta_1} z^{-1}}$$

show that $|h(0)| \leq |h_1(0)|$ for any θ_1 and $|a_1| < 1$. All systems are causal.

Exercise 5.8. A signal $x(n)$ has passed trough a media whose influence can be described by the transfer function

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

and the signal $r(n) = x(n) * h(n)$ is obtained. Find a causal and stable system to process $r(n)$ in order to obtain $\left|Y(e^{j\omega})\right| = \left|X(e^{j\omega})\right|$.

5.8 SOLUTIONS

Solution 5.1. For this system we can write

$$H(s) = \frac{\frac{1}{LC}}{s^2 + s\frac{R}{L} + \frac{1}{LC}} = \frac{25}{s^2 + 8s + 25} = \frac{25}{(s+4+3j)(s+4-j3)} = \frac{j\frac{25}{6}}{s+4+j3} + \frac{-j\frac{25}{6}}{s+4-j3}.$$

The poles are mapped using

$$s_i \to z_i = e^{s_i}.$$

The discrete-time system is

$$\begin{aligned} H(z) &= \frac{j\frac{25}{6}}{1 - e^{-(4+j3)}z^{-1}} + \frac{-j\frac{25}{6}}{1 - e^{-(4-j3)}z^{-1}} \\ &= \frac{\frac{25}{3}e^{-4}z^{-1}\sin(3)}{1 - 2e^{-4}\cos 3z^{-1} + e^{-8}z^{-2}}, \end{aligned}$$

with the corresponding difference equation

$$y(n) = \frac{25}{3}e^{-4}\sin(3)x(n-1) + 2e^{-4}\cos(3)y(n-1) - e^{-8}y(n-2).$$

The output signal values can be calculated for any input signal using this difference equation. For $x(n) = \delta(n)$ the impulse response follows. The impulse response can be obtained in a closed form from

$$H(z) = j\frac{25}{6}\sum_{n=0}^{\infty} e^{-(4+j3)n}z^{-n} - j\frac{25}{6}\sum_{n=0}^{\infty} e^{-(4-j3)n}z^{-n}$$

as

$$h(n) = \frac{25}{6}e^{-4n}(je^{-j3n} - je^{j3n})u(n) = \frac{25}{3}e^{-4n}\sin(3n)u(n).$$

There is no correction term since $\lim_{z\to\infty} H(z) = 0$.

Solution 5.2. The system is not of lowpass type. For $s \to \infty$ we get $H(s) \to 1$. Thus, the impulse invariance method cannot be used. The bilinear transform can be used. It produces

$$H(z) = \frac{4\frac{(1-z^{-1})^2}{(1+z^{-1})^2} - 6\frac{1-z^{-1}}{1+z^{-1}} + 3}{4\frac{(1-z^{-1})^2}{(1+z^{-1})^2} + 6\frac{1-z^{-1}}{1+z^{-1}} + 3} = \frac{13z^{-2} - 2z^{-1} + 1}{z^{-2} - 2z^{-1} + 13}.$$

Solution 5.3. For the system

$$y''(t) + \frac{3}{2}y'(t) + \frac{1}{2}y(t) = x(t)$$

the transfer function is

$$H(s) = \frac{1}{s^2 + \frac{3}{2}s + \frac{1}{2}}.$$

The corresponding discrete-time system is obtained using

$$s \to \frac{1-z^{-1}}{\Delta t} = 10(1-z^{-1})$$

as

$$H(z) = \frac{1}{100(1-z^{-1})^2 + \frac{3}{2}10(1-z^{-1}) + \frac{1}{2}} = \frac{1}{100z^{-2} - 215z^{-1} + \frac{231}{2}}.$$

The difference equation for this system is given by

$$y(n) = \frac{2}{231}x(n) + \frac{430}{231}y(n-1) - \frac{200}{231}y(n-2).$$

Solution 5.4. The transfer function can be written in the form

$$H(s) = -\frac{1+j}{s+1-j} - \frac{1-j}{s+1+j}.$$

Using the invariance impulse method, the transfer function of the discrete-time system follows

$$H(z) = -\frac{2 - 2(\cos(1)+\sin(1))e^{-1}z^{-1}}{1 - 2\cos(1)e^{-1}z^{-1} + e^{-2}z^{-2}}.$$

The bilinear transform produces

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

Solution 5.5. (a) The transfer function

$$H(s) = \frac{(1+4s)}{(s+1/2)(s+1)^3}$$

is expanded into partial fractions, appropriate for the impulse invariance method, as

$$H(s) = \frac{k_1}{s+1/2} + \frac{k_2}{(s+1)} + \frac{k_3}{(s+1)^2} + \frac{k_4}{(s+1)^3}$$

with $k_1 = H(s)(s+1/2)|_{s=-1/2} = -8$ and $k_4 = H(s)(s+1)^3\big|_{s=-1} = 6$. By equating the coefficients with s^3 to 0, we get the relation $k_1 + k_2 = 0$. A similar relation follows for the coefficients with s^2 in the form $3k_1 + 5k_2/2 + k_3 = 0$ or $k_1/2 + k_3 = 0$. Then, $k_2 = 8$ and $k_3 = 4$. With

$$\frac{k_i}{s-s_i} \to \frac{k_i}{1-e^{s_i}z^{-1}}$$

and

$$\frac{1}{m!}\frac{d^m}{ds_i^m}\frac{k_i}{s-s_i} \to \frac{1}{m!}\frac{d^m}{ds_i^m}\{\frac{k_i}{1-e^{s_i}z^{-1}}\}$$

we get the discrete-time system transfer function

$$H(z) = \frac{-8}{1-e^{-1/2}z^{-1}} + \frac{8}{1-e^{-1}z^{-1}} + \frac{d}{ds_1}\left(\frac{4}{1-e^{s_1}z^{-1}}\right)\Big|_{s_1=-1} + \frac{d^2}{2!ds_1^2}\left(\frac{6}{1-e^{s_1}z^{-1}}\right)\Big|_{s_1=-1}$$
$$= \frac{-8}{1-e^{-1/2}z^{-1}} + \frac{8}{1-e^{-1}z^{-1}} + \frac{4e^{-1}z^{-1}}{(1-e^{-1}z^{-1})^2} + \frac{3e^{-2}z^{-2}+3e^{-1}z^{-1}}{(1-e^{-1}z^{-1})^3}$$
$$= \frac{-5.83819z^{-3}-9.68722z^{-2}+22.0531z^{-1}}{(z^{-1}-e)^3(z^{-1}-e^{1/2})}$$

(b) The discrete-time system, obtained using the bilinear transform, is given by

$$H(z) = \frac{(1+8\frac{1-z^{-1}}{1+z^{-1}})}{(2\frac{1-z^{-1}}{1+z^{-1}}+1/2)(2\frac{1-z^{-1}}{1+z^{-1}}+1)^3} = \frac{-14z^{-4}-24z^{-3}+12z^{-2}+40z^{-1}+18}{3z^{-4}-32z^{-3}+126z^{-2}-216z^{-1}+135}.$$

(c) The matched z-transform produces

$$H(z) = \frac{4(1-e^{-1/4}z^{-1})}{\left(1-e^{-1/2}z^{-1}\right)\left(1-e^{-1}z^{-1}\right)^3}.$$

Solution 5.6. Since the bilinear transform is used, we have to pre-modify the system according to

$$\Omega_d = \frac{2}{\Delta t}\tan(\frac{\Omega_1\Delta t}{2}) = 2.0 = 0.6366\pi.$$

The frequency value is shifted from $\Omega_1 = 0.5\pi$ to $\Omega_d = 0.6366\pi$. The modified system is

$$H_d(s) = \frac{2Q\Omega_d}{s^2+2\Omega_d Qs+\Omega_d^2+Q^2}.$$

Now, using $s = 2\frac{1-z^{-1}}{1+z^{-1}}$, the corresponding discrete-time system is obtained,

$$H(z) = \frac{2Q\Omega_d}{\left(2\frac{1-z^{-1}}{1+z^{-1}}\right)^2+2\Omega_d Q\left(2\frac{1-z^{-1}}{1+z^{-1}}\right)+\Omega_d^2+Q^2}.$$

The bilinear transform returns the pre-modified frequency to the desired one.

Solution 5.7. The poles of $H(s)H(-s)$ for a continuous-time second-order ($N = 2$) Butterworth filter are

$$s_k = \Omega_c e^{j(2\pi k+\pi)/2N+j\pi/2} = 2\pi f_c e^{j(2\pi k+\pi)/4+j\pi/2},$$

where

$$f_c = \frac{2}{\Delta t}\tan(2\pi f_{ac}\Delta t/2)/(2\pi) = 4.6253\text{kHz}.$$

With $k = 0,1,2,3$ follows

$$s_k = 2\pi f_c(\pm\frac{\sqrt{2}}{2} \pm j\frac{\sqrt{2}}{2}).$$

For a stable system, the poles satisfy $\text{Re}\{s_p\} < 0$, thus

$$s_{0,1} = 2\pi f_c(-\frac{\sqrt{2}}{2} \pm j\frac{\sqrt{2}}{2}).$$

The transfer function $H(s)$ is

$$H_a(s) = \frac{s_1 s_2}{(s-s_0)(s-s_1)} = \frac{4\pi^2 f_c^2}{s^2 + 2\pi f_c \sqrt{2} s + 4\pi^2 f_c^2}.$$

Using the bilinear transform with $\Delta t = 50 \cdot 10^{-6}$, we get the corresponding discrete-time system transfer function,

$$H(z) = \frac{1.0548(1+z^{-1})^2}{5.1066 - 1.8874 z^{-1} + z^{-2}}.$$

This filter has -3 dB attenuation at $\omega = 0.4\pi$, corresponding to $\Omega = 0.4\pi/\Delta t = 2\pi \times 4 \times 10^3$.

b) The discrete highpass filter is obtained by the shift corresponding to

$$H_h(e^{j\omega}) = H(e^{j(\omega+\pi)}).$$

This shift corresponds to the impulse response modulation $h_h(n) = (-1)^n h(n)$ or to the substitution of z by $-z$ in the transfer function,

$$H_H(z) = \frac{1.0548(1-z^{-1})^2}{5.1066 + 1.8874 z^{-1} + z^{-2}}.$$

The critical frequency of the highpass filter, $H_H(z)$, is $\omega_c = 0.6\pi$ or $f_{ac} = 6$ kHz.

Solution 5.8. For the continuous-time system the design frequencies are

$$f_p = 1 \text{ kHz}$$
$$f_s = 2 \text{ kHz}.$$

They correspond to

$$\Omega_p = 2\pi 10^3 \text{ rad/s}$$
$$\Omega_s = 4\pi\, 10^3 \text{ rad/s}.$$

The discrete-time frequencies are obtained from $\omega = \Omega\Delta t = \Omega/10^4$ as

$$\omega_p = 0.2\pi$$
$$\omega_s = 0.4\pi.$$

The frequencies for the filter design, that will be mapped to ω_s and ω_p, after the bilinear transform is used, are

$$\Omega_{pd} = \frac{2}{\Delta t}\tan(0.2\pi/2) = \frac{0.6498}{\Delta t}$$
$$\Omega_{sd} = \frac{2}{\Delta t}\tan(0.4\pi/2) = \frac{1.4531}{\Delta t}.$$

The filter order follows from

$$N = \frac{1}{2}\frac{\log \frac{1-10^{0.1a_p}}{1-10^{0.1a_s}}}{\log \frac{\Omega_{pd}}{\Omega_{sd}}} = 1.368.$$

We assume $N = 2$.

Since the frequency for -3 dB attenuation is given, the design cutoff frequency is

$$\Omega_{cd} = \Omega_{pd} = \frac{0.6498}{\Delta t}.$$

The poles of the filter transfer function, for $N = 2$ and Ω_{cd}, are

$$s_{0,1} = \frac{0.6498}{\Delta t}\left(-\frac{\sqrt{2}}{2} \pm j\frac{\sqrt{2}}{2}\right)$$

with the transfer function

$$H(s) = \frac{s_0 s_1}{(s-s_0)(s-s_1)} = \frac{\frac{1}{\Delta t^2}0.4223}{s^2 + 0.919s\frac{1}{\Delta t} + 0.4223\frac{1}{\Delta t^2}}.$$

Mapping of this system into the discrete-time domain using the bilinear transform,

$$s = \frac{2}{\Delta t}\frac{1-z^{-1}}{1+z^{-1}},$$

produces the second-order Butterworth filter

$$H(z) = \frac{0.067569(1+z^{-1})^2}{1 - 1.14216z^{-1} + 0.412441z^{-2}}.$$

Solution 5.9. The Butterworth filter order is

$$N = \frac{1}{2}\frac{\log\frac{1-10^{0.1a_p}}{1-10^{0.1a_s}}}{\log\frac{\Omega_p}{\Omega_s}} = 2.335.$$

with $\Omega_p = \omega_p/\Delta t$, $\Omega_s = \omega_s/\Delta t$, and $\Delta t = 1$. Assume $N = 3$.

The cutoff frequency Ω_c, where the amplitude of the frequency response is attenuated for 3 dB ($a_p = -3$dB), is

$$\Omega_c = \frac{\Omega_p}{\sqrt[2N]{10^{0.1a_p} - 1}} = 0.109345\pi = 0.3435.$$

The transfer function $H(s)$ poles are

$$s_{1,2} = -0.17175 \pm j0.29748$$
$$s_0 = -\Omega_c = -0.3435.$$

The transfer function form is

$$\begin{aligned} H(s) &= \frac{-s_0 s_1 s_2}{(s-s_0)(s-s_1)(s-s_2)} = \frac{0.0405}{(s+0.3435)(s^2+0.3435s+0.1178)} \\ &= \frac{k_1}{s-s_0} + \frac{k_2}{s-s_1} + \frac{k_3}{s-s_2} \\ &= \frac{0.3435}{s+0.3435} - \frac{0.17175 - j0.09916}{s+0.17175+j0.29748} - \frac{0.17175+j0.09916}{s+0.17175-j0.29748}. \end{aligned}$$

The coefficients k_i are calculated from

$$k_i = H(s)(s-s_i)\,|_{s=s_i}\,.$$

Using the impulse invariance method, mapping from the continuous-time domain to the discrete-time domain, is done according to

$$\frac{k_i}{s-s_i} \rightarrow \frac{\Delta t k_i}{1-e^{s_i \Delta t} z^{-1}}.$$

The discrete-time filter transfer function is

$$H(z) = \frac{-0.0253z^{-2} - 0.0318z^{-1}}{-1.98774 + 4.61093z^{-1} - 3.68033z^{-2} + z^{-3}}.$$

Solution 5.10. The transfer function of the highpass filter is

$$H_H(s) = H(\frac{1}{s}),$$

with $s = \frac{2}{\Delta t}\frac{1-z^{-1}}{1+z^{-1}} = \frac{2}{\Delta t}\frac{z-1}{z+1}$ and $\Delta t = 2$. The corresponding lowpass filter would be

$$H_L(z) = H(s)|_{s=\frac{z-1}{z+1}} = H(\frac{z-1}{z+1}).$$

The discrete-time highpass filter is

$$H_H(z) = H_H(s)|_{s=\frac{z-1}{z+1}} = H(\frac{1}{s})\Big|_{s=\frac{z-1}{z+1}}$$

$$H_H(z) = H(\frac{z+1}{z-1}).$$

Obviously $H_H(z) = H_L(-z)$. It means that a discrete highpass system can be realized by replacing z with $-z$ in the transfer function. For $\Delta t \neq 2$ a scaling is present as well.

Solution 5.11. a) The frequency relation with $\Delta t = 0.001$ s produces a lowpass filter with $\Omega_p = \omega_p/\Delta t = 150\,\pi\,\text{rad/s}$. For $\Delta t = 0.1$ *s* the frequency is $\Omega_p = \omega_p/\Delta t = 1.5\,\pi\,\text{rad/s}$.

b) For $\Delta t = 0.001$ s a bandpass filter is obtained for the range 200π rad/s $\leq \Omega \leq 250\pi$ rad/s, while $\Delta t = 0.1$ *s* produces a bandpass filter with 2π rad/s $\leq \Omega \leq 2.5\pi$ rad/s.

c) For $\Delta t = 0.001$ s a highpass filter has the frequency $\Omega_p = 350\,\text{rad/s}$, while for $\Delta t = 0.1$ s the highpass filter has critical frequency $\Omega_p = 3.5\,\text{rad/s}$.

For the impulse invariance method starting design frequencies should be equal to the calculated analog frequencies. If the bilinear transform is used calculated analog frequencies Ω_p should be pre-modified to Ω_m according to $\Omega_m = \frac{2}{\Delta t}\tan\frac{\Omega_p \Delta t}{2}$.

Solution 5.12. The impulse response of the passband filter is $h_B(n) = 2h(n)\cos(\omega_c n)$. The z-transform of the impulse response is

$$H_B(z) = \sum_{n=-\infty}^{\infty} 2h(n)\cos(\omega_c n)z^{-n} = \sum_{n=-\infty}^{\infty} h(n)(e^{-j\omega_c}z)^{-n} + \sum_{n=-\infty}^{\infty} h(n)(e^{j\omega_c}z)^{-n}$$

$$H_B(z) = H(e^{-j\omega_c}z) + H(e^{j\omega_c}z) = \frac{2(1-\alpha)(1-\alpha\cos\omega_c z^{-1})}{1-2\alpha\cos\omega_c z^{-1} + \alpha^2 z^{-2}}.$$

Solution 5.13. The causal system

$$H_1(z) = \frac{2 - 3z^{-1} + 2z^{-2}}{(1 - 2z^{-1})^2}$$

is not stable since it has a second-order pole at $z = 2$. This system may be stabilized, keeping the same amplitude of the frequency response, using a second-order allpass system with zero at $z = 2$

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

The new system has the transfer function

$$H_1(z) = \frac{2 - 3z^{-1} + 2z^{-2}}{(z^{-1} - 2)^2}.$$

The causal system $H_2(z)$ has the pole at $z = 4$. It can be stabilized using the allpass system

$$H_A(z) = \frac{z^{-1} - \frac{1}{4}}{1 - \frac{1}{4}z^{-1}} = \frac{4 - z}{4z - 1}.$$

The transfer function of a stable system is

$$H_2(z) = \frac{z}{(4z - 1)(1/3 - z)}.$$

Solution 5.14. For the z-transform

$$R(z) = \frac{(z - \frac{1}{4})(z^{-1} - \frac{1}{4})(z + \frac{1}{2})(z^{-1} + \frac{1}{2})}{(z + \frac{4}{5})(z^{-1} + \frac{4}{5})(z - \frac{3}{7})(z^{-1} - \frac{3}{7})}$$

and

$$R(z) = H(z)H^*(\frac{1}{z^*}).$$

the minimum phase system is a part of $R(z)$ whose all zeros and poles are located inside the unit circle, meaning that $H(z)$ system and its inverse system $1/H(z)$ can be causal and stable. Therefore,

$$H(z) = \frac{(z - \frac{1}{4})(z + \frac{1}{2})}{(z + \frac{4}{5})(z - \frac{3}{7})}.$$

It is easy to check that $H^*(\frac{1}{z^*})$ is equal to the remaining terms in $R(z)$, since

$$H^*(\frac{1}{z^*}) = \frac{(\frac{1}{z^*} - \frac{1}{4})^*(\frac{1}{z^*} + \frac{1}{2})^*}{(\frac{1}{z^*} + \frac{4}{5})^*(\frac{1}{z^*} - \frac{3}{7})^*} = \frac{(z^{-1} - \frac{1}{4})(z^{-1} + \frac{1}{2})}{(z^{-1} + \frac{4}{5})(z^{-1} - \frac{3}{7})}.$$

Here we used, for example, $(\frac{1}{z^*} - \frac{1}{4})^* = \left(\frac{1}{z^*}\right)^* - \frac{1}{4} = \frac{1}{z} - \frac{1}{4}$.

Solution 5.15. The received signal should be processed by the inverse system

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

However, this system has two poles outside the unit circle since

$$H_i(z) = \frac{z - \frac{1}{2}}{(4 - z)(1/3 - z)(z - 1.2071)(z - 0.2071)}.$$

These poles have to be compensated, keeping the same amplitude, using two first-order allpass systems. The resulting system transfer function is

$$H_i(z)\frac{z - 4}{1 - 4z}\frac{z - 1.2071}{1 - 1.2071z} = \frac{z - \frac{1}{2}}{(1/3 - z)(z - 0.2071)(1 - 4z)(1 - 1.2071z)}.$$

Chapter 6

Realization of Discrete Systems

LINEAR discrete-time systems may, in general, be described by a difference equation relating the output signal with the input signal at the considered instant and the previous values of the output and input signal. The transfer function can be written in various forms producing different system realizations. Some of them will be presented next. Symbols that are used in the realizations are presented in Fig. 6.1.

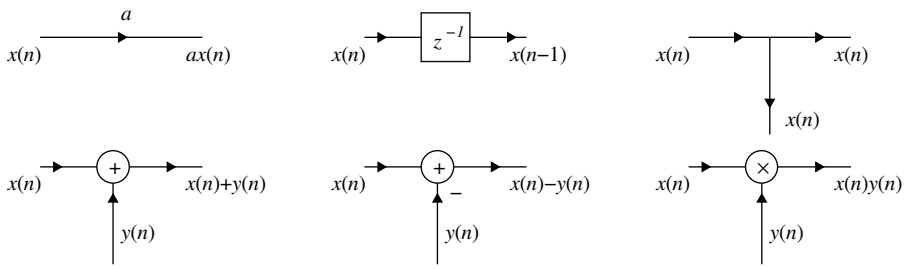


Figure 6.1 Symbols representing particular digital systems and their functions in the realization of discrete-time systems.

6.1 REALIZATION OF IIR SYSTEMS

A system that includes recursions of the output signal values results in an infinite impulse response (IIR). These systems will be presented first.

6.1.1 Direct realization I

Consider a discrete system described by a linear difference equation

$$y(n) = A_1 y(n-1) + \cdots + A_N y(n-N) + B_0 x(n) + B_1 x(n-1) + \cdots + B_M x(n-M). \quad (6.1)$$

The second-order system, as a special case of the system in (6.1), will be presented first. Its implementation is shown in Fig. 6.2. A general system described by (6.1) can be implemented as in Fig. 6.3. This form is a direct realization I of a discrete-time system.

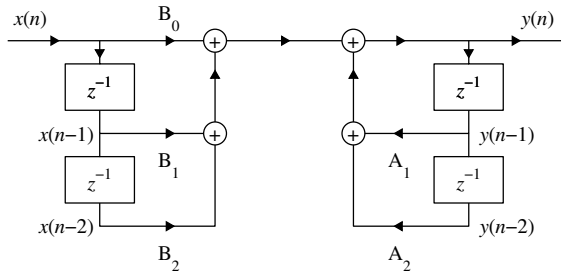


Figure 6.2 Direct form implementation of the second-order system.

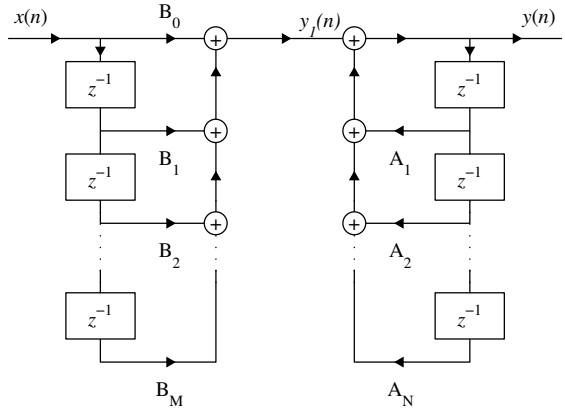


Figure 6.3 Direct form I implementation of a discrete-time system.

6.1.2 Direct realization II

Direct realization I consists of two system blocks connected in cascade. The first block implements the non-recursive part of the difference equation

$$y_1(n) = B_0 x(n) + B_1 x(n-1) + \cdots + B_M x(n-M),$$

while the second block corresponds to the recursive relation

$$y(n) = A_1 y(n-1) + \cdots + A_N y(n-N) + y_1(n),$$

with the output from the first block, $y_1(n)$, being the input signal to the second block. The cascade of these two block is shown in Fig. 6.3. The transfer functions of these blocks are

$$H_1(z) = B_0 + B_1 z^{-1} + \cdots + B_M z^{-M}$$

and

$$H_2(z) = \frac{1}{1 - A_1 z^{-1} - \cdots - A_N z^{-N}}.$$

The overall transfer function is

$$H(z) = H_1(z)H_2(z) = H_2(z)H_1(z) = \frac{B_0 + B_1 z^{-1} + \cdots + B_M z^{-M}}{1 - A_1 z^{-1} - \cdots - A_N z^{-N}}. \tag{6.2}$$

It means that these two blocks can interchange their positions. After the positions are interchanged then, using the same delay systems, we get the resulting system in the direct realization II form, presented in Fig. 6.4. This system uses a reduced number of delay blocks in the realization.

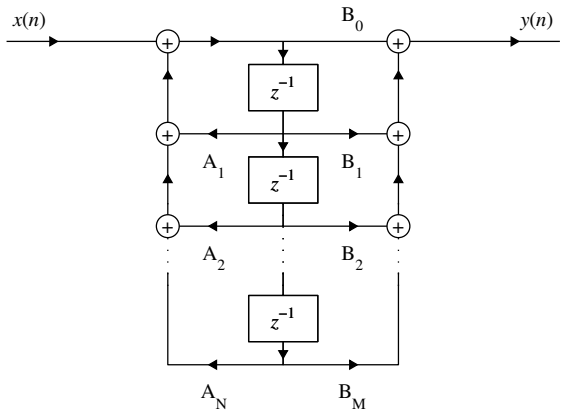


Figure 6.4 Direct realization II of a discrete-time system.

Example 6.1. Find the transfer function of the discrete-time system presented in Fig. 6.5.

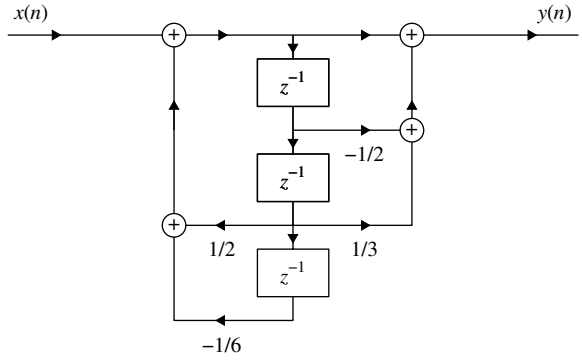


Figure 6.5 A discrete-time system.

★The system can be recognized as a direct realization II form. After its blocks are separated and interchanged, the system in a form shown in Fig. 6.6 is obtained.

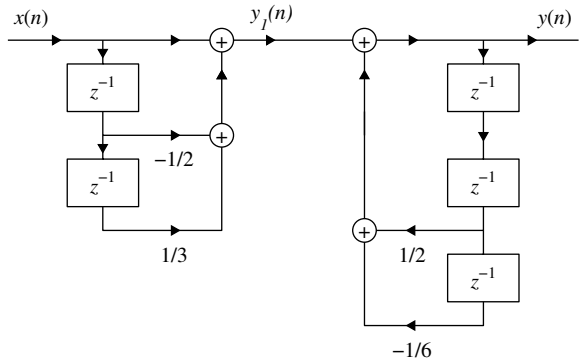


Figure 6.6 The system from Fig. 6.5, with interchanged blocks.

The output of the first block is

$$y_1(n) = x(n) - \frac{1}{2}x(n-1) + \frac{1}{3}x(n-2). \tag{6.3}$$

The transfer function of this block is

$$H_1(z) = 1 - \frac{1}{2}z^{-1} + \frac{1}{3}z^{-2}.$$

The output of the second block is described by the following difference equation

$$y(n) = \frac{1}{2}y(n-2) - \frac{1}{6}y(n-3) + y_1(n). \tag{6.4}$$

The transfer function of then second block is

$$H_2(z) = \frac{1}{1 - \frac{1}{2}z^{-2} + \frac{1}{6}z^{-3}}.$$

The difference equation for the whole system is obtained after $y_1(n)$ from (6.3) is replaced into (6.4)

$$y(n) = \frac{1}{2}y(n-2) - \frac{1}{6}y(n-3) + x(n) - \frac{1}{2}x(n-1) + \frac{1}{3}x(n-2).$$

The system transfer function of the whole system is

$$H(z) = H_1(z)H_2(z) = \frac{1 - \frac{1}{2}z^{-1} + \frac{1}{3}z^{-2}}{1 - \frac{1}{2}z^{-2} + \frac{1}{6}z^{-3}}.$$

6.1.3 Sensitivity of the System Poles/Zeros

Systems with a large number of elements in a recursion may be sensitive to the errors due to the coefficients deviations. Deviations of the coefficients from the true values are caused by finite length registers used to memorize them in a computer. Influence of the finite register lengths to the signal and system realization will be studied later, as a part of random disturbance analysis. Here, we will only consider influence of this effect to the system coefficients, since it may influence the way how to realize a discrete-time system.

For the first-order system with a real-valued pole

$$H(z) = \frac{1}{1 + A_1 z^{-1}} = \frac{1}{1 - z_{p1} z^{-1}}$$

the error in coefficient A_1 is the same as the error in the system pole z_{p1}. If the coefficient is quantized with a step Δ, then the error in the pole location is of order Δ. The same holds for the system zeros.

For a second-order system with real-valued coefficients and a pair of complex-conjugated poles

$$H(z) = \frac{1}{1 + A_1 z^{-1} + A_2 z^{-2}} = \frac{1}{(1 - z_{p1} z^{-1})(1 - z_{p2} z^{-1})}$$

the relation between the coefficients and the real and imaginary parts of the poles $z_{p1/2} = x_p \pm j y_p$ is

$$\begin{aligned} H(z) &= \frac{1}{1 - 2x_p z^{-1} + (x_p^2 + y_p^2) z^{-2}} \\ A_1 &= -2x_p \\ A_2 &= x_p^2 + y_p^2. \end{aligned}$$

The error in coefficient A_1 defines the error in the real part of the pole x_p.

When the coefficient A_2 takes discrete values $A_2 = m\Delta$, with $A_1 \sim x_p = n\Delta$, then the imaginary part of the poles may take the values $y_p = \pm\sqrt{A_2 - x_p^2} = \pm\sqrt{m\Delta - n^2\Delta^2}$ with $n^2 \leq mN$. For small n, that is, for small real part of the pole, $y_p = \pm\sqrt{\Delta m}$. For N discretization levels, assuming that the poles are within the unit circle $x_p^2 + y_p^2 \leq 1$, the first discretization step is changed from $1/N$ order to $1/\sqrt{N}$ order. The error, in this case, could significantly be increased. The changes in y_p due to the discretization of A_2 may be large.

The quantization of x_p and y_p as a result of quantization of $-A_1/2$ and $A_2 = x_p^2 + y_p^2$ is shown in Fig. 6.7, for the case of $N = 16$ and $N = 32$ quantization levels. We see that the error in y_p, when it assumes small values, can be very large. We can conclude that the poles close to the unit circle with larger imaginary values y_p are less sensitive to the errors. The highest error could appear if the second-order real-valued pole (with $y_p = 0$) were implemented using the second-order system.

We have concluded that the poles close to the real axis (small y_p) are sensitive to the error in coefficients even in the second-order systems. The sensitivity increases with the system order, since the higher powers in the polynomial increase the maximum possible error.

Consider a general form of a polynomial in the transfer function, written in two forms

$$P(z) = z^M + z^{M-1} A_1 + \cdots + A_M$$

and

$$P(z) = (z - z_1)(z - z_2) \ldots (z - z_M).$$

If the coefficients $A_1, A_2, \ldots, A_M$ are changed for small $\Delta A_1, \Delta A_2, \ldots, \Delta A_M$ (due to quantization) then the pole position (without loss of generality and for notation simplicity consider the pole z_1) is

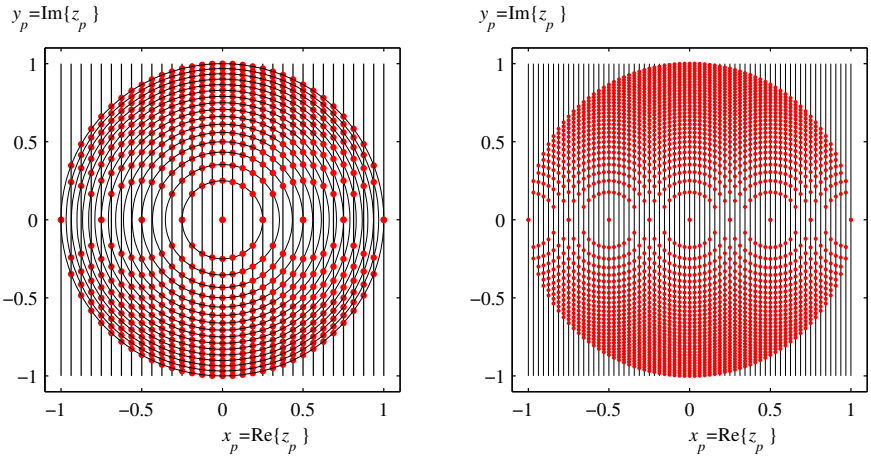


Figure 6.7 Quantization of the real part and the imaginary part, $x_p = \mathrm{Re}\{z_p\}$ and $y_p = \mathrm{Im}\{z_p\}$, of poles (zeros) as a result of the quantization in 16 levels (left) and 32 levels (right) of the coefficients $A_1 = -2x_p$ and $A_2 = x_p^2 + y_p^2$.

changed for

$$\Delta z_1 \cong \left[\frac{\partial z_1}{\partial A_1}\Delta A_1 + \frac{\partial z_1}{\partial A_2}\Delta A_2 + \cdots + \frac{\partial z_1}{\partial A_M}\Delta A_M\right]_{|z=z_1}. \tag{6.5}$$

Since there is no a direct relation between z_1 and A_1 we will find $\partial z_1/\partial A_i$ using

$$\frac{\partial P(z)}{\partial A_i}_{|z=z_1} = \frac{\partial P(z)}{\partial z_1}\frac{\partial z_1}{\partial A_i}_{|z=z_1}.$$

From this relation follows

$$\frac{\partial z_1}{\partial A_i}_{|z=z_1} = \frac{\frac{\partial P(z)}{\partial A_i}_{|z=z_1}}{\frac{\partial P(z)}{\partial z_1}_{|z=z_1}} = \frac{z_1^{M-i}}{-(z_1-z_2)(z_1-z_3)\ldots(z_1-z_M)}.$$

The coefficients $\partial z_1/\partial A_{i|z=z_1}$ could be large, especially in the case when there are close poles, with a small distance $(z_i - z_k)$.

Example 6.2. Consider the discrete-time system

$$H(z) = \frac{1}{P(z)} \tag{6.6}$$

$$P(z) = (z-\frac{12}{27})(z-\frac{7}{29})(z-\frac{111}{132})(z-\frac{95}{101})$$

$$\cong (z-0.4444)(z-0.2414)(z-0.8409)(z-0.9406) \tag{6.7}$$

In the realization of this system the coefficients are rounded to two decimal positions, with the absolute error up to 0.005. Find the poles of the system with rounded coefficients.

★The system denominator is

$$P(z) \cong z^4 - 2.4673z^3 + 2.1200z^2 - 0.7336z + 0.0849.$$

With coefficients rounded to two decimal positions we get

$$\hat{P}(z) = z^4 - 2.47z^3 + 2.12z^2 - 0.73z + 0.08$$

with poles

$$\hat{P}(z) = (z - 0.5370)(z - 0.2045)(z - 0.7285)(z - 1).$$

The poles of this function with rounded coefficients can differ significantly from the original pole values in (6.7). The maximum error in poles is $0.8409 - 0.7285 = 0.1124$. One pole is on the unit circle making the system with rounded coefficients unstable, in contrast to the stable original system. Note that if the system is written as a product of the first-order functions in the denominator and every pole value is rounded to two decimals

$$H(z) = \frac{1}{(z - \frac{7}{29})(z - \frac{12}{27})(z - \frac{111}{132})(z - \frac{95}{101})}$$
$$P(z) \cong (z - 0.24)(z - 0.44)(z - 0.84)(z - 0.94)$$

the poles will differ from the original ones for no more than 0.005.

If the poles are grouped into the second-order terms (what should be done if the coefficients were complex-conjugate in order to avoid calculation with complex valued coefficients), then

$$P(z) \cong (z^2 - 0.6858z + 0.1073)(z^2 - 1.7815z + 0.7910).$$

If the coefficients are rounded to two decimal positions

$$\hat{P}(z) = (z^2 - 0.69z + 0.11)(z^2 - 1.78z + 0.79)$$

we will get

$$\hat{P}(z) = (z - 0.25)(z - 0.44)(z - 0.8442)(z - 0.9358)$$

with maximum error of 0.01.

The pole values are illustrated in Fig. 6.8.

The sensitivity analysis for this example can be done for each of the poles. Assume that the poles are denoted by $z_1 = 12/27$, $z_2 = 7/29$, $z_3 = 111/132$, and $z_4 = 95/101$. Then,

$$(z_1 - z_2)(z_1 - z_3)(z_1 - z_4) = 0.0399$$
$$\frac{\partial z_1}{\partial A_1}_{|z=z_1} = \frac{z_1^{4-1}}{-(z_1 - z_2)(z_1 - z_3)(z_1 - z_4)} = -2.1979$$
$$\frac{\partial z_1}{\partial A_2}_{|z=z_1} = \frac{z_1^{4-2}}{-(z_1 - z_2)(z_1 - z_3)(z_1 - z_4)} = -4.9452,$$
$$\frac{\partial z_1}{\partial A_3}_{|z=z_1} = -11.1267,$$
$$\frac{\partial z_1}{\partial A_4}_{|z=z_1} = -25.0350$$

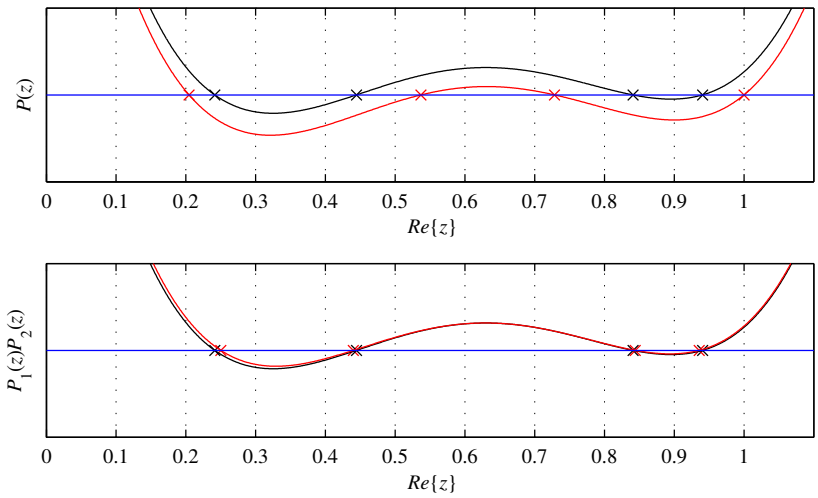


Figure 6.8 Poles for a system with errors in coefficients: for the fourth-order polynomial (top) and the product of the two second-order polynomials (bottom).

with the errors in the coefficients

$$\Delta A_1 = -2.4673 - (-2.47) = 0.0027,$$
$$\Delta A_2 = 2.12 - 2.12 = 0,$$
$$\Delta A_3 = -0.7336 - (-0.73) = -0.0036,$$
$$\Delta A_4 = 0.0849 - 0.08 = 0.0049.$$

Replacing these values into (6.5) the approximation of the error is

$$\Delta z_1 \cong 0.0878.$$

The true error is $\Delta z_1 = 0.0926$. A small difference is due to the linear approximation, assuming small ΔA_i. The obtained result is a good estimate of an order of error for the pole z_1. The error in z_1 is about 18.5 time greater than the maximum error in the coefficients A_i, which is of order 0.005.

6.1.4 Cascade Realization

A transfer function of the discrete-time system in (6.2), with $M = N$, might be written as a product of the first-order subsystems

$$H(z) = k\frac{1 - z_{o0}z^{-1}}{1 - z_{p0}z^{-1}} \times \frac{1 - z_{o1}z^{-1}}{1 - z_{p1}z^{-1}} \cdots \times \frac{1 - z_{oN}z^{-1}}{1 - z_{pN}z^{-1}}.$$

Commonly real-valued signals are processed and the poles and zeros in the transfer function are in complex-conjugated pairs. In that case it is better to group these pairs into second-order systems to

avoid complex calculations. The transfer function is then of the form

$$H(z) = \frac{B_{00} + B_{10}z^{-1} + B_{20}z^{-2}}{1 - A_{10}z^{-1} - A_{20}z^{-2}} \times \cdots \times \frac{B_{0K} + B_{1K}z^{-1} + B_{2K}z^{-2}}{1 - A_{1K}z^{-1} - A_{2K}z^{-2}} = H_0(z)H_1(z)\ldots H_K(z),$$

where

$$H_i(z) = \frac{B_{0i} + B_{1i}z^{-1} + B_{2i}z^{-2}}{1 - A_{1i}z^{-1} - A_{2i}z^{-2}}$$

are second-order systems with real-valued coefficients. The whole system may be realized as a cascade of lower-order (first or second-order) systems, Fig. 6.9. Of course, if there are some real-valued poles then there is no need to group them. It is better to keep the realization order of the subsystems as low as possible, as shown in Section 6.1.3.

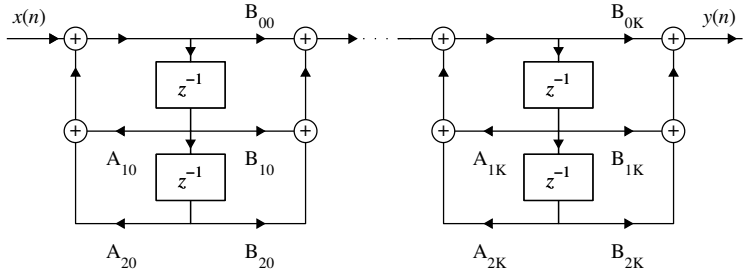


Figure 6.9 Cascade realization of a discrete-time system.

In the realization, the second-order subsystems are commonly used. Note that it is possible to realize these second-order subsystems using the first-order systems with real-valued coefficients x_{pL} and y_{pL} that are real and imaginary parts of the complex-conjugated pair of poles, $z_{pL} = x_{pL} \pm jy_{pL}$, respectively. To this aim consider first an example.

Example 6.3. Find the transfer function of a system with a feedback shown in Fig. 6.10.

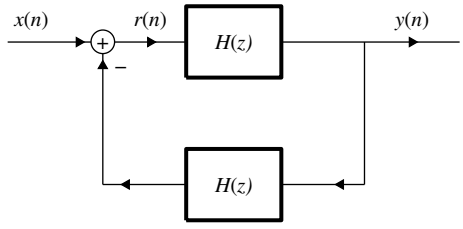


Figure 6.10 System with a feedback.

★The z-transform of the signal at the output of adder the in the system form Fig. 6.10 is

$$R(z) = X(z) - H(z)Y(z).$$

The z-transform of the output signal is equal to

$$Y(z) = H(z)R(z) = H(z)X(z) - H^2(z)Y(z).$$

The transfer function of this system is

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

Let us now consider a realization of the second-order subsystem of the form

$$Q_i(z) = \frac{y_{pL}z^{-1}}{1 + A_{1i}z^{-1} + A_{2i}z^{-2}}.$$

Using the real and imaginary part of the complex-conjugate poles $z_{pL} = x_{pL} \pm jy_{pL}$, the transfer function, $Q_i(z)$, can be expressed as

$$\begin{aligned} Q_i(z) &= \frac{y_{pL}z^{-1}}{1 - 2x_{pL}z^{-1} + x_{pL}^2z^{-2} + y_{pL}^2z^{-2}} = \frac{y_{pL}z^{-1}}{(1 - x_{pL}z^{-1})^2 + y_{pL}^2z^{-2}} \\ &= y_{pL}z^{-1}\frac{1}{(1 - x_{pL}z^{-1})^2}\frac{1}{1 + \left(\frac{y_{pL}z^{-1}}{1 - x_{pL}z^{-1}}\right)^2} = \frac{H(z)H_2(z)}{1 + H^2(z)} \end{aligned}$$

where

$$H(z) = \frac{y_{pL}z^{-1}}{1 - x_{pL}z^{-1}} \quad \text{and} \quad H_2(z) = \frac{1}{1 - x_{pL}z^{-1}}.$$

Therefore, the second-order system can be implemented as in Fig. 6.11, using the first-order systems shown in Fig. 6.12. In this case there is no grouping of the coefficients into the second-order polynomials.

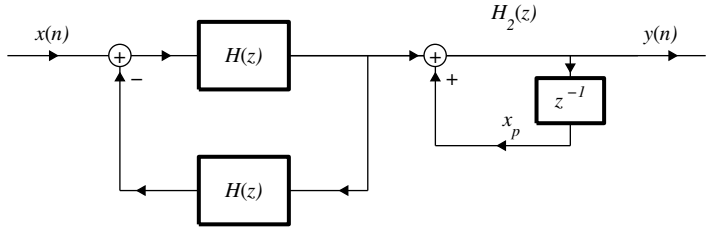


Figure 6.11 Complete second-order subsystem with the complex-conjugate pair of poles realized using the first-order systems.

The error in one coefficient (real or imaginary part of a pole) does not influence the other coefficients. However, if an error in the signal calculation happens in one cascade, then it will propagate as an input to the following cascades. In that sense, it would be the best to order cascades in such a way that the lowest probability of an error appears in the first cascade. From the analysis of error we can conclude that the cascades with the poles and zeros close to the origin are more sensitive to the error and should be used in later cascade stages.

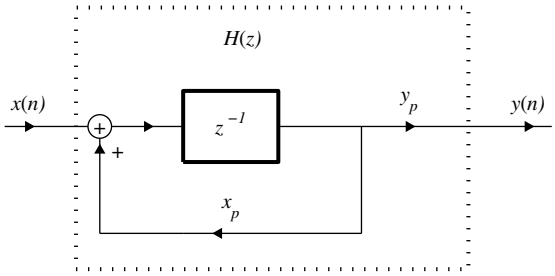


Figure 6.12 First-order system used in the realization of the second-order system with the complex-conjugate pair of poles.

Example 6.4. For the system

$$H(z) = \frac{1.4533(1+z^{-1})^3}{(-0.8673z^{-1}+3.1327)(3.0177z^{-2}-5.434z^{-1}+7.54)}$$
$$= 0.0615\frac{1+z^{-1}}{1-0.2769z^{-1}} \times \frac{1+2z^{-1}+z^{-2}}{1-0.7207z^{-1}+0.4002z^{-2}}$$

present the cascade realization using:

(a) both the first and the second-order systems;

(b) the first-order systems with real-valued coefficients only.

★(a) Realization of the system $H(z)$ when both the first and the second-order subsystems can used is done according to the system transfer function as in Fig. 6.13.

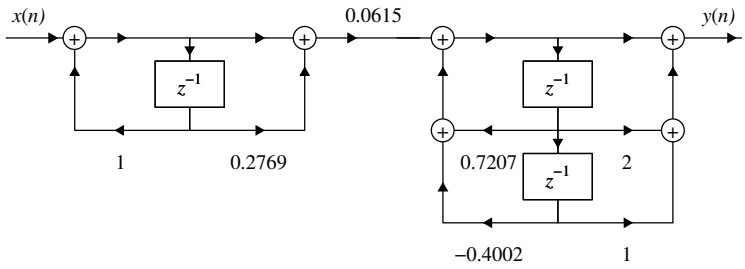


Figure 6.13 Cascade realization of a system.

(b) For the first-order subsystems, the realization should be done based on

$$H(z) = 0.0615\frac{1+z^{-1}}{1-0.2769z^{-1}} \times (1+z^{-1}) \times (1+z^{-1}) \times \frac{1}{1-0.7207z^{-1}+0.4002z^{-2}},$$

with

$$\frac{1}{1-0.7207z^{-1}+0.4002z^{-2}} = \frac{1}{(1-(0.3603+j0.5199)z^{-1})(1-(0.3603-j0.5199)z^{-1})}$$
$$= \frac{1}{1-2\times 0.3603z^{-1}+0.3603^2z^{-2}+0.5199^2z^{-2}}$$
$$= \frac{1}{0.5199^2z^{-2}+(1-0.3603z^{-1})^2} = \frac{1}{(1-0.3603z^{-1})^2}\frac{1}{1+(\frac{0.5199z^{-1}}{1-0.3603z^{-1}})^2}.$$

In this way, the system can be written and realized in terms of the first-order subsystems,

$$H(z) = 0.0615\frac{1+z^{-1}}{1-0.2769z^{-1}}\frac{1+z^{-1}}{1-0.3603z^{-1}}\frac{1+z^{-1}}{1-0.3603z^{-1}}\frac{1}{1+\frac{0.5199z^{-1}}{1-0.3603z^{-1}}\frac{0.5199z^{-1}}{1-0.3603z^{-1}}}$$

as shown in Fig. 6.14.

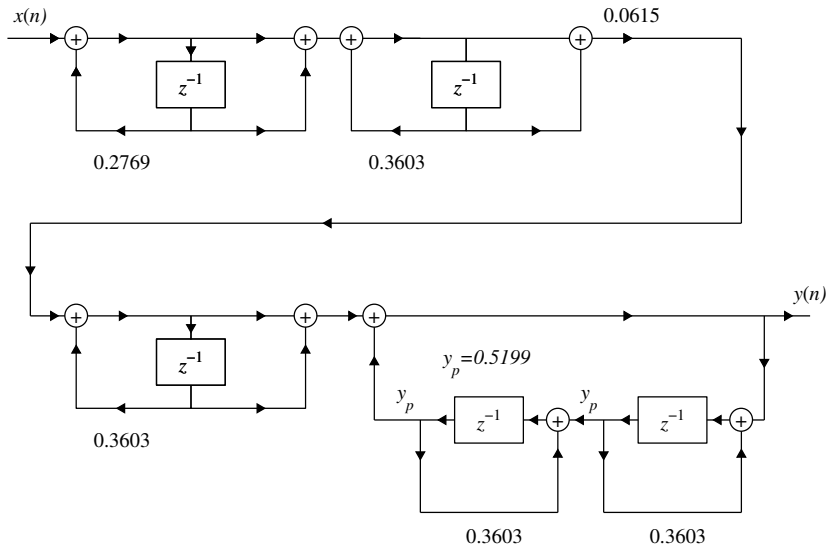


Figure 6.14 Discrete-time system realized using the first-order subsystems.

6.1.5 Parallel realization

This realization is implemented based on a transfer function written in the form

$$H(z) = \frac{B_{00}+B_{10}z^{-1}+B_{20}z^{-2}}{1-A_{10}z^{-1}-A_{20}z^{-2}} + \cdots + \frac{B_{0K}+B_{1K}z^{-1}+B_{2K}z^{-2}}{1-A_{1K}z^{-1}-A_{2K}z^{-2}} = H_0(z)+\cdots+H_K(z).$$

In the case of a parallel realization the error in one subsystem does not influence the other subsystems. If an error in the signal calculation appears in one parallel subsystem, then it will influence the output signal, but will not influence the outputs of other parallel subsystems.

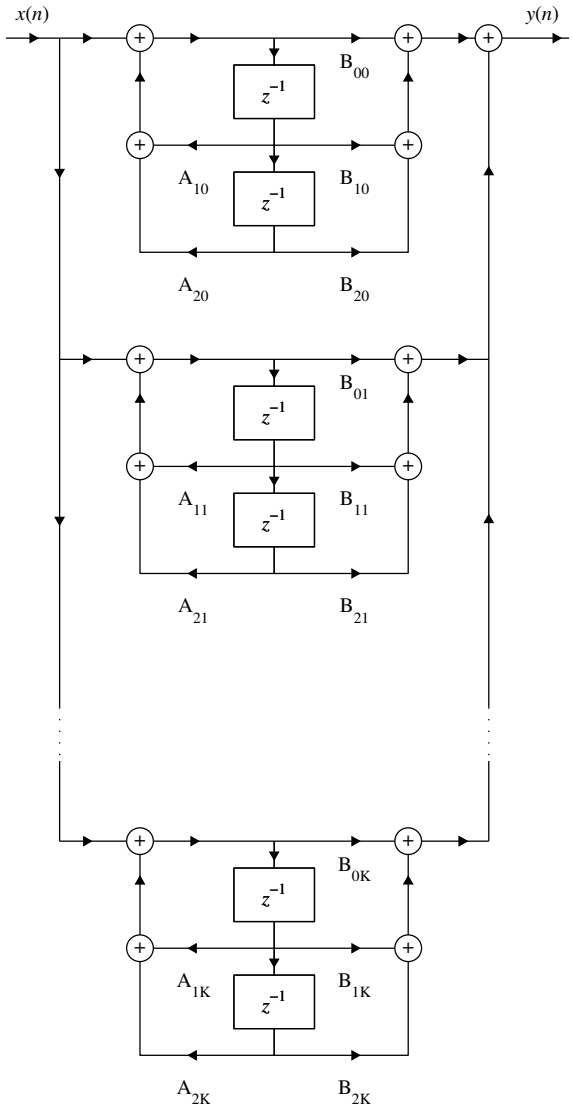


Figure 6.15 Parallel realization of a discrete-time system.

Example 6.5. For the system

$$H(z) = \frac{-0.7256 + 0.2542z^{-1}}{1 - 1.1078z^{-1} + 0.5482z^{-2}} + \frac{0.7256 - 0.084z^{-1}}{1 - 0.9246z^{-1} + 0.2343z^{-2}}$$

present a parallel and a cascade realization using the second-order subsystems.

★ The parallel realization follows directly from the system transfer function definition. It is presented in Fig. 6.16.

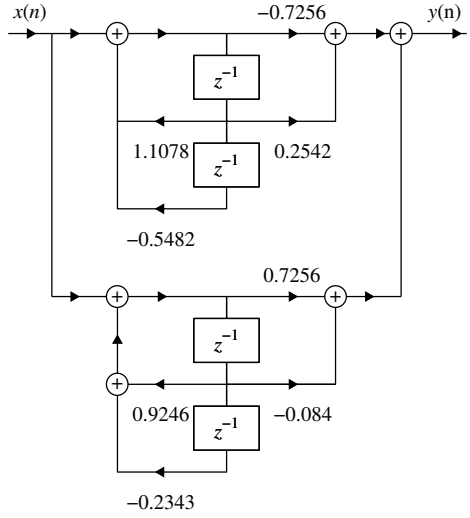


Figure 6.16 Parallel realization of a discrete-time system.

For the cascade realization, the system transfer function should be written in a form of the product of the second-order transfer functions,

$$H(z) = \frac{0.0373z^{-1} + 0.0858z^{-2} + 0.0135z^{-3}}{\left(1 - 1.1078z^{-1} + 0.5482z^{-2}\right)\left(1 - 0.9246z^{-1} + 0.2343z^{-2}\right)}$$
$$= \frac{z^{-1}}{1 - 1.1078z^{-1} + 0.5482z^{-2}} \frac{0.0373 + 0.0858z^{-1} + 0.0135z^{-2}}{1 - 0.9246z^{-1} + 0.2343z^{-2}}.$$

The cascade realization is presented in Fig. 6.17.

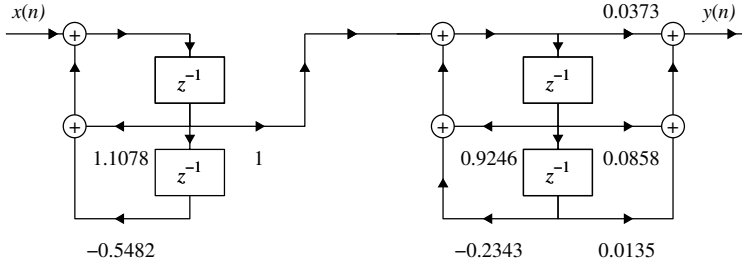


Figure 6.17 Cascade realization of a discrete system.

6.1.6 Inverse realization

For each of the previous realization, an inverse form may be implemented by switching the input and the output signal and changing the flow directions of the signal. As an example, consider the direct realization II from Fig. 6.4. This realization, with separated delay circuits is shown in Fig. 6.18. Its inverse form is presented in Fig. 6.19.

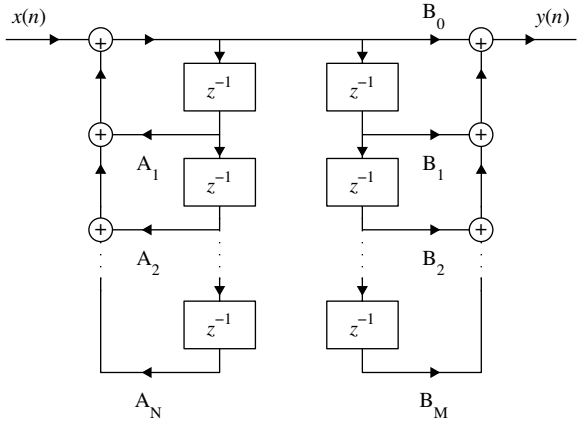


Figure 6.18 Direct realization II with separated delay circuits.

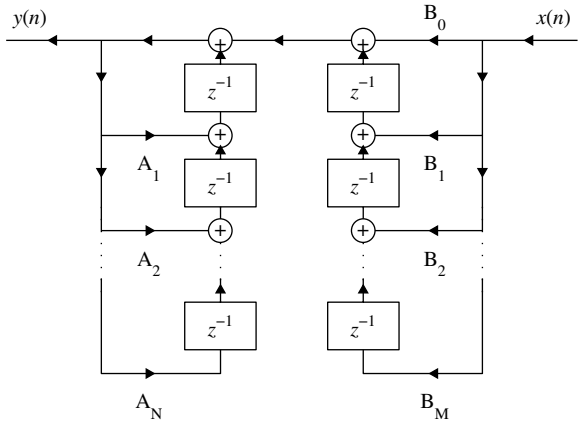


Figure 6.19 Inverse realization of the direct realization II.

It is easy to conclude that the inverse realization of the direct realization II has the same transfer function as the direct realization I. Since both realization I and realization II have the same transfer functions it follows that the inverse realization has the same transfer function as the original realization.

6.2 FIR SYSTEMS AND THEIR REALIZATIONS

In general, transfer functions of discrete-time systems are obtained in the form of a ratio of two polynomials. The polynomial in the transfer function denominator defines poles. In the time domain, this means a recursive relation, relating the output signal at the current instant with the previous output signal values. Realization of this kind of systems is efficient, as described in the previous section. When the output signal is a linear combination of the input signal, $x(n)$, and its delayed versions, $x(n-m)$, only, the systems does not have recursions. Its difference equations is

$$y(n) = B_0 x(n) + B_1 x(n-1) + \cdots + B_M x(n-M).$$

This system is characterized with a finite impulse response, and it is referred to as the FIR system. This system is always stable. The FIR systems can also have a linear phase.

6.2.1 Linear Phase Systems and Group Delay

In an implementation of a discrete-time system it is important to modify the amplitude of the Fourier transform of the input signal in a desired way. At the same time, we should take care about the phase function changes in the input signal. In an ideal case of the signal filtering, the phase function of the input signal should remain the same, meaning a zero-phase transfer function. A linear phase form of the transfer function

$$\arg\{H(e^{j\omega})\} = \arctan\{\frac{\mathrm{Im}\{H(e^{j\omega})\}}{\mathrm{Re}\{H(e^{j\omega})\}}\} = -\omega q \tag{6.8}$$

is also acceptable in these systems. They will have a constant group delay

$$\tau_g = -\frac{d(\arg\{H(e^{j\omega})\})}{d\omega} = q$$

and will not distort the impulse response with respect to the zero-phase system. The impulse response will only be delayed in time for a constant q.

Example 6.6. Consider an input signal of the form

$$x(n) = \sum_{m=1}^{M} A_m e^{j(\omega_m n + \theta_m)}.$$

After passing through a system with the frequency response $H(e^{j\omega})$, this signal is changed to

$$y(n) = \sum_{m=1}^{M} A_m |H(e^{j\omega_m})| e^{j(\omega_m n + \theta_m + \arg\{H(e^{j\omega_m})\})}.$$

In general, the phase of every signal component is changed in a different way for $\arg\{H(e^{j\omega_m})\}$, causing the signal distortion due to different delays corresponding to different frequencies. If the phase function of the frequency response is linear then all signal component phases are changed in the same way for $\arg\{H(e^{j\omega_m})\} = -\omega_m q$. They corresponding to a constant delay for all components. A delayed signal, without distortion, is obtained

$$y(n) = y_0(n-q) = \sum_{m=1}^{M} A_m |H(e^{j\omega_m})| e^{j(\omega_m (n-q) + \theta_m)},$$

where $y_0(n)$ would be the response if the phase of the transfer function were 0. In the case of a linear phase $\arg\{H(e^{j\omega})\} = -\omega q$ the phase delay

$$\tau_\varphi = -\frac{\arg\{H(e^{j\omega})\}}{\omega} = q$$

and the group delay τ_g are the same. In general, the group delay and the phase delay are different. The group delay, as the notion dual to the instantaneous frequency, is introduced and discussed in the first chapter of this book.

Consider a system with a real-valued impulse response $h(n)$. Its frequency response is

$$H(e^{j\omega}) = \sum_{n=0}^{N-1} h(n)e^{-j\omega n} = \sum_{n=0}^{N-1} h(n)\cos(\omega n) - j\sum_{n=0}^{N-1} h(n)\sin(\omega n). \tag{6.9}$$

Combining the linear phase condition (6.8) with the form in (6.9), we get

$$-\tan(\omega q) = \frac{\mathrm{Im}\{H(e^{j\omega})\}}{\mathrm{Re}\{H(e^{j\omega})\}} = -\frac{\sum_{n=0}^{N-1} h(n)\sin(\omega n)}{\sum_{n=0}^{N-1} h(n)\cos(\omega n)},$$

or

$$\sum_{n=0}^{N-1} h(n)[\sin(\omega q)\cos(\omega n) - \cos(\omega q)sin(\omega n)] = 0.$$

The last equation can be written as

$$\sum_{n=0}^{N-1} h(n)\sin(\omega(n-q)) = 0. \tag{6.10}$$

The middle point of the interval where $h(n) \neq 0$ is $n = (N-1)/2$. If $q = (N-1)/2$, then $\sin(\omega(n-q))$ is an odd function with respect to $n = (N-1)/2$. The summation (6.10) is zero if the impulse response $h(n)$ is an even function with respect to $n = (N-1)/2$. Hence, the solution to (6.10) is

$$q = \frac{N-1}{2}$$
$$h(n) = h(N-1-n),\ 0 \leq n \leq N-1.$$

Since the Fourier transform is unique, this is the unique solution for the linear phase condition. It is illustrated for an even and odd N in Fig. 6.20. From the symmetry condition, it is easy to conclude that there is no a causal linear phase system with an infinite impulse response.

6.2.2 Windows

When a system obtained from the design procedure is an IIR system and the requirement is to implement it as an FIR system, in order to get a linear phase or to guaranty the system stability (when small changes of the coefficients are possible), then the most obvious way is to truncate the desired impulse response $h_d(n)$ of the resulting IIR system. The impulse response of the FIR system is

$$h(n) = \begin{cases} h_d(n), & \text{for } 0 \leq n \leq N-1 \\ 0, & \text{elsewhere.} \end{cases}$$

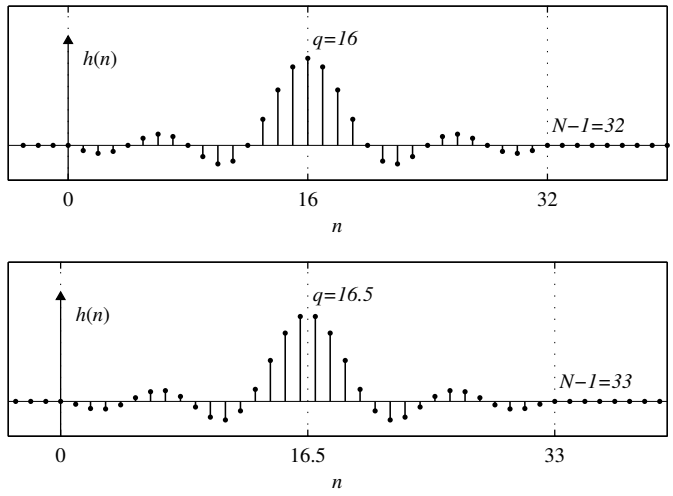


Figure 6.20 The impulse response of a system with the linear phase for an odd and even N.

This form can be written as

$$h(n) = h_d(n)w(n),$$

where

$$w(n) = \begin{cases} 1 \text{ for } 0 \leq n \leq N-1 \\ \\ 0 \quad \text{elsewhere} \end{cases}$$

is a rectangular window function. In the Fourier domain, the desired impulse response truncation by a window function will mean a convolution of the desired frequency response with the frequency response of the window function

$$H(e^{j\omega}) = H_d(e^{j\omega}) *_\omega W(e^{j\omega}).$$

Since the rectangular window function has the Fourier transform of the form

$$W(e^{j\omega}) = \sum_{n=0}^{N-1} e^{-j\omega n} = e^{-j\omega(N-1)/2} \frac{\sin(\omega N/2)}{\sin(\omega/2)},$$

its convergence is slow, with significant oscillations. This oscillations will cause oscillations in the resulting frequency response $H(e^{j\omega})$, Fig. 6.21. By increasing the number of samples N, the convergence speed will increase. However the oscillations amplitude will remain the same, Figs.6.21 (d) and (f). Even with $N \to \infty$ the amplitude oscillations will remain, Figs.6.21 (b). This effect is called the Gibbs phenomenon.

Example 6.7. The desired frequency response of a system is $H_d(e^{j\omega})$, with the IIR $h_d(n)$ for $-\infty < n < \infty$. Find the FIR system impulse response $h_c(n)$ that approximates the desired transfer function with a minimum mean absolute squared error.

★The mean squared absolute error is

$$e^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|H_d(e^{j\omega}) - H_c(e^{j\omega})\right|^2 d\omega.$$

According to Parseval's theorem

$$e^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|H_d(e^{j\omega}) - H_c(e^{j\omega})\right|^2 d\omega = \sum_{n=-\infty}^{\infty}|h_d(n) - h_c(n)|^2.$$

Without loss of generality, assume that the most significant values of $h_d(n)$ are within $-N/2 \le n \le N/2-1$. The impulse response $h_c(n)$ can take nonzero values only within $-N/2 \le n \le N/2-1$. Therefore,

$$e^2 = \sum_{n=-N/2}^{N/2-1}|h_d(n) - h_c(n)|^2 + \sum_{n=-\infty}^{-N/2-1}|h_d(n)|^2 + \sum_{n=N/2}^{\infty}|h_d(n)|^2.$$

Since the last two terms are $h_c(n)$ independent and all three terms are non negative, the error e^2 is minimum if

$$h_c(n) = h_d(n),\ -N/2 \le n \le N/2-1.$$

If we want to have a causal realization of the FIR system then

$$h(n) = h_c(n - N/2).$$

A shift in time does not change the amplitude of the desired frequency response, since $\left|H(e^{j\omega})\right| = \left|H_c(e^{j\omega})\right|$.

In order to reduce the oscillations in the frequency response amplitude other windows are introduced. They are presented within the introductory chapters, trough the examples. Here we will list the basic windows (more details on the window functions will be given in Part V).

Triangular (Bartlett) window is defined as

$$w(n) = \begin{cases} 1 - \frac{|n+1-N/2|}{N/2}, & \text{for } 0 \le n \le N-1 \\ 0, & \text{elsewhere.} \end{cases}$$

Avoiding window discontinuities at the ending points, the convergence of its transform is improved. Since this window may be considered as a convolution of the two rectangular windows

$$w(n) = \frac{1}{N/2}[u(n) - u(n-N/2)] *_n [u(n) - u(n-N/2)]$$

its the Fourier transform is the product of the corresponding rectangular window Fourier transforms

$$W(e^{j\omega}) = \frac{1}{N/2}e^{-j\omega(N/2-1)}\frac{\sin^2(\omega N/4)}{\sin^2(\omega/2)}.$$

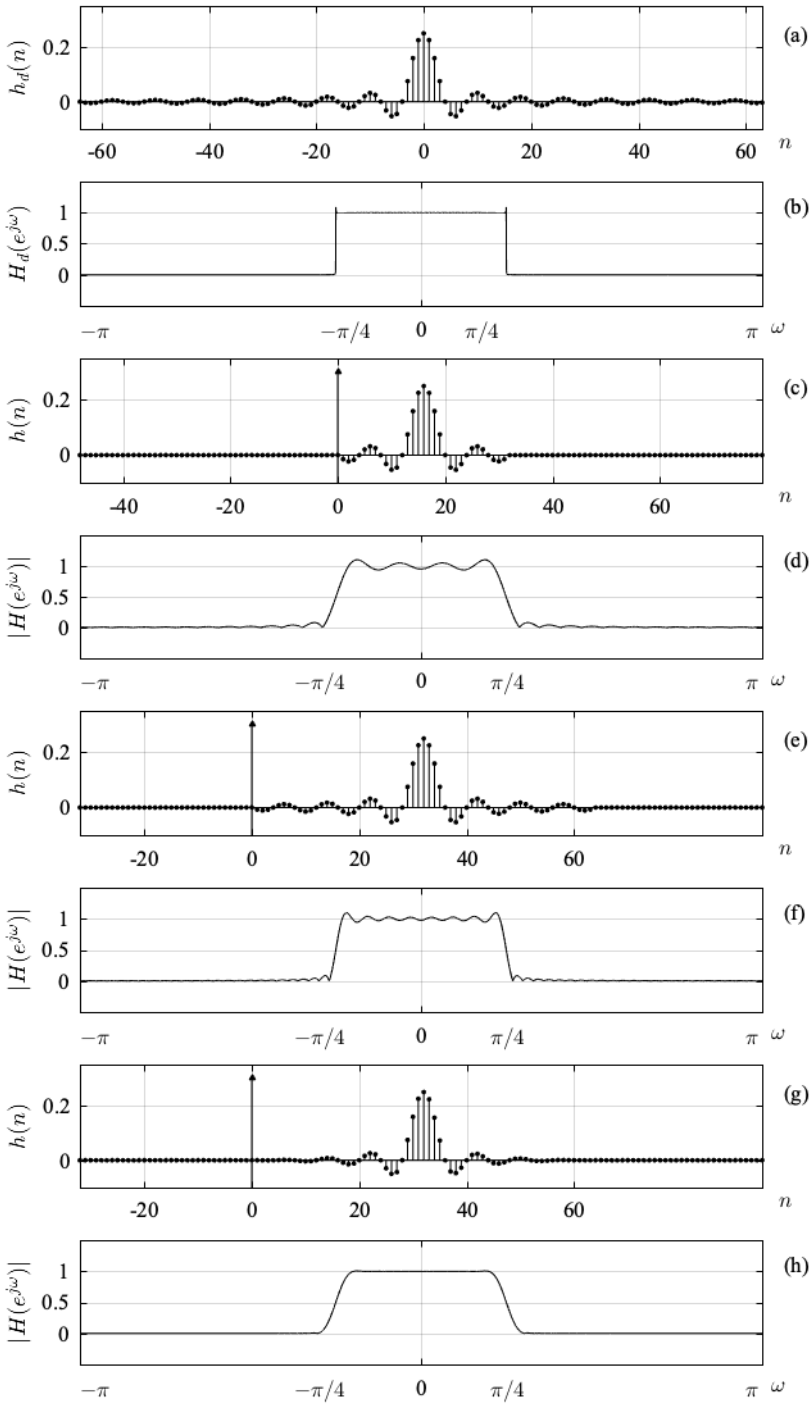


Figure 6.21 Impulse response of a FIR system obtained by truncating the desired IIR response (a), (b) using two rectangular windows of different widths (c)-(f), and using a Hann(ing) window (g),(h).

Hann(ing) window defined by

$$w(n) = \begin{cases} \frac{1}{2}\left(1 + \cos((n - N/2)\frac{2\pi}{N}\right)), & \text{for } 0 \le n \le N-1 \\ 0, & \text{elsewhere} \end{cases}$$

would be continuous in the continuous-time domain. In that domain its first derivative would be continuous as well. Thus, its Fourier domain convergence is further improved with respect to the rectangular and the Bartlett windows. The Fourier transform of this window is related to the Fourier transform of the rectangular window as $W(e^{j\omega})/2 + W(e^{j(\omega+2\pi/N)})/4 + W(e^{j(\omega-2\pi/N)})/4$.

Hamming window is a slight modification of the Hann(ing) window

$$w(n) = \begin{cases} 0.52 + 0.48\cos((n - N/2)\frac{2\pi}{N}) & \text{for } 0 \le n \le N-1 \\ 0, & \text{elsewhere.} \end{cases}$$

It loses the continuity property (in the continuous-time domain). Its convergence for very large values of ω will be slower than in the Hann(ing) window case. However, as it will be shown later, its coefficients are derived in such a way that the first side-lobe is canceled out at its mid point. Then, the immediate convergence, after the main lobe, is much better than in the Hann(ing) window case.

Other windows are derived with other different constraints. Some of them will be reviewed in Part V of this book as well.

6.2.3 Design of a FIR System in the Frequency Domain

Suppose that the desired system frequency response is given in the frequency domain. If we want to get an N point FIR system that approximates the desired frequency response, then it can be obtained by sampling the desired frequency response $H_d(e^{j\omega})$ at $\omega = 2\pi k/N$, $k = 0,1,2,\ldots,N-1$, that is

$$H(k) = H_d(e^{j\omega})_{|\omega=2\pi k/N}.$$

Then, the impulse response of the FIR system is

$$h(n) = \text{IDFT}\{H(k)\}.$$

This procedure is illustrated on a lowpass filter design, Fig. 6.22. Note that at the discontinuity points, high oscillations will occur in the resulting $H(e^{j\omega})$. The oscillations can be avoided by smoothing the transition intervals. Smoothing by a Hann(ing) window in the frequency domain is shown in Fig. 6.23.

6.2.4 Realizations of the FIR systems

The FIR systems can be realized in the same way as the IIR systems presented in the previous section, without using the recursive coefficients. A common way of presenting the direct realization of the FIR system is shown in Fig. 6.24. It is often referred to as an adder with the weighting coefficients $h(n)$.

A realization of liner phase FIR system that uses the coefficients symmetry $h(0) = h(N-1)$, $h(2) = h(N-2)$, ... property is shown in Fig. 6.25.

Realization of a frequency sampled FIR filter may be done using the relation between the $z-$transform and the DFT of a signal.

If we want to realize a FIR system with N nonzero samples, then it can be expressed in term of the DFT of the frequency response (samples of the transfer function $H(z)$ along the unit circle) as

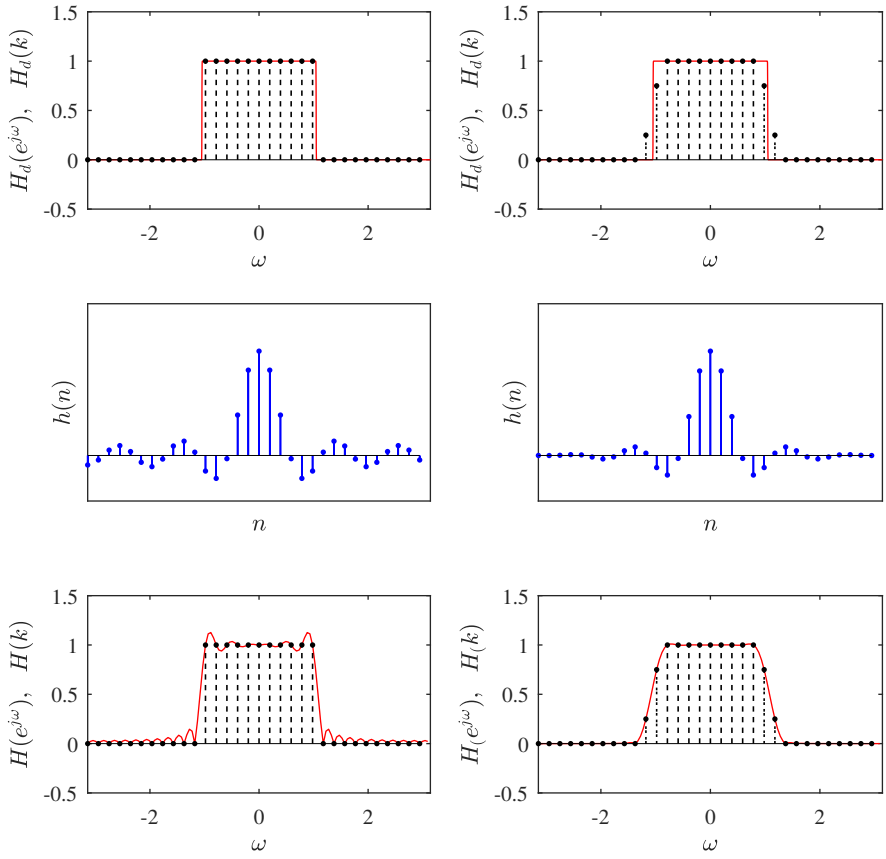


Figure 6.22 Realization of a FIR system with N samples in time, obtained by sampling the desired frequency response with N samples. A direct sampling (left) and the sampling with smoothed transition (right),

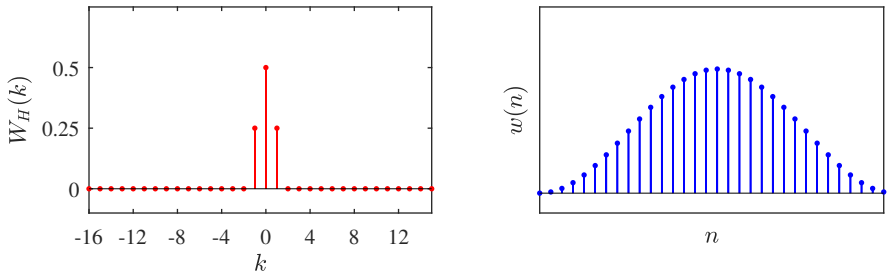


Figure 6.23 A Hann(ing) window for smoothing the frequency response in the frequency domain (left) and in the time domain (right).

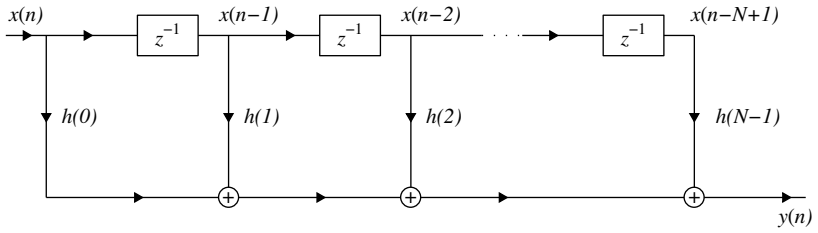


Figure 6.24 Direct realization of a FIR system.

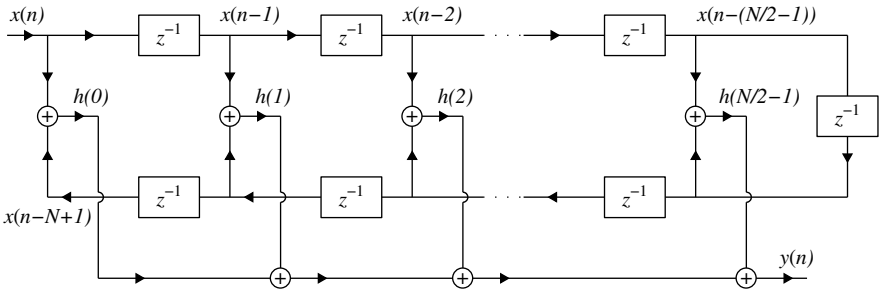


Figure 6.25 Direct realization of a FIR system with a linear phase.

follows. For a FIR filter we may write

$$H(k) = \sum_{n=0}^{N-1} h(n) e^{-j2\pi nk/N}$$

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

Then, the transfer function $H(z)$ calculated using the values of $h(n)$, $0 \leq n \leq N-1$, is

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

with $H(k) = H(z)$ for $z = \exp(j2\pi k/N)$, and $k = 0,1,2,\ldots,N-1$.

Example 6.8. For a system whose impulse response is the Hamming window function of the length $N = 32$ present the FIR filter based realization.

★For the Hamming window with $N = 32$, the impulse response is given by

$$h(n) = 0.52 + 0.48\cos((n-16)\frac{\pi}{16}), \quad \text{for} \quad 0 \le n \le 31.$$

The DFT values are $H(0) = 0.52 \times 32$, $H(1) = -0.24 \times 32$, $H(31) = H(-1) = -0.24 \times 32$, and $H(k) = 0$ for other k within $0 \le k \le 31$. Therefore,

$$\begin{aligned} H(z) &= \frac{1}{32}\frac{1-z^{-32}}{1-z^{-1}}H(0) - \frac{1}{32}\frac{1-z^{-32}e^{j2\pi}}{1-z^{-1}e^{j2\pi/32}}H(1) - \frac{1}{32}\frac{1-z^{-32}e^{-j2\pi}}{1-z^{-1}e^{-j2\pi/32}}H(31) \\ &= \frac{1}{32}(1-z^{-32})\left(\frac{H(0)}{1-z^{-1}} - \frac{2H(1)(1-\cos(\pi/16)z^{-1})}{1-2\cos(\pi/16)z^{-1}+z^{-2}}\right). \end{aligned}$$

This is a cascade of he system

$$H_1(z) = (1-z^{-32})/32$$

and the system $H_2(z) + H_3(z)$ where

$$H_2(z) = H(0)/(1-z^{-1})$$

and

$$H_3(z) = -2H(1)\frac{1-\cos(\pi/16)z^{-1}}{1-2\cos(\pi/16)z^{-1}+z^{-2}}.$$

Example 6.9. For the system whose frequency response $H_d(j\Omega)$ in the continuous-time domain is

$$H_d(j\Omega) = \pi - |\Omega|,$$

for $|\Omega| \le \pi$, with the corresponding $H_d(e^{j\omega})$ in the discrete-time domain ($\Delta t = 1$ is assumed, Fig. 6.26) find the FIR filter impulse response with $N = 7$ and $N = 8$:

(a) Sampling of the desired frequency response $H_d(e^{j\omega})$ in the frequency domain.

(b) Calculating $h_d(n) = \text{IFT}\{H_d(e^{j\omega})\}$ and taking its N the most significant values, $h(n) = h_d(n)$ for $-N/2 \le n \le N/2-1$ and $h(n) = 0$ elsewhere (using rectangular window).

(c) Comment the error in both cases.

★(a) Sampling in the frequency domain is illustrated in Fig. 6.26. The values of the FIR system frequency response, in this case, are the samples of $H_d(e^{j\omega})$,

$$H(k) = H_d(e^{j\omega})\Big|_{\omega=2\pi k/N} = \begin{cases} \pi(1-2\frac{k}{N}), \text{for } 0 \le k < N/2 \\ \\ \pi(2\frac{k}{N}-1), \text{for } N/2 \le k \le N-1. \end{cases}.$$

The sampling is illustrated in the second row of Fig. 6.26 for $N = 7$ and $N = 8$. Impulse response of the FIR filter is

$$h(n) = \text{IDFT}\{H(k)\} = \frac{1}{N}\sum_{k=0}^{N-1} H(k)e^{j2\pi nk/N}.$$

For $N = 7$

$$h(n) = \frac{\pi}{7} + \frac{10\pi}{49}\cos(\frac{2\pi}{7}n) + \frac{6\pi}{49}\cos(2\frac{2\pi}{7}n) + \frac{2\pi}{49}\cos(3\frac{2\pi}{7}n),\ \ 0 \le n \le 6.$$

For $N = 8$

$$h(n) = \frac{\pi}{8} + \frac{3\pi}{16}\cos(\frac{2\pi}{8}n) + \frac{\pi}{8}\cos(2\frac{2\pi}{8}n) + \frac{\pi}{16}\cos(3\frac{2\pi}{8}n),\ \ 0 \le n \le 7.$$

These impulse responses are shown in Fig. 6.26 (third row). The frequency response of the FIR filter is

$$H(e^{j\omega}) = \mathrm{FT}\{h(n)\}.$$

Its values are equal to the desired frequency response at the sampling points

$$H(e^{j\omega})\Big|_{\omega=2\pi k/N} = H_d(e^{j\omega})\Big|_{\omega=2\pi k/N}.$$

Outside these points, the frequency responses significantly differ (calculate, for example the values $H(e^{j0})$, $H(e^{j\pi/2})$, and $H(e^{j\pi})$). Here, there is no significant discontinuity in the frequency response. It means that the frequency response smoothing, using a window (Hann(ing) or Hamming window in the time domain), would not significantly improve the result.

(b) The impulse response of the desired system is

$$h_d(n) = \mathrm{IFT}\{H_d(e^{j\omega})\} = \frac{1}{2\pi}\int_{-\pi}^{\pi}(\pi - |\omega|)\,e^{j\omega n}d\omega$$

$$= \frac{2}{2\pi}\int_{0}^{\pi}(\pi - \omega)\cos(\omega n)d\omega = \frac{1-\cos(n\pi)}{\pi n^2}.$$

Using the first $N = 7$ samples in the time domain we get

$$h(n) = \begin{cases} \frac{1-\cos(n\pi)}{\pi n^2}, \text{ for } -3 \le n \le 3 \\ 0. \qquad\quad \text{elsewhere.} \end{cases}$$

For $N = 8$, the impulse response $h(n)$ is the same with $-4 \le n \le 3$.

The frequency response of this FIR filter is

$$H(e^{j\omega}) = \mathrm{FT}\{h(n)\},$$

and it is shown in Fig. 6.27.

(c) The error in the frequency sampling approach in (a) is zero at the desired frequency points. However, since the frequency response is equal to the samples of the impulse response of an infinite duration there will be aliasing of the impulse response, resulting in the error outside the sampling points. For the case of windowing the impulse response in (b), the aliasing in the frequency response is avoided since the impulse response is truncated. However, the truncation causes an error in the resulting frequency response. In this case the error distribution is not the same as in the case in (a). The mean squared error E_r is calculated and presented in Fig. 6.28, along with the errors in the absolute value of the frequency responses in both cases. As expected from the theory, the impulse response truncation produced lower mean squared error in the estimation.

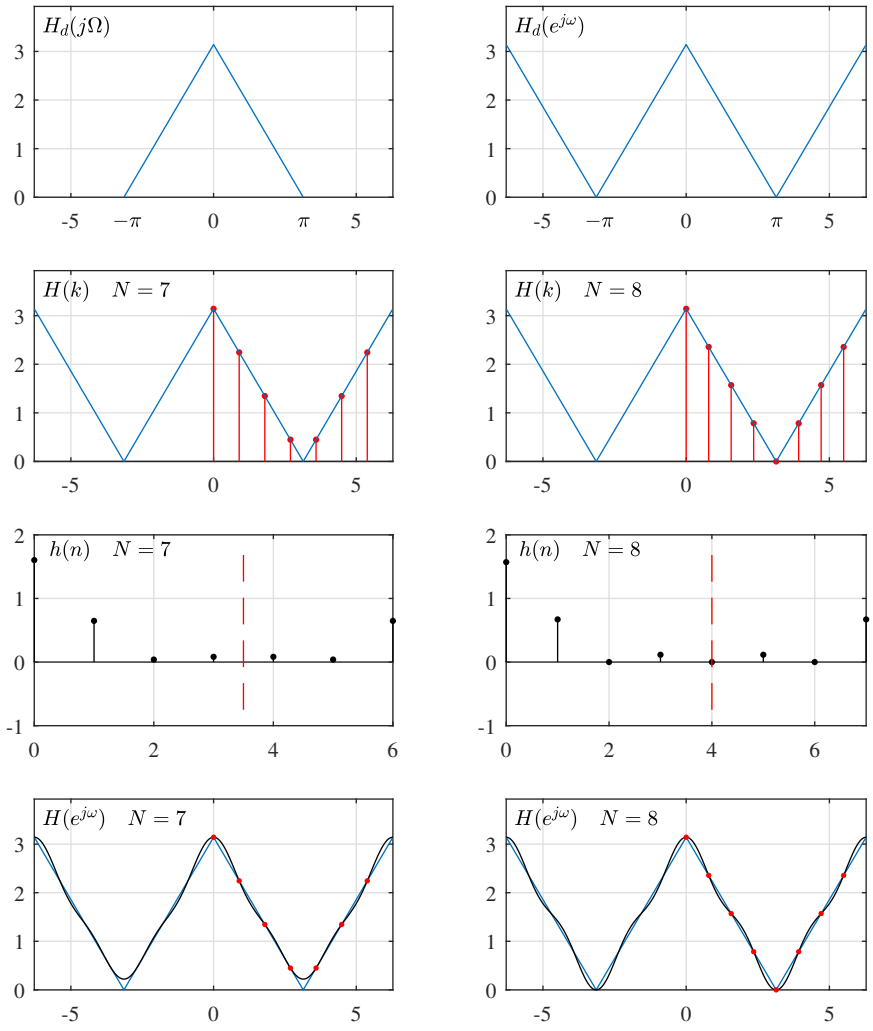


Figure 6.26 Design of a FIR filter by the frequency sampling of the desired frequency response.

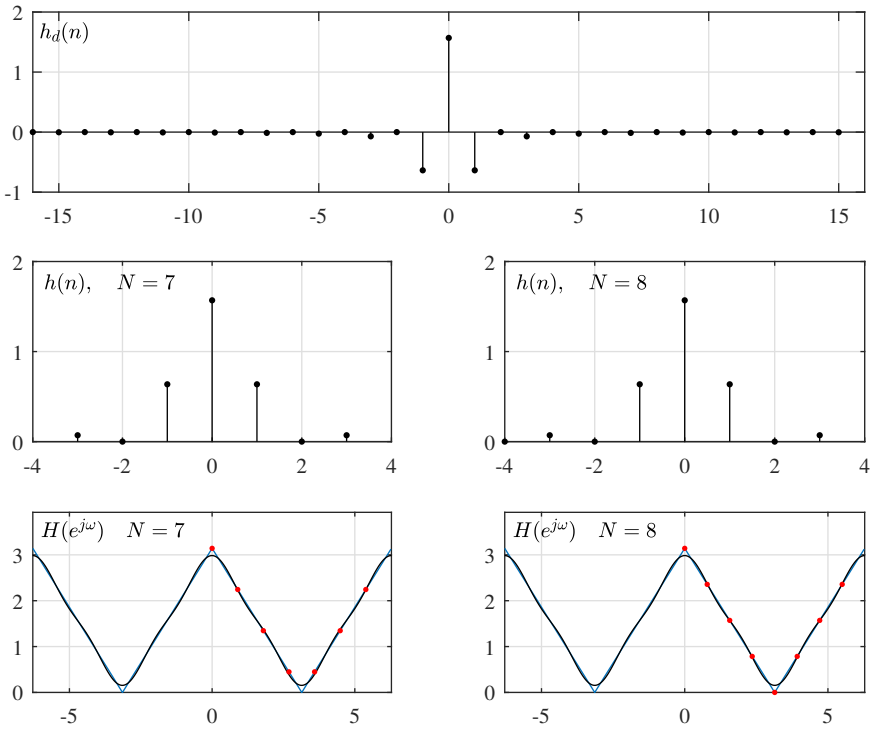


Figure 6.27 Design of a FIR filter by windowing the impulse response of an IIR filter.

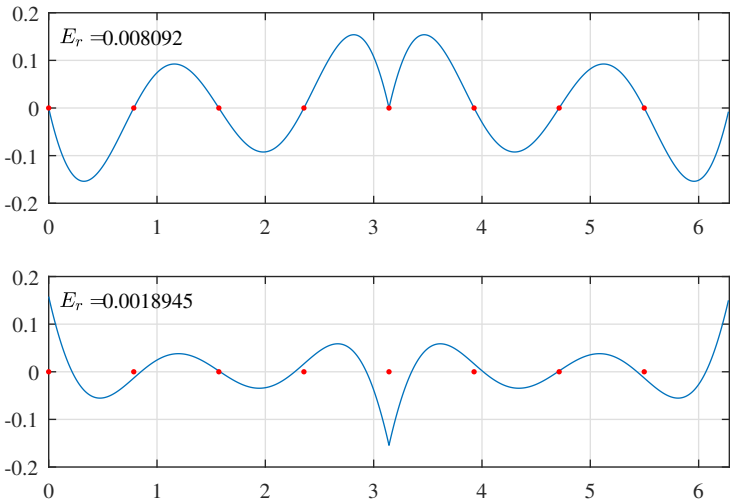


Figure 6.28 Error in the case of the frequency response sampling (top) and the IIR impulse response truncation (bottom), along with the corresponding mean square error (E_r) value.

6.3 PROBLEMS

Problem 6.1. For the system whose transfer function is

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

plot the cascade, parallel and direct realization.

Problem 6.2. Given the discrete system with

$$y(n) = x(n) + x(n-1) + x(n-2) + y(n-1) - y(n-2) - 3y(n-3).$$

Plot its direct realization I, direct realization II, parallel realization, and cascade realization.

Problem 6.3. Find the transfer function of the discrete system presented in Fig. 6.29.

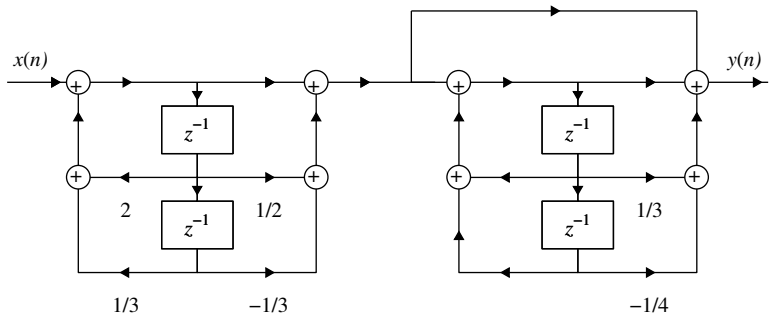


Figure 6.29 Discrete-time system

Problem 6.4. Find the transfer function of the discrete system presented in Fig. 6.30.

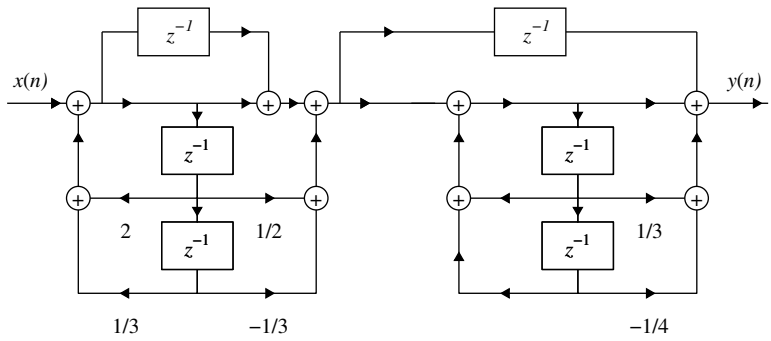


Figure 6.30 Discrete-time system

Problem 6.5. For the system

$$H(z) = \frac{1 - 0.2z^{-1} + 0.02z^{-2}}{1 - 1.7z^{-1} + 1.285z^{-2}} \frac{1 - 1.8z^{-1} + 1.45z^{-2}}{1 - 0.1z^{-1} + 0.0125z^{-2}},$$

present its cascade realization. Order the subsystems in the system so that the subsystem which is less sensitive to possible quantization comes first.

Problem 6.6. If the transfer function of the system is

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

plot its cascade and parallel realization. Write down the difference equation which describes this system.

Problem 6.7. For the system defined by the transfer function

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

plot the cascade realization.

Problem 6.8. For the system presented in Fig. 6.31 find the transfer function.

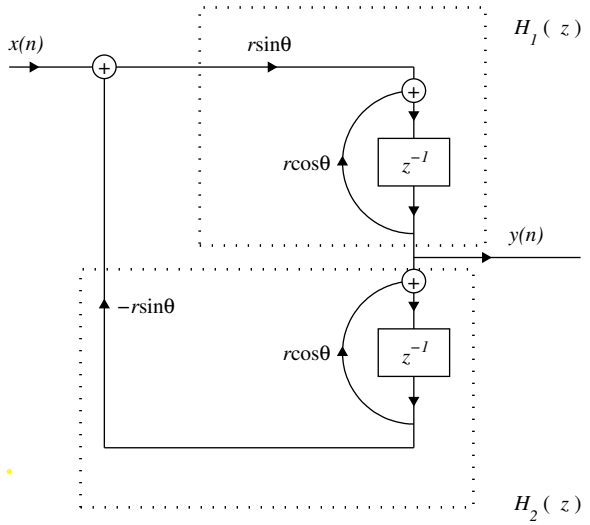


Figure 6.31 Discrete-time system

Problem 6.9. The discrete system is defined by the following two equations

$$y(n) + \frac{1}{4}y(n-1) + w(n) + \frac{1}{2}w(n-1) = \frac{2}{3}x(n)$$

$$y(n) - \frac{5}{4}y(n-1) + 2w(n) - 2w(n-1) = -\frac{5}{3}x(n),$$

where $x(n)$ is the input signal, $y(n)$ is the output signal, and $w(n)$ is a signal within the system. What is the frequency and impulse response of the system?

Problem 6.10. Show that the FIR system

$$H(z) = \frac{1 + 2z - z^2 + 4z^3 - z^4 + 2z^5 + z^6}{z^6}$$

has a linear phase function. Find its group delay.

Problem 6.11. Let $h(n)$ be an impulse response of a causal system with the Fourier transform $H(e^{j\omega})$. A real-valued output signal $y_1(n) = x(n) * h(n)$ of this system is reversed, $r(n) = y_1(-n)$, and passed through the same system, resulting in the output signal $y_2(n) = r(n) * h(n)$. The final output is reversed again $y(n) = y_2(-n)$. Find the phase of the frequency response function of the overall system.

Problem 6.12. For the system whose frequency response in the continuous-time domain is

$$H_d(j\Omega) = \begin{cases} 2 & \text{for } |\omega| < \frac{\pi}{2} \\ 1 & \text{for } \frac{\pi}{2} < |\omega| < \frac{3\pi}{4} \\ 0 & \text{elsewhere,} \end{cases}$$

with the corresponding $H_d(e^{j\omega})$ in the discrete-time domain obtained with $\Delta t = 1$, find the FIR filter impulse response with $N = 15$ and $N = 14$:

(a) Sampling the desired frequency response $H_d(e^{j\omega})$ in the frequency domain,

(b) Calculating $h_d(n) = \text{IFT}\{H_d(e^{j\omega})\}$ and taking its N the most significant values, $h(n) = h_d(n)$ for $-N/2 \leq n \leq N/2 - 1$ and $h(n) = 0$ elsewhere (rectangular window).

(c) Comment the sources of error in both cases.

6.4 EXERCISE

Exercise 6.1. Given the discrete system with

$$y(n) = x(n) - x(n-1) + x(n-2) + \frac{1}{2}y(n-1) - \frac{1}{3}y(n-2) - \frac{1}{4}y(n-3),$$

plot its direct realization I, direct realization II, parallel realization, and cascade realization.

Exercise 6.2. For the system whose transfer function is

$$H(z) = \frac{z^2 - 2}{(z-1)(z-2)}$$

plot the direct realization I, direct realization II, parallel realization, and cascade realization.

Exercise 6.3. For the system whose transfer function is

$$H(z) = \frac{3z^{-2} + 6}{z^{-3} - 2z^{-2} + 3z^{-1} - 6}$$

a) plot the direct realization I, direct realization II, cascade realization, and parallel realization.
b) Find $\sum_{n=-\infty}^{\infty} h(n)$, where $h(n)$ is the impulse response of the system.

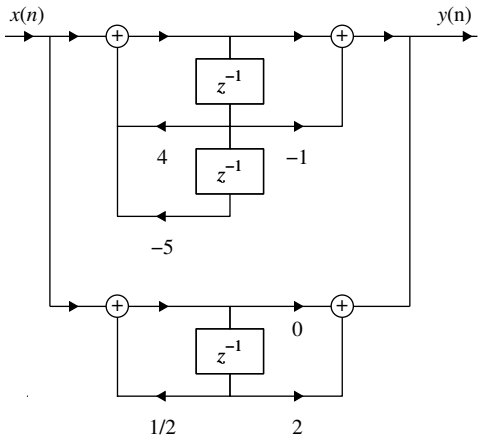


Figure 6.32 Discrete-time system.

Exercise 6.4. Find the impulse response of the discrete system presented in Fig. 6.32.

Exercise 6.5. Using the impulse invariance method with the sampling interval $\Delta t = 0.1$, transform the continuous-time system given with the transfer function

$$H(s) = \frac{1+5s}{8+2s+5s^2}$$

into a discrete-time system, and plot the direct and the cascade realization of the system. Is the obtained discrete-time system stable?

Exercise 6.6. Using the bilinear transform with the sampling interval $\Delta t = 1$, transform the system given with the transfer function

$$H(s) = \frac{2+s}{8+2s+5s^2}$$

into a discrete-time system, and plot the direct and the cascade realization of the system. Is the obtained discrete system stable?

Exercise 6.7. Using the bilinear transform, with the sampling interval $\Delta t = 0.2$ transform the continuous-time system given with the transfer function

$$H(s) = \frac{3s+6}{(s+1)(s+3)}$$

into discrete-time system, and plot its direct realization II.

Exercise 6.8. For the system whose frequency response in the continuous-time domain is

$$H_d(j\Omega) = \begin{cases} 2 - \frac{|\Omega|}{\pi/2} \text{ for } |\omega| < \frac{\pi}{2} \\ 0 \qquad\qquad\quad \text{elsewhere} \end{cases}$$

with the corresponding $H_d(e^{j\omega})$ in the discrete-time domain obtained for $\Delta t = 1$, and presented in Fig. 6.33, find the FIR filter impulse response with $N = 7$ and $N = 8$:

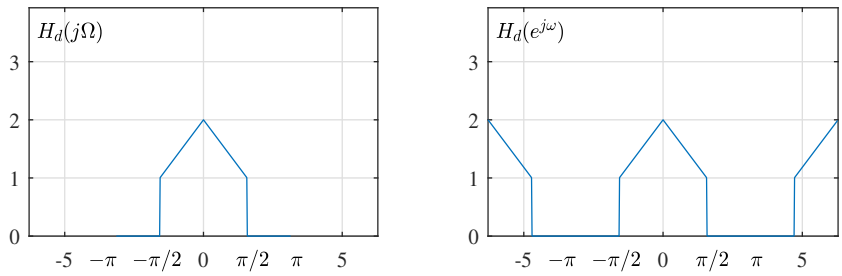


Figure 6.33 The desired system in the continuous-time domain (left) and discrete-time domain (right).

(a) Sampling the desired frequency response $H_d(e^{j\omega})$ in the frequency domain,

(b) Calculating $h_d(n) = \text{IFT}\{H_d(e^{j\omega})\}$ and taking its N the most significant values, $h(n) = h_d(n)$ for $-N/2 \leq n \leq N/2-1$ and $h(n) = 0$ elsewhere.

(c) Comment the sources of the error in both cases.

6.5 SOLUTIONS

Solution 6.1. In order to plot the direct form of realization, the transfer function should be written in a form suitable for this type of realization,

$$H(z) = \frac{16(z+1)z^2}{(4z^2-2z+1)(4z+3)} = \frac{1+z^{-1}}{(1-\frac{1}{2}z^{-1}+\frac{1}{4}z^{-2})(1+\frac{3}{4}z^{-1})}$$
$$= \frac{1+z^{-1}}{1+\frac{1}{4}z^{-1}-\frac{1}{8}z^{-2}+\frac{3}{16}z^{-3}}. \tag{6.11}$$

According to the previous relation, direct realizations I and II follow. They are presented in Fig. 6.34 and Fig. 6.35, respectively.

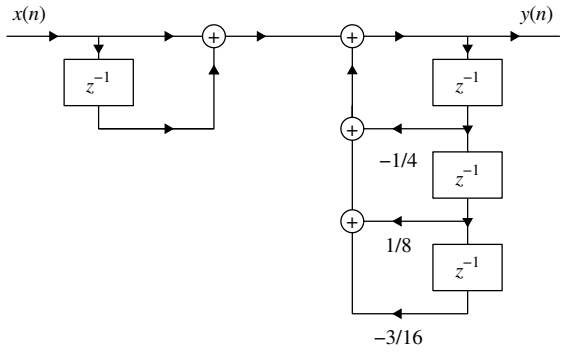


Figure 6.34 Direct realization I of the discrete-time system in (6.11).

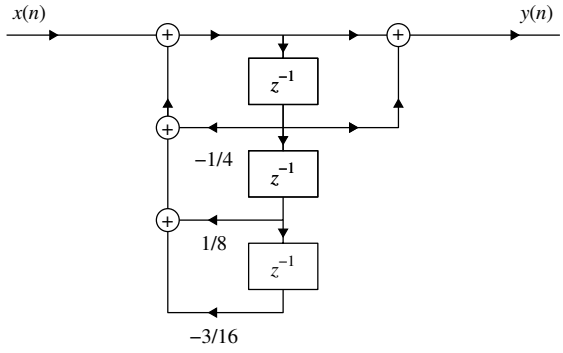


Figure 6.35 Direct realization II of the discrete-time system in (6.11).

For the cascade realization, the transfer function is written in the form

$$H(z) = \frac{1+z^{-1}}{(1-\frac{1}{2}z^{-1}+\frac{1}{4}z^{-2})(1+\frac{3}{4}z^{-1})} = \frac{1+z^{-1}}{1-\frac{1}{2}z^{-1}+\frac{1}{4}z^{-2}} \frac{1}{1+\frac{3}{4}z^{-1}} = H_1(z)H_2(z).$$

The cascade realization, implemented as a product of two blocks, has the form shown in Fig. 6.36.

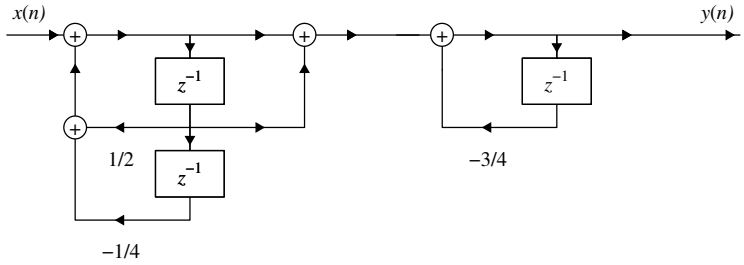


Figure 6.36 Cascade realization of the discrete-time system in (6.11).

In order to plot a parallel realization, the transfer function should be written in a form of partial fractions expansion, which is suitable for this type of realization,

$$H(z) = \frac{1+z^{-1}}{(1-\frac{1}{2}z^{-1}+\frac{1}{4}z^{-2})(1+\frac{3}{4}z^{-1})} = \frac{Az^{-1}+B}{1-\frac{1}{2}z^{-1}+\frac{1}{4}z^{-2}} + \frac{C}{1+\frac{3}{4}z^{-1}}.$$

Calculating the coefficients $A = 1/19$, $B = 22/19$ and $C = -3/19$, we get

$$H(z) = \frac{\frac{22}{19}+\frac{1}{19}z^{-1}}{1-\frac{1}{2}z^{-1}+\frac{1}{4}z^{-2}} + \frac{-\frac{3}{19}}{1+\frac{3}{4}z^{-1}}.$$

This equations is used to plot the parallel realization, Fig. 6.37.

Solution 6.2. Using the z-transform properties, the given difference equation can be written as

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

According to the definition of the transfer function, we get

$$H(z) = \frac{Y(z)}{X(z)} = \frac{1+z^{-1}+z^{-2}}{1-z^{-1}+z^{-2}+3z^{-3}}. \tag{6.12}$$

Direct realizations I and II, presented in Fig. 6.38 and Fig. 6.39, respectively, follow from the previous equation.

For the cascade realization, the transfer function should be written in the form of a product of two blocks

$$H(z) = \frac{1+z^{-1}+z^{-2}}{1-2z^{-1}+3z^{-2}} \frac{1}{1+z^{-1}} = H_1(z)H_2(z).$$

This form is now suitable for the cascade realization given in Fig. 6.40.

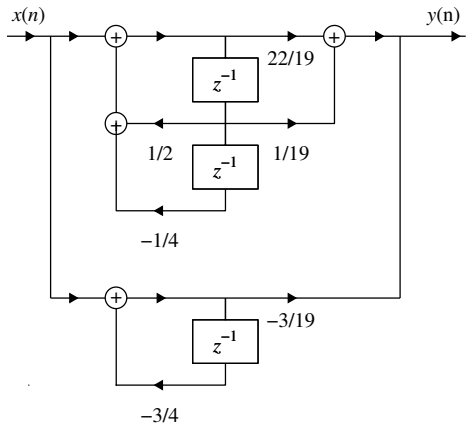


Figure 6.37 Parallel realization of the discrete-time system in (6.11).

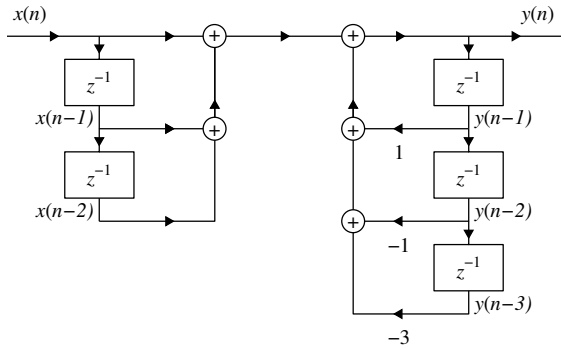


Figure 6.38 Direct realization I of the discrete-time system in (6.12).

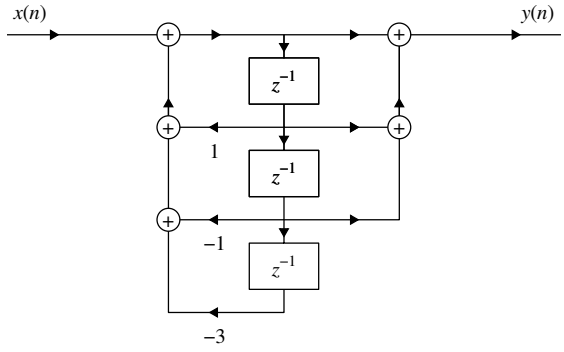


Figure 6.39 Direct realization II of the discrete-time system in (6.12).

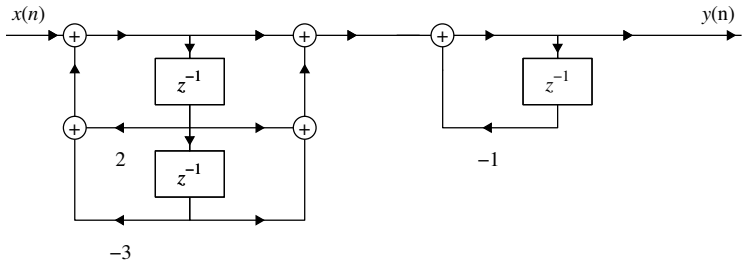


Figure 6.40 Cascade realization of the discrete-time system in (6.12).

For the parallel realization, we will write the transfer function in the form of partial fractions with real-valued coefficients

$$H(z) = \frac{\frac{1}{6}}{1+z^{-1}} + \frac{\frac{1}{2}z^{-1}+\frac{5}{6}}{1-2z^{-1}+3z^{-2}}.$$

Its realization is now straightforward.

Solution 6.3. The system can be recognized as a cascade of two subsystems and its transfer function can be written as a product of the transfer functions of these two blocks

$$H(z) = H_1(z)H_2(z),$$

where $H_1(z)$ denotes the first block and $H_2(z)$ denotes the second block. The first subsystem with the transfer function $H_1(z)$ can be considered as a direct realization II, with the input to output relation

$$y_1(n) = 2y_1(n-1) + \frac{1}{3}y_1(n-2) + x(n) + \frac{1}{2}x(n-1) - \frac{1}{3}x(n-2),$$

as shown in Fig. 6.41. Using the z-transform properties, its transfer function is

$$H_1(z) = \frac{Y_1(z)}{X(z)} = \frac{1+\frac{1}{2}z^{-1}-\frac{1}{3}z^{-2}}{1-2z^{-1}-\frac{1}{3}z^{-2}}.$$

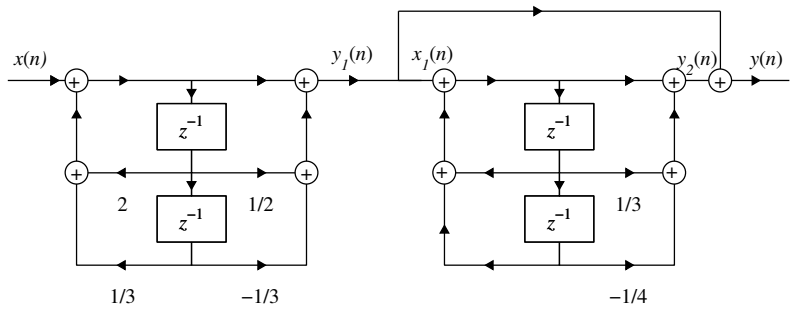


Figure 6.41 A discrete-time system.

Now consider the second block whose transfer function is $H_2(z)$. This block can be considered as a parallel realization of two blocks, $H_2(z) = H_{21}(z) + H_{22}(z)$ where $H_{21}(z) = 1$.

The second transfer function is the transfer function that corresponds to a direct realization II, of a subsystem described by

$$y_2(n) = y_2(n-1) + y_2(n-2) + x_1(n) + \frac{1}{3}x_1(n-1) - \frac{1}{4}x_1(n-2).$$

Thus, the transfer function of this subsystem is

$$H_{22}(z) = \frac{Y_2(z)}{X_1(z)} = \frac{1 + \frac{1}{3}z^{-1} - \frac{1}{4}z^{-2}}{1 - z^{-1} - z^{-2}}.$$

The transfer function of the second block is now

$$H_2(z) = H_{21}(z) + H_{22}(z) = 1 + \frac{1 + \frac{1}{3}z^{-1} - \frac{1}{4}z^{-2}}{1 - z^{-1} - z^{-2}}.$$

Finally, the transfer function of the whole system is

$$H(z) = H_1(z)H_2(z) = \frac{1 + \frac{1}{2}z^{-1} - \frac{1}{3}z^{-2}}{1 - 2z^{-1} - \frac{1}{3}z^{-2}} \left(1 + \frac{1 + \frac{1}{3}z^{-1} - \frac{1}{4}z^{-2}}{1 - z^{-1} - z^{-2}}\right).$$

Solution 6.4. This realization can be considered as a cascade realization of two blocks $H_1(z)$ and $H_2(z)$,

$$H(z) = H_1(z)H_2(z).$$

The first block is a direct realization II, whose transfer function is

$$H_1(z) = \frac{1 + (\frac{1}{2} + 1)z^{-1} - \frac{1}{3}z^{-2}}{1 - 2z^{-1} - \frac{1}{3}z^{-2}}.$$

Previous relation holds since the upper delay block (above the obvious direct realization II block) has the same input and output as the first delay block below it.

The block with transfer function $H_2(z)$ can be considered as a parallel realization of two blocks, similarly as in previous example, with, $H_{21}(z)$ and $H_{22}(z)$, defined by

$$H_{21}(z) = \frac{1 + \frac{1}{3}z^{-1} - \frac{1}{4}z^{-2}}{1 - z^{-1} - z^{-2}},$$

and

$$H_{22}(z) = z^{-1}.$$

Hence, the transfer function of the second block is

$$H_2(z) = H_{21}(z) + H_{22}(z) = \frac{1 + \frac{1}{3}z^{-1} - \frac{1}{4}z^{-2}}{1 - z^{-1} - z^{-2}} + z^{-1}.$$

Now, the resulting transfer function can be written in the form

$$H(z) = H_1(z)H_2(z) == \frac{1 + (\frac{1}{2} + 1)z^{-1} - \frac{1}{3}z^{-2}}{1 - 2z^{-1} - \frac{1}{3}z^{-2}} \left(\frac{1 + \frac{1}{3}z^{-1} - \frac{1}{4}z^{-2}}{1 - z^{-1} - z^{-2}} + z^{-1}\right).$$

Solution 6.5. The transfer function can be written as

$$H(z) = H_1(z)H_2(z).$$

It can be expressed, using the roots of the numerator and denominator polynomials, as

$$H(z) = \frac{\left(1-(0.1+j0.1)z^{-1}\right)\left(1-(0.1-j0.1)z^{-1}\right)}{\left(1-(0.85-j0.75)z^{-1}\right)\left(1-(0.85+j0.75)z^{-1}\right)} \times \frac{\left(1-(0.9+j0.8)z^{-1}\right)\left(1-(0.9-j0.8)z^{-1}\right)}{\left(1-(0.05-j0.1)z^{-1}\right)\left(1-(0.05+j0.1)z^{-1}\right)}.$$

The subsystems should be positioned as

$$H_1(z) = \frac{1-1.8z^{-1}+1.45z^{-2}}{1-1.7z^{-1}+1.285z^{-2}} \qquad H_2(z) = \frac{1-0.2z^{-1}+0.02z^{-2}}{1-0.1z^{-1}+0.0125z^{-2}},$$

since the zero-pole pairs with small values of imaginary parts should come later. They are more sensitive to the quantization of coefficients and they will more probably cause this kind of error. Larger imaginary parts of roots are less sensitive to these effects, as discussed in Section 6.1.3. The cascade realization is given in Fig. 6.42.

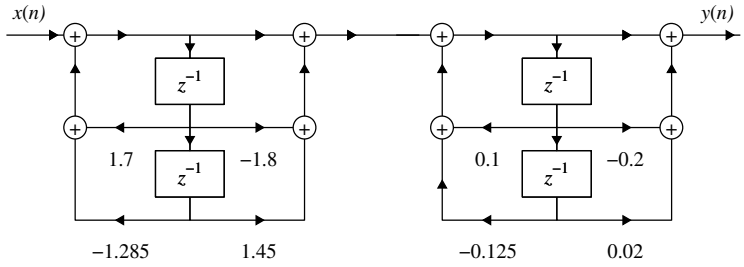


Figure 6.42 Cascade realization of the system with blocks ordered in such a way that the whole system is less sensitive to possible quantization error

Solution 6.6. For a cascade realization, the transfer function is expressed in the form

$$H(z) = \frac{1}{1-\frac{1}{2}z^{-1}+\frac{1}{4}z^{-2}} \frac{1+z^{-1}}{1+\frac{3}{4}z^{-1}}. \tag{6.13}$$

Its realization is presented in Fig. 6.43

For a parallel realization, the transfer function can be expanded into partial fractions form

$$H(z) = \frac{\frac{22}{19}+\frac{1}{19}z^{-1}}{1-\frac{1}{2}z^{-1}+\frac{1}{4}z^{-2}} + \frac{-\frac{3}{19}z^{-1}}{1+\frac{3}{4}z^{-1}}. \tag{6.14}$$

This realization is shown in Fig. 6.44.

The transfer function can be written in the form

$$H(z) = \frac{1+z^{-1}}{1+\frac{1}{4}z^{-1}-\frac{1}{8}z^{-2}+\frac{3}{16}z^{-3}}.$$

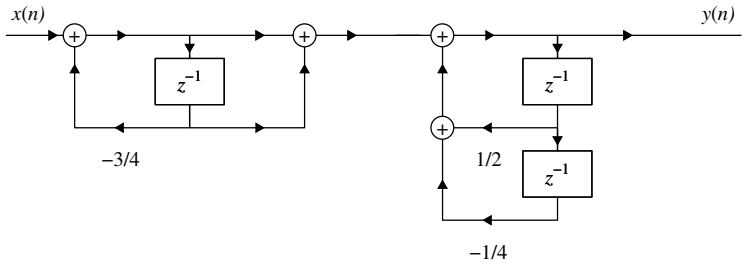


Figure 6.43 A cascade realization of the system in (6.13).

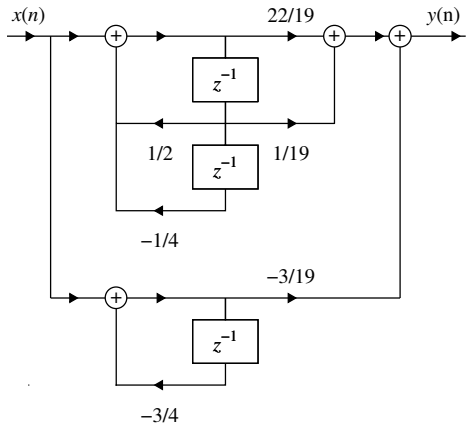


Figure 6.44 Parallel realization of the discrete-time system in (6.14).

The difference equation describing this system is

$$y(n) = x(n) + x(n-1) - \frac{1}{4}y(n-1) + \frac{1}{8}y(n-2) - \frac{3}{16}y(n-3).$$

Solution 6.7. The transfer function form which corresponds to the cascade realization of the system is

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

In order the use the smallest number of the delay circuits, it can be expressed in the form

$$H(z) = H_1(z)H_2(z) = \frac{1}{(1+z^{-1})}\,\frac{(1+z^{-2})}{(1+z^{-1}+z^{-2})}. \tag{6.15}$$

This form corresponds to the cascade realization presented in Fig. 6.45.

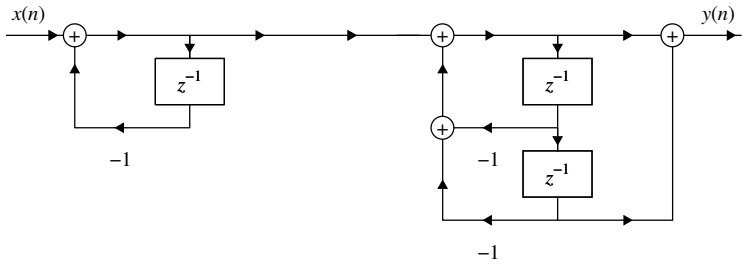


Figure 6.45 Cascade realization of the discrete-time system in (6.15).

Solution 6.9. The z-transforms of these difference equations are

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

$$Y(z)(1-\frac{5}{4}z^{-1})+2W(z)(1-z^{-1})=-\frac{5}{3}X(z).$$

By eliminating $W(z)$ we get

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

The transfer function is

$$H(z)=\frac{Y(z)}{X(z)}=\frac{3-\frac{1}{2}z^{-1}}{1-\frac{3}{4}z^{-1}+\frac{1}{8}z^{-2}},$$

with the difference equation describing this system

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

The frequency response is

$$H(e^{j\omega})=\frac{3-\frac{1}{2}e^{-j\omega}}{1-\frac{3}{4}e^{-j\omega}+\frac{1}{8}e^{-j2\omega}}.$$

Based on

$$H(z)=\frac{Y(z)}{X(z)}=\frac{3-\frac{1}{2}z^{-1}}{1-\frac{3}{4}z^{-1}+\frac{1}{8}z^{-2}}=\frac{4}{1-\frac{1}{2}z^{-1}}-\frac{1}{1-\frac{1}{4}z^{-1}},$$

the impulse response is

$$h(n)=[4(1/2)^n-(1/4)^n]u(n).$$

Solution 6.8. The transfer function of the subsystem denoted by $H_1(z)$ follows from

$$y(n)=r\sin\theta x_1(n-1)+r\cos\theta y(n-1)$$

where $x_1(n)$ is the input signal to this subsystem, whose transfer function is

$$H_1(z) = \frac{Y(z)}{X_1(z)} = \frac{z^{-1} r \sin\theta}{1 - r\cos\theta z^{-1}}.$$

The transfer function of the other subsystem is

$$H_2(z) = -\frac{z^{-1} r \sin\theta}{1 - r\cos\theta z^{-1}}.$$

For the feedback holds

$$H_1(z)(X(z) + Y(z)H_2(z)) = Y(z).$$

This relation produces

$$H(z) = \frac{Y(z)}{X(z)} = \frac{H_1(z)}{1 - H_1(z)H_2(z)} = \frac{z^{-1} r \sin\theta (1 - r\cos\theta z^{-1})}{1 - 2r\cos\theta z^{-1} + r^2 z^{-2}}.$$

Solution 6.10. The system impulse response is

$$h(n) = \delta(n) + 2\delta(n-1) - \delta(n-2) + 4\delta(n-3) - \delta(n-4) + 2\delta(n-5) + \delta(n-6).$$

This impulse response satisfies the property

$$h(n) = h(N-1-n), \quad 0 \le n \le N-1,$$

with $N = 7$, which implies the phase function linearity. Thus, the group delay q is

$$q = \frac{N-1}{2} = 3.$$

Solution 6.11. Let $h(n)$ be an impulse response of a causal system with the Fourier transform $H(e^{j\omega})$. A real-valued output signal $y_1(n) = x(n) * h(n)$ of this system is reversed, $r(n) = y_1(-n)$, and passed through the same system, resulting in the output signal $y_2(n) = r(n) * h(n)$. The final output is reversed again $y(n) = y_2(-n)$. Find the phase of the frequency response function of the overall system. The frequency domain form of the system $y_1(n) = x(n) * h(n)$ is

$$Y_1(e^{j\omega}) = H(e^{j\omega})X(e^{j\omega}).$$

For the operation $r(n) = y_1(-n)$ in the time domain, the frequency domain form is

$$R(e^{j\omega}) = Y_1^*(e^{j\omega}) = H^*(e^{j\omega})X^*(e^{j\omega})$$

When this signal passes through the same system $y_2(n) = r(n) * h(n)$ we have

$$Y_2(e^{j\omega}) = R(e^{j\omega})H(e^{j\omega}) = H^*(e^{j\omega})H(e^{j\omega})X^*(e^{j\omega})$$

Finally, the signal is reversed, $y(n) = y_2(-n)$, producing

$$Y(e^{j\omega}) = Y_2^*(e^{j\omega}) = H(e^{j\omega})H^*(e^{j\omega})X(e^{j\omega}).$$

So we get

$$Y(e^{j\omega}) = |H(e^{j\omega})|^2 X(e^{j\omega}).$$

Obviously, the phase function of the overall system, $|H(e^{j\omega})|^2$, is equal to zero for all ω.

Solution 6.12. (a) Values of the FIR filter, obtained by sampling the frequency response in the frequency domain are

$$H(k) = H_d(e^{j\omega})\Big|_{\omega=2\pi nk/N}.$$

This sampling is illustrated in the second row of Fig. 6.46 for $N = 15$ and $N = 14$.

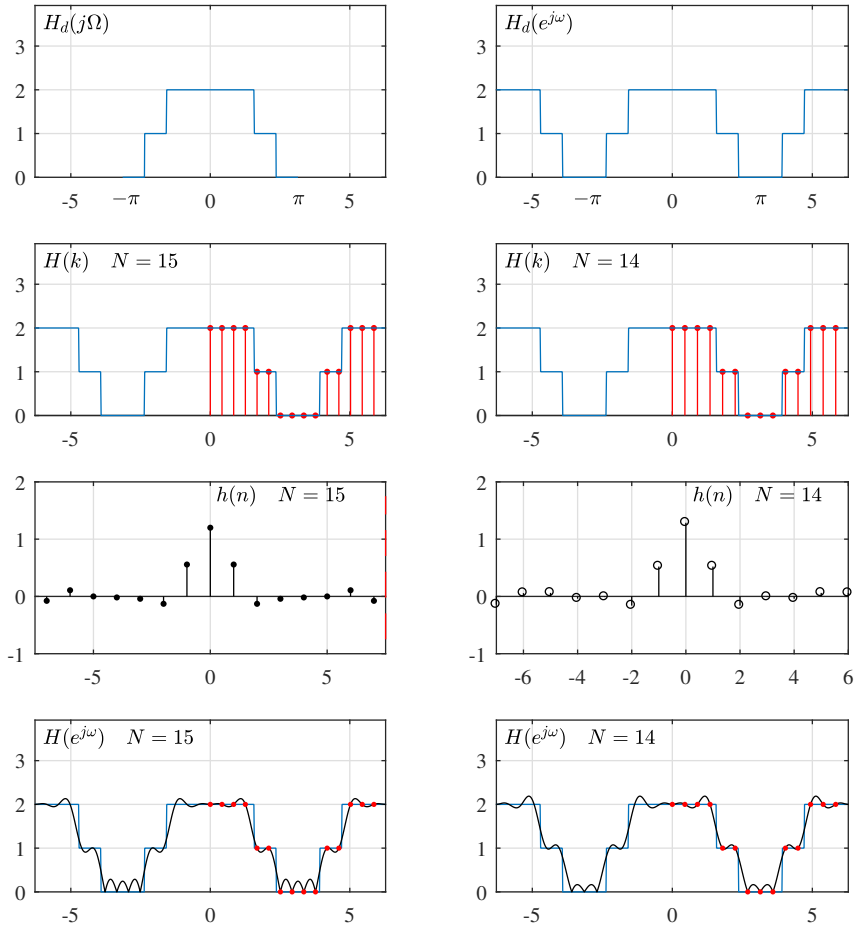


Figure 6.46 Design of the FIR filter using the frequency sampling of the desired frequency response.

The impulse response of the FIR filter is calculated as

$$h(n) = \text{IDFT}\{H(k)\} = \frac{1}{N}\sum_{k=0}^{N-1} H(k)e^{j2\pi nk/N}.$$

It is shown in Fig. 6.46 (third row). The frequency response of the FIR filter is

$$H(e^{j\omega}) = \text{FT}\{h(n)\}.$$

Its values are equal to the desired frequency response at the sampling points

$$H(e^{j\omega})\Big|_{\omega=2\pi k/N} = H_d(e^{j\omega})\Big|_{\omega=2\pi k/N}.$$

(b) The impulse response of the desired system is

$$h_d(n) = \text{IFT}\{H_d(e^{j\omega})\} = \frac{\sin(n\pi/2)}{\pi n} + \frac{\sin(3n\pi/4)}{\pi n}.$$

Using the first $N = 15$ samples in the discrete-time domain we get

$$h(n) = \begin{cases} h_d(n), & \text{for } -7 \le n \le 7 \\ 0, & \text{elsewhere} \end{cases}$$

or for $N = 16$

$$h(n) = \begin{cases} h_d(n), & \text{for } -8 \le n \le 7 \\ 0. & \text{elsewhere.} \end{cases}$$

The frequency response of this FIR filter is

$$H(e^{j\omega}) = \text{DFT}\{h(n)\}.$$

It is shown in Fig. 6.47.

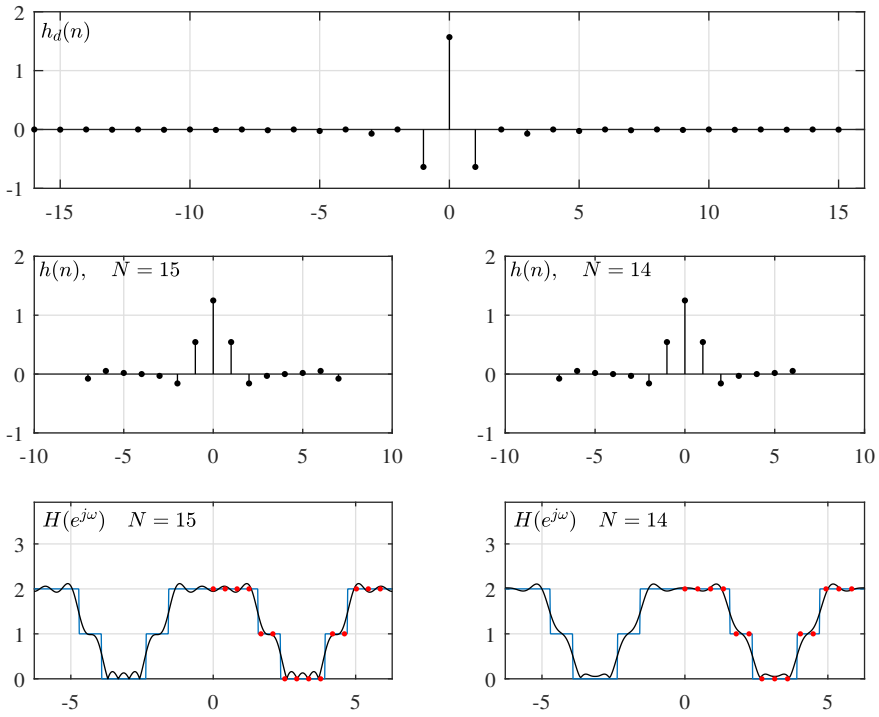


Figure 6.47 The FIR filter design using N the most significant values of the impulse response (window approach).

(c) The error value, as a function of the frequency ω, along with the mean squared absolute error E_r is shown in Fig. 6.48.

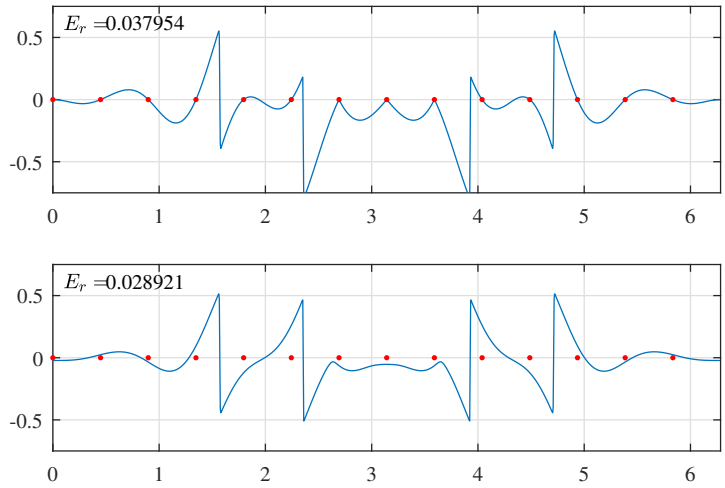


Figure 6.48 Error in the case of the frequency response sampling (top) and the IIR impulse response truncation (bottom), along with the corresponding mean square error (E_r) value.

Part III

Random Discrete-Time Signals and Systems

Chapter 7

Discrete-Time Random Signals

RANDOM signal values cannot be defined by simple deterministic 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 second-order statistics will be considered.

7.1 BASIC STATISTICAL DEFINITIONS

Statistics is a science or practice dealing with the collection, analysis, interpretation, and presentation of numerical data, inferring parameters from the whole set of data or their representative sample. A *statistic* is a single numerical fact obtained from the analysis of the considered set of data and used for the whole data set description.

7.1.1 Mean Value – Sample Average

The first-order statistics is the starting point in describing random signals. The mean value, or the data *sample average*, of a random signal is one of the parameters of this statistics. If we have a set of signal samples,

$$\mathbb{X} = \{x(n)|\ \ n = 1,2,\ldots,N\}, \tag{7.1}$$

the mean value of this set of signal samples is calculated as

$$\hat{\mu}_x = \text{mean}\{x(n)|\ \ n = 1,2,\ldots,N\} = \frac{1}{N}(x(1) + x(2) + \cdots + x(N)). \tag{7.2}$$

For notation simplicity, we will also use $\hat{\mu}_x = \text{mean}\{x(n)\}$, meaning the mean of the dataset $\{x(n)\}$ for all indices n where the signal is available.

To distinguish the calculated (statistically estimated) value $\hat{\mu}$ of a signal parameter from the true one μ (if all possible signal realizations were available) we will use the hat ($\hat{}$) symbol.

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.

Table 7.1
A realization of random signal

54	62	58	51	70	43	99	52	57	76
56	53	38	61	28	69	87	41	72	80
23	26	66	47	69	71	69	81	68	79
31	55	52	23	60	34	83	39	66	59
37	12	54	42	67	95	89	67	42	63
35	55	54	55	49	77	18	64	73	70
67	56	42	66	50	47	49	25	50	57
61	84	48	67	71	74	35	59	60	42
40	77	52	63	57	42	44	64	36	71
66	39	50	31	11	75	45	62	60	55

★The realization of signal $x(n)$ defined in Table 7.1 is presented in Fig. 7.1.

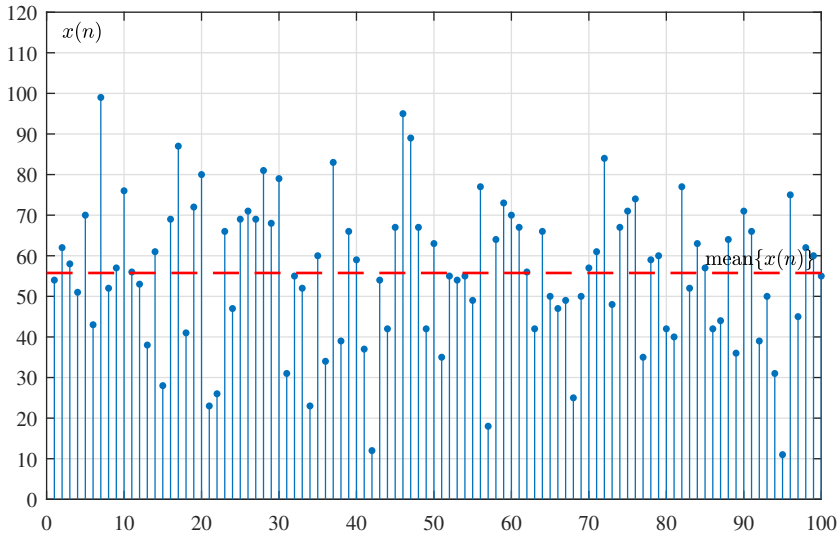


Figure 7.1 A realization of the random signal $x(n)$.

The mean value of all signal samples is

$$\hat{\mu}_x = \frac{1}{100}\sum_{n=1}^{100} x(n) = 55.76.$$

From Table 7.1 or its visualized presentation in Fig. 7.1, we can conclude that, for example, there is no any 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 the result is shown in Fig. 7.2. This kind of random signal presentation is called a *histogram* of $x(n)$, with the defined intervals.

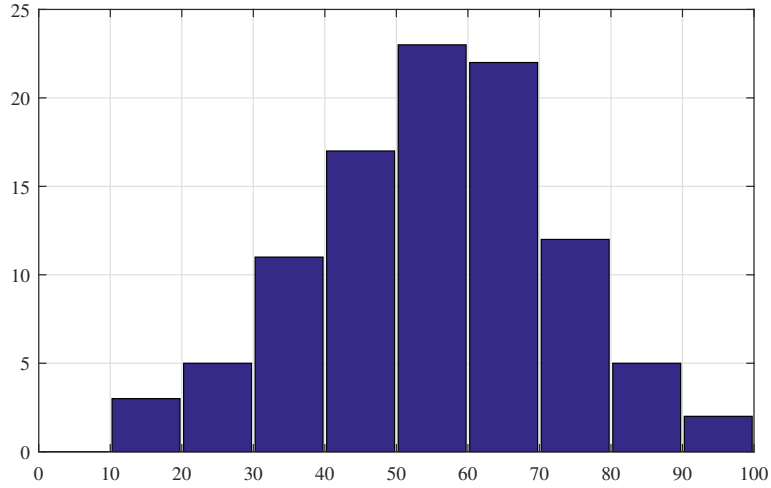


Figure 7.2 Histogram of the random signal $x(n)$ from Fig. 7.1, with 10 intervals defined by $[10i+1, 10i+10]$, $i = 0,1,2,\ldots,9$.

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 $\text{int}\{\cdot\}$ denotes the nearest integer. This means that $y(n) = 1$ for $1 \leq x(n) \leq 10$, $y(n) = 2$ for $11 \leq x(n) \leq 20$, $\ldots$, $y(n) = i$ for $10(i-1)+1 \leq x(n) \leq 10i$, up to $i = 10$. What is the set of possible values of $y(n)$? Find and graphically present the number of occurrences of every possible value of $y(n)$ 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. It takes the values from the set $\{2,3,4,5,6,7,8,9,10\}$.

For the signal $y(n)$, instead of the histogram, we can plot a diagram of the number of occurrences of every value that $y(n)$ can take, as in Fig. 7.4. The mean value of $y(n)$ is

$$\hat{\mu}_y = \frac{1}{100}\sum_{n=1}^{100} y(n) = 6.13.$$

The mean value can also be written, by grouping the same values of $y(n)$, as

$$\hat{\mu}_y = \frac{1}{100}(1\cdot n_1 + 2\cdot n_2 + 3\cdot n_3 + \cdots + 10\cdot n_{10}) =$$
$$= 1\cdot\frac{n_1}{N} + 2\cdot\frac{n_2}{N} + 3\cdot\frac{n_3}{N} + \cdots + 10\cdot\frac{n_{10}}{N},$$

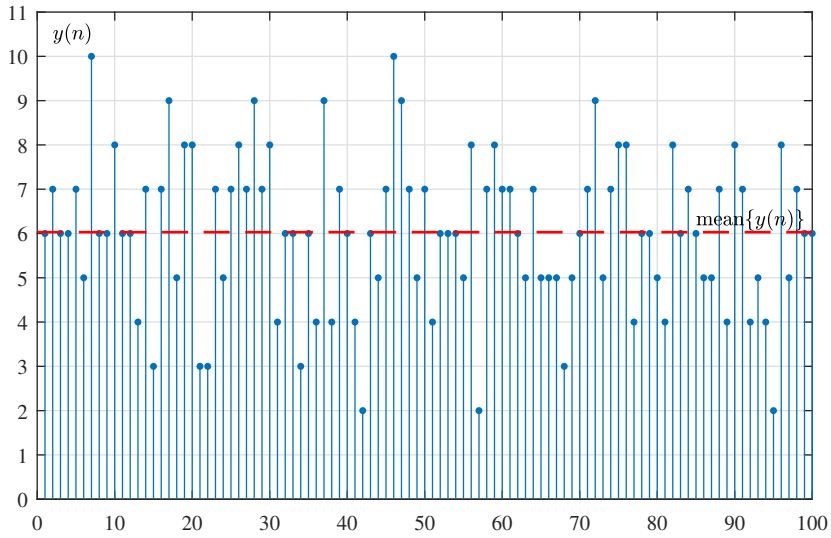


Figure 7.3 Random signal $y(n)$.

where $N = 100$ is the total number of the available 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 every outcome value i, then

$$P_y(i) = \frac{n_i}{N}$$

can be considered as an estimate of the probability that the value i appears. In that sense

$$\hat{\mu}_y = 1 \cdot P_y(1) + 2 \cdot P_y(2) + 3 \cdot P_y(3) + \cdots + 10 \cdot P_y(10) = \sum_{i=1}^{10} y(i) P_y(i)$$

with

$$\sum_{i=1}^{10} P_y(i) = \sum_{i=1}^{10} \frac{n_i}{N} = 1.$$

Values of the probability estimates $P_y(i)$ are shown in Fig. 7.4.

In general, the mean value for every 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 the argument $n = 1,2,3,\ldots,N$ is the cardinal number of the day within a year and $i = 1,2,\ldots,M$ is the index of realization (year index). The mean value is then calculated as

$$\hat{\mu}_x(n) = \frac{1}{M}(x_1(n) + x_2(n) + \cdots + x_M(n)) = \frac{1}{M}\sum_{i=1}^{M} x_i(n), \tag{7.3}$$

for every n. In this case, we have a set (a signal) of mean values $\{\hat{\mu}_x(n)\}$, for $n = 1,2,\ldots,365$.

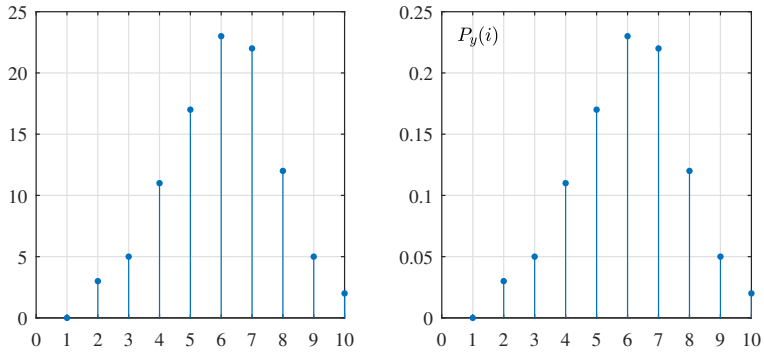


Figure 7.4 Number of appearances of every possible value of $y(n)$ (left) and the estimates of the probabilities that the random signal $y(n)$ takes a value $i = 1,2,\ldots,10$ (right).

Example 7.3. Consider the signal $x(n)$ whose realizations are given in Table 7.2. The values of $x(n)$ are equal to the monthly average of the maximum daily temperatures in a city measured from year 2001 to 2015. Find the mean of this temperature for each month over the considered period of years. What is the mean value of the temperature over all months and years? What is the mean temperature for every year?

★The signal for years 2001 to 2007 is given in Fig. 7.5. The mean temperature for the nth month, over the considered years, is

$$\hat{\mu}_x(n) = \frac{1}{15}\sum_{i=1}^{15} x_{20i}(n),$$

where the notation $20i$ is symbolic in the sense that 2001, 2002, ...2015 holds for $i = 01,02,\ldots,15$. The mean value signal $\hat{\mu}_x(n)$ is shown in the last panel of Fig. 7.5. The mean value over all months and years is

$$\hat{\mu}_x = \frac{1}{15\cdot 12}\sum_{n=1}^{12}\sum_{i=1}^{15} x_{20i}(n) = 19.84.$$

The mean value for each of the considered years is

$$\hat{\mu}_x(20i) = \frac{1}{12}\sum_{n=1}^{12} x_{20i}(n).$$

The mean value calculated as the sample average is commonly used due it calculation simplicity. Later, it will be shown that the sample average is optimal in the estimation of the true mean value of a signal sample when its realizations are corrupted by a quite common disturbance called Gaussian noise (it is interesting not notice that Gauss introduced his famous distribution as the best framework for the sample average estimator, see Section 7.4.5).

Table 7.2

Average of maximum temperature values within months over 15 years, 2001-2015.

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
10	4	18	17	22	29	30	28	27	17	17	5
6	7	11	23	22	32	35	33	22	26	22	8
10	11	10	16	21	26	32	31	23	19	17	4
3	11	13	19	22	26	34	29	26	22	12	9
7	10	13	21	27	29	30	34	24	20	16	11
7	11	17	17	27	25	37	34	33	22	14	14
7	12	13	19	23	32	34	38	21	21	12	10
12	5	9	20	21	37	34	34	27	22	20	7
7	12	13	23	27	33	29	31	25	21	6	11
8	12	10	17	27	33	38	32	23	20	15	9
8	10	13	24	23	33	33	31	27	21	16	8
4	6	15	18	25	26	27	33	23	23	13	11
3	6	16	17	27	28	30	32	29	24	12	10
11	12	14	18	22	29	34	34	23	21	20	11
6	13	8	22	22	29	30	34	23	18	15	8

The mean value, calculated as the sample average using (7.2) or (7.3), is the result of the following minimization problem. Given a set of the random sample $x(n)$ realizations, $\{x_i(n)\}$, where $i = 1,2,\ldots,M$ is the index of a realization (in (7.2), $x_i(n) = x(i)$). The aim is to estimate the true mean signal value $\mu(n)$ by $\hat{\mu}(n)$, such that its squared distance (deviation) from the available realizations $x_i(n)$, $i = 1,2,\ldots,M$, is minimum, that is

$$\begin{aligned}\hat{\mu}_x(n) &= \min_{\alpha}\Big((x_1(n)-\alpha)^2+(x_2(n)-\alpha)^2+\cdots+(x_M(n)-\alpha)^2\Big)\\ &= \min_{\alpha}||\mathbf{x}-\alpha||_2^2 = \min_{\alpha} f(\alpha),\end{aligned}$$

where $\mathbf{x} = [x_1(n), x_2(n),\ldots,x_M(n)]^T$ and $f(\alpha) = ||\mathbf{x}-\alpha||_2$ is the two-norm of the vector $\mathbf{x}-\alpha$. The result of this minimization is obtained from

$$\frac{d}{d\alpha}\Big((x_1(n)-\alpha)^2+(x_2(n)-\alpha)^2+\cdots+(x_M(n)-\alpha)^2\Big) = 0 \tag{7.4}$$

in the form given by (7.3) or (7.2).

The sample average estimation is very sensitive to possible wrongly recorded realizations of a sample $x(n)$ or to the realization with a very high disturbance due to some exceptional circumstances. These signal realizations, which significantly differ from the true value of the signal sample, are called *outliers*, in contrast to the realizations with relatively small errors called *inliers*.

The sample average calculation will produce a completely wrong (unbounded) result if at least one outlier happens in the considered set of realizations $\{x_i(n)\}$. The smallest possible fraction of samples needed to be replaced by outliers, in order to make an estimator unbounded, is called the *breakdown point* of the estimator. For the sample average (7.3) or (7.2), the breakdown point is the

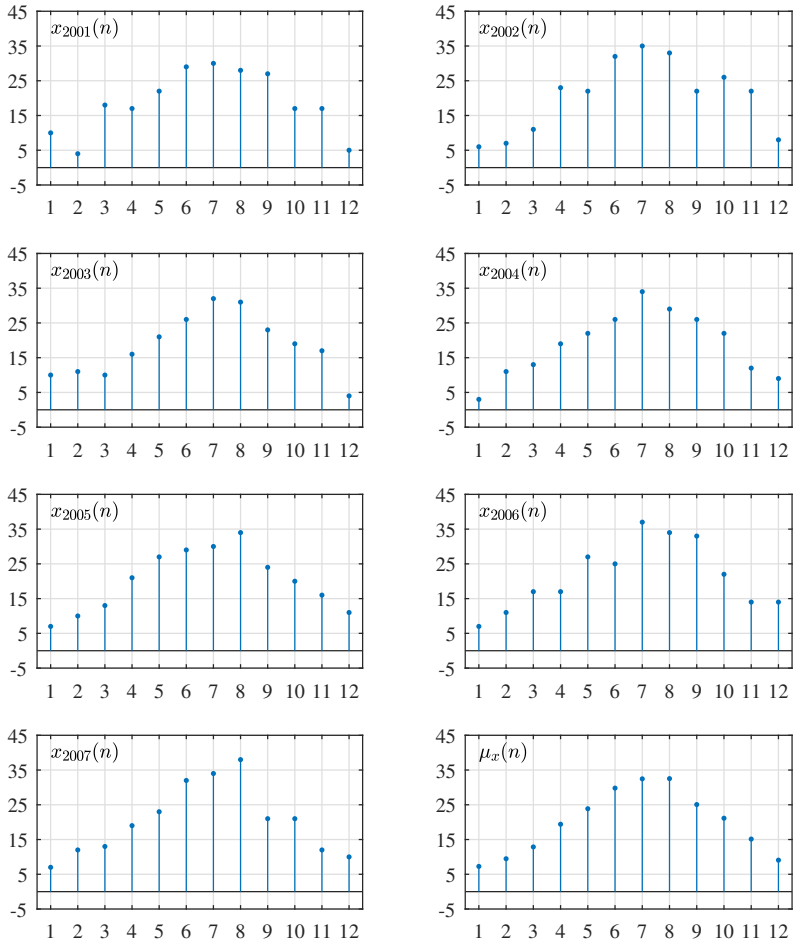


Figure 7.5 Several realizations of a random signal $x_{20i}(n)$, for $i = 01,02,\ldots,07$ and the mean value $\mu_x(n)$ for every sample (month) over 15 available realizations.

smallest possible, $1/M$ (where M is the number of available realizations), since only one sample (outlier) can make it unbounded.

The estimators which are robust to possible outliers in the data are defined and analyzed within robust statistics. The simplest tool in this area will be considered next.

7.1.2 Median

In addition to the sample average, a *sample median* is used as a statistic to describe of a set of random values. The median of a dataset is a value in the middle of the set of available samples, after the members of the set are sorted. If we denote the sorted values of $x(n)$ as $s(n)$

$$s(n) = \text{sort}\{x(n)\},\ n = 1,2,\ldots,N$$

then the median value is

$$\text{median}\{x(n)|\ n=1,2,\ldots,N\} = s\left(\frac{N+1}{2}\right), \text{ for an odd } N.$$

If N is even, then the median is defined as the mean value of two samples the nearest to $(N+1)/2$,

$$\text{median}\{x(n)|\ n=1,2,\ldots,N\} = \frac{s\left(\frac{N}{2}\right)+s\left(\frac{N}{2}+1\right)}{2}, \text{ for an even } N.$$

Example 7.4. Find the median of the sets
(a) $\mathbb{A}=\{-1,\ 1,-2,\ 4,\ 6,-9,\ 0\}$,
(b) $\mathbb{B}=\{-1,\ 1,-1367,\ 4,\ 35,-9,\ 0\}$, and
(c) The signal $x(n)$ from Example 7.1.

★(a) After sorting the values in the set $\mathbb{A}$ we get $\mathbb{A}=\{-9,-2,-1,\ 0,\ 1,\ 4,\ 6\}$. Therefore, $\text{median}(\mathbb{A})=0$.

(b) Similarly, $\text{median}(\mathbb{B})=0$. The mean values of these data would significantly differ.

(c) The sorted values of $x(n)$ are shown 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 the sorted samples 50 and 51. Thus, the median is defined here as the mean value of the 50th and 51st sorted sample.

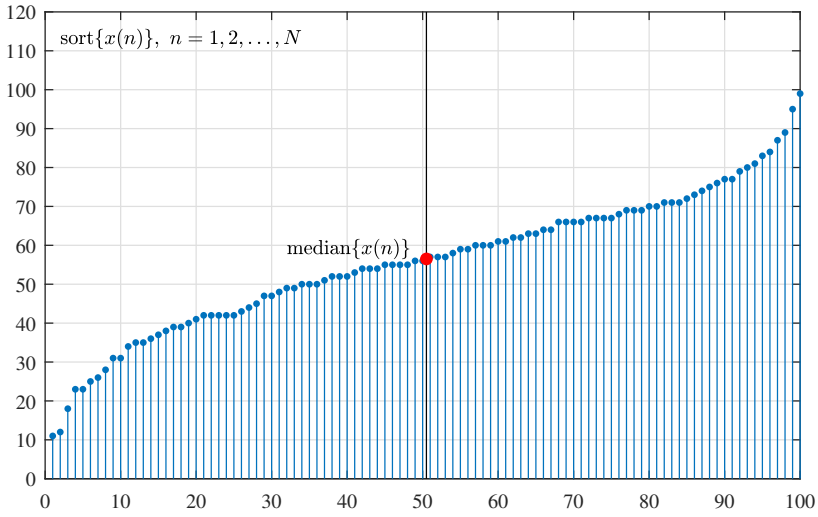


Figure 7.6 Sorted values and the median of $x(n)$.

The median will not be influenced by a possible small number of big outliers (signal values being significantly different from the values in the rest of the data). In the worst case, we have to replace $N/2$ of the realizations in order to be certain that the middle signal sample is among the

outliers and the median result will not be an inlier. Therefore, the breakdown point of this estimator is $(N/2)/N = 1/2$.

The sample average estimator was introduced by minimizing the squared distance (deviation) from the available realizations $x_i(n)$. Since the square of large errors is very large, this kind of estimator is highly influenced by the outliers. A common way to reduce the influence of large errors is to use the absolute value of the difference, instead of the squared distance in the minimization function (7.4), that is,

$$\min_{\alpha}\Big(|x_1(n)-\alpha|+|x_2(n)-\alpha|+\cdots+|x_M(n)-\alpha|\Big).$$

The same holds for the case when we consider $x(n)$, $n = 1,2,\ldots,N$, with

$$\min_{\alpha}\Big(|x(1)-\alpha|+|x(2)-\alpha|+\cdots+|x(N)-\alpha|\Big).$$

Next, we will show that the result of this minimization is the median of the considered set

$$\underset{i=1,2,\ldots,M}{\text{median}}\{x_i(n)\} = \min_{\alpha}\Big(|x_1(n)-\alpha|+|x_2(n)-\alpha|+\cdots+|x_M(n)-\alpha|\Big),$$

where $\underset{i=1,2,\ldots,M}{\text{median}}\{x_i(n)\}$ is used to denote $\text{median}\{x_i(n)|\ i = 1,2,\ldots,M\}$.

Consider the cost function

$$f(\alpha) = |x_1(n)-\alpha|+|x_2(n)-\alpha|+\cdots+|x_M(n)-\alpha| = ||\mathbf{x}-\alpha||_1 \tag{7.5}$$

and assume, without loss of generality, that the samples in $\mathbf{x} = [x_1(n), x_2(n), \ldots, x_M(n)]^T$ are already sorted, $x_1(n) \le x_2(n) \le \ldots, x_M(n)$ and that M is odd. The minimum of this function cannot be obtained as in (7.4) since this function is not differentiable at the points $\alpha = x_1(n)$, $\alpha = x_2(n)$, ..., $\alpha = x_M(n)$. However, the function $f(\alpha)$ is differentiable for all other values of α and continuous for any α. We will use this property to establish the intervals of α where it decreases and increases. The derivative of the function $|x_i(n)-\alpha|$ is equal to

$$\frac{d|x_i(n)-\alpha|}{d\alpha} = \begin{cases} -1, & \text{for } \alpha < x_i(n) \\ 1, & \text{for } \alpha > x_i(n). \end{cases}$$

Therefore, the derivative of $f(\alpha)$ within the interval on the left of the smallest signal value, $\alpha < x_1(n)$, is equal to the sum of derivatives $d|x_i(n)-\alpha|/d\alpha = -1$ of all terms, and it is equal to $df(\alpha)/d\alpha = -M$. If we move to the right along the α axis, to the interval $x_1(n) < \alpha < x_2(n)$, then the derivative of $|x_1(n)-\alpha|$ is changed to 1, while all other $M-1$ terms, have the derivatives equal to -1. This means that $df(\alpha)/d\alpha = -M+2$, in this interval. If we continue and move next to the interval $x_2(n) < \alpha < x_3(n)$, and so on, we get

$$\frac{df(\alpha)}{d\alpha} = \begin{cases} -M, & \text{for } \alpha < x_1(n) \\ -M+2, & \text{for } x_1(n) < \alpha < x_2(n) \\ \vdots & \\ -1, & \text{for } x_{(M-1)/2}(n) < \alpha < x_{(M+1)/2}(n) \\ 1, & \text{for } x_{(M+1)/2}(n) < \alpha < x_{(M+3)/2}(n) \\ \vdots & \\ M, & \text{for } \alpha > x_M(n). \end{cases}$$

as illustrated in Fig. 7.7 for $\mathbf{x} = [x_1(n), x_2(n), \ldots, x_7(n)]^T = [-0.9, -0.5, 0, 0.2, 0.7, 0.8, 1]^T$.

Obviously, the cost function $f(\alpha)$ is a decreasing function, $df(\alpha)/d\alpha < 0$, for $\alpha < x_{(M+1)/2}(n)$ and an increasing function, $df(\alpha)/d\alpha > 0$, for $\alpha > x_{(M+1)/2}(n)$. Since the function $f(\alpha)$ is continuous, this proves that

$$\underset{i=1,2,\ldots,M}{\text{median}}\{x_i(n)\} = \min_{\alpha}(f(\alpha)).$$

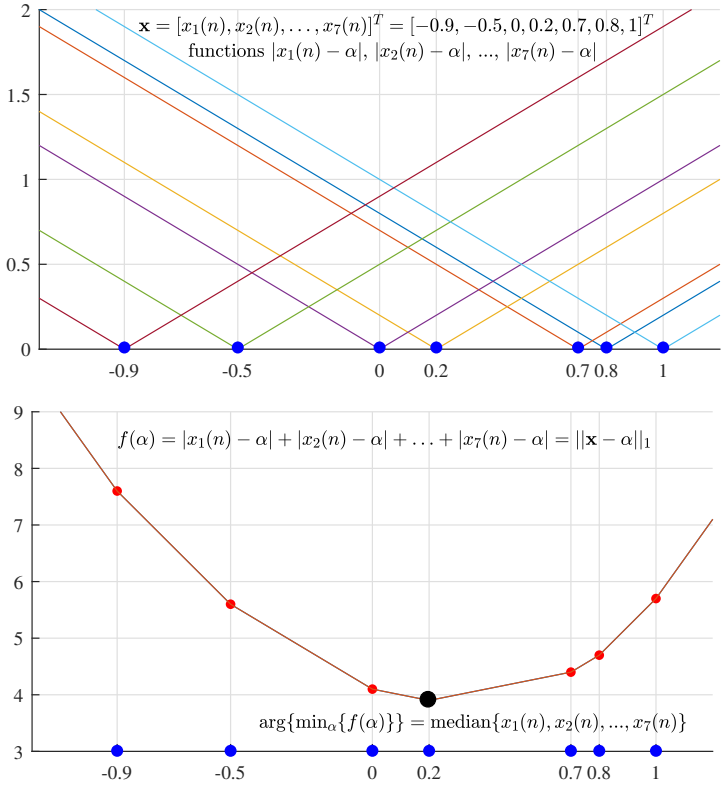


Figure 7.7 Median as the solution to the L_1-norm minimization problem.

When M is even, then $df(\alpha)/d\alpha = 0$ will be obtained for the interval $x_{M/2}(n) < \alpha < x_{M/2+1}(n)$. This means that the cost function decreases for $\alpha < x_{M/2}(n)$, it is a constant within the interval $x_{M/2}(n) < \alpha < x_{M/2+1}(n)$, and then increases for $\alpha > x_{M/2+1}(n)$. In the case of an even M, the mean value of $x_{M/2}(n)$ and $x_{M/2+1}(n)$ is used as the sample median.

In some cases the number of outliers is small. Then, the median will neglect many inlier signal values that could produce a good estimate of the mean value. In these 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 (trimmed) mean, for an odd N, as

$$\text{LSmean}\{x(n)|\ n = 1,2,\ldots,N\} = \frac{1}{2L+1}\sum_{i=-L}^{L} s\left(\frac{N+1}{2} + i\right).$$

With $L = (N-1)/2$, all signal values are used and $\text{LSmean}\{x(n)|\ n = 1,2,\ldots,N\}$ is the standard mean of a signal. With $L = 0$, the value of $\text{LSmean}\{x(n)\|\ n = 1,2,\ldots,N\}$ is equal to the sample median. In general, this way of the mean estimation is the L-statistics (α-trimmed) based estimation.

7.1.3 Variance and Standard Deviation

The next important parameter in statistics is a measure of the deviation of realizations of a random sample from the mean value. The most commonly used parameter for the description of this statistical property is the standard deviation (called the *spread*) or its squared value called the variance. For a random signal $x(n)$ whose values are available in M realizations, the variance is calculated as the mean squared deviation of the signal values from the corresponding true mean values, $\mu_x(n)$,

$$\hat{\sigma}_x^2(n) = \frac{1}{M}\left(|x_1(n)-\mu_x(n)|^2 + \cdots + |x_M(n)-\mu_x(n)|^2\right). \tag{7.6}$$

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,

$$\hat{\sigma}_x(n) = \sqrt{\frac{1}{M}\left(|x_1(n)-\mu_x(n)|^2 + \cdots + |x_M(n)-\mu_x(n)|^2\right)}. \tag{7.7}$$

If the mean value is estimated using the same set of data, $\hat{\mu}_x(n) = \frac{1}{M}\sum_{i=1}^{M} x_i(n)$, the previous estimate (which assumes the true mean value $\mu_x(n)$) tends to produce lower values of the standard deviation (biased standard deviation). Thus, an adjusted version, the *sample standard deviation*, is used as an unbiased spread measure,

$$\hat{\sigma}_x(n) = \sqrt{\frac{1}{M-1}\left(|x_1(n)-\hat{\mu}_x(n)|^2 + \cdots + |x_M(n)-\hat{\mu}_x(n)|^2\right)}. \tag{7.8}$$

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 (see Problem 7.2).

Example 7.5. For the signal $x(n)$ from Example 7.1 calculate the mean value and the variance. Compare it with the mean value and the variance of the signal $z(n)$ given in Table 7.3.

Table 7.3

Random signal $z(n)$

55	57	56	54	59	52	66	54	56	60
55	55	51	56	48	59	63	52	59	61
47	48	58	53	58	59	59	61	58	61
49	55	54	47	56	50	62	51	58	56
50	44	55	50	58	58	63	58	52	57
50	55	55	55	53	60	46	57	59	59
58	55	58	58	54	53	54	48	54	56
57	62	53	58	59	60	50	56	56	50
51	60	54	57	55	52	52	57	50	59
58	51	54	49	44	60	52	57	56	55

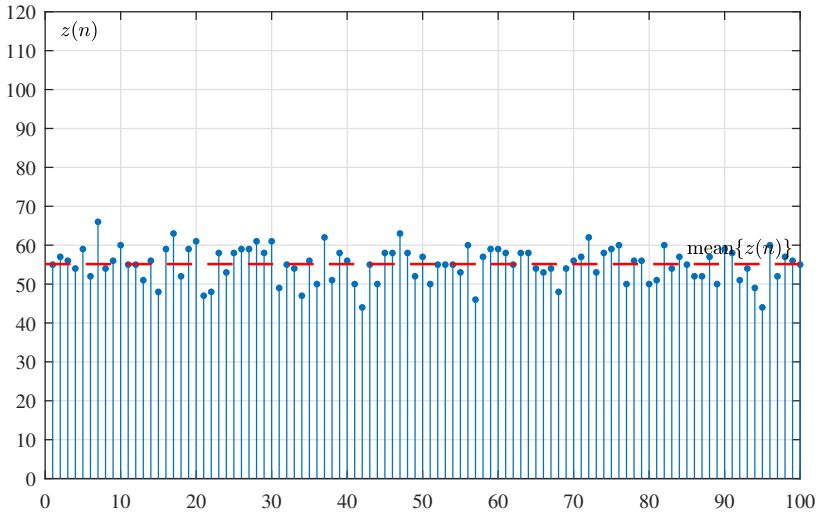


Figure 7.8 Random signal $z(n)$ from Table 7.3.

★The mean value and the variance of signal $x(n)$ are $\hat{\mu}_x = 55.76$ and $\hat{\sigma}_x^2 = 314.3863$. The standard deviation is $\hat{\sigma}_x = 17.7309$. It is a measure of the signal value deviations from the mean value. For the signal $z(n)$, the mean value is $\hat{\mu}_z = 55.14$ (very close to $\hat{\mu}_x$), while the variance is $\hat{\sigma}_z^2 = 18.7277$ and the standard deviation is $\hat{\sigma}_z = 4.3275$. Deviations of $z(n)$ from its mean value are much smaller. If the signals $x(n)$ and $z(n)$ were the measurements of the same physical value, then the individual measurements from $z(n)$ would be more reliable than the individual measurements from $x(n)$.

If we denote the sample standard deviation of the data set $\{x_i(n)\}$, where $i = 1,2,\ldots,M$, as $S(x_1(n),x_2(n)),\ldots,x_M(n)) = \sigma_x(n)$, then it satisfies the *scale property*

$$S(ax_1(n)+b,ax_2(n)+b,\ldots,ax_M(n)+b) = |a|S(x_1(n),x_2(n)),\ldots,x_M(n)).$$

The proof is simple, using (7.8) and the property that the mean value of $y_i(n) = ax_i(n)+b$ is $\hat{\mu}_y(n) = a\hat{\mu}_x(n)+b$.

The sample standard deviation is sensitive to outliers. This can be concluded form its definition (7.8). For the spread estimation, when outliers can be expected, the *median absolute deviation* (MAD) can be used as its robust measure. The MAD is defined as

$$MAD_x(n) = \underset{j=1,2,\ldots,M}{\text{median}}\{\left|x_j(n) - \underset{i=1,2,\ldots,M}{\text{median}}\{x_i(n)\}\right|\},$$

by analogy with the variance definition in (7.6), $\sigma_x^2(n) = \underset{j=1,2,\ldots,M}{\text{mean}}\{\left|x_j(n) - \underset{i=1,2,\ldots,M}{\text{mean}}\{x_i(n)\}\right|^2\}$.
The MAD value is related to the sample standard deviation as

$$MAD_x(n) = 0.6745\sigma_x(n),$$

for the Gaussian random variable (see Problem 7.11). The breakdown point for the MAD is the same as for the sample median.

7.1.4 Linear Regression Analysis

The regression analysis deals with the random variable modeling and it is widely used in various areas, including machine learning and data prediction. The most common model is the linear regression, where it has been assumed that the outcome random variable fits the linear model of the independent (also random) variable. Within the signal processing framework, we will consider a continuous-time random signal $x(t)$ sampled at random instants t_n. In linear regression, the signal model is a linear function

$$x(t_n) = at_n + b + \varepsilon(t_n), \;\; i = 1,2,\ldots,N$$

where $\varepsilon(t_n)$ is a random variable that describes the deviations of the individual realizations, $x(t_n)$, from the assumed linear model, $at_n + b$, with constant parameters a and b. The values of $\varepsilon(t_n)$ are unknown.

The aim is to estimate the linear model parameters a and b from the available data and to use them for prediction or classification of new data values. Since the values of $x(t_n)$ and t_n are available, the error function is formed as

$$e(n) = x(t_n) - at_n - b.$$

The cost function, that will be used in the minimization process, is

$$J(a,b) = f(x(t_n) - at_n - b).$$

The most common form of the cost function is defined as the sum of the squared values of the error function,

$$J(a,b) = \sum_{n=1}^{N} e^2(n) = \sum_{n=1}^{N} \Big(x(t_n) - at_n - b\Big)^2. \tag{7.9}$$

This cost function is optimal if the measurement disturbances $e(n) = \varepsilon(t_n)$ are Gaussian distributed.

The minimization of this function (*least squares* - LS minimization) is done using

$$\frac{\partial J(a,b)}{\partial a} = -2\sum_{n=1}^{N} t_n\Big(x(t_n) - at_n - b\Big) = 0$$

and

$$\frac{\partial J(a,b)}{\partial b} = -2\sum_{n=1}^{N} \Big(x(t_n) - at_n - b\Big) = 0$$

The system of equations

$$\hat{a}\sum_{n=1}^{N} t_n^2 + \hat{b}\sum_{n=1}^{N} t_n = \sum_{n=1}^{N} t_n x(t_n) \tag{7.10}$$

$$\hat{a}\sum_{n=1}^{N} t_n + \hat{b}N = \sum_{n=1}^{N} x(t_n) \tag{7.11}$$

that can be written in the form $\mathbf{A}[\hat{a} \;\; \hat{b}]^T = \mathbf{B}$, produces the estimates $\hat{a}$ and $\hat{b}$ of the linear regression model parameters a and b, $[\hat{a} \;\; \hat{b}]^T = \mathbf{A}^{-1}\mathbf{B}$. After the system of equations is solved, the values of parameters are estimated as

$$\hat{a} = \frac{\hat{\mu}_{xt} - \hat{\mu}_x\hat{\mu}_t}{\hat{\mu}_{t^2} - \hat{\mu}_t^2} \quad \text{and} \quad \hat{b} = \hat{\mu}_x - \hat{a}\hat{\mu}_t,$$

where $\hat{\mu}_{xt} = \text{mean}\{x(t_n)t_n\}$, $\hat{\mu}_x = \text{mean}\{x(t_n)\}$, $\hat{\mu}_t = \text{mean}\{t_n\}$, and $\hat{\mu}_{t^2} = \text{mean}\{t_n^2\}$.

Example 7.6. The random signal $x(t)$, whose behavior is expected to be linear, is sampled at the instants t_n

$$[t_1, t_2, \ldots, t_N]^T = [0.95,\ 0.23,\ 0.61,\ 0.49,\ 0.89, 0.76,\ 0.46,\ 0.02]^T.$$

The obtained signal values, $x(t_n)$, are

$$[x(t_1), x(t_2), \ldots, x(t_N)]^T = [4.81,\ 3.13,\ 4.25,\ 4.04,\ 4.55,\ 4.76,\ 4.16,\ 3.03]^T.$$

Find the linear regression model using the least squares approach. What is the prediction of the signal value $x(t)$ at $t = 1.1$?

★Elements of the matrices $\mathbf{A}$ and $\mathbf{B}$ in the system of equations

$$\mathbf{A}\begin{bmatrix}\hat{a} & \hat{b}\end{bmatrix}^T = \mathbf{B}$$

used for the model parameters a and b estimation, are defined by (7.10) and (7.11),

$$\mathbf{A} = \begin{bmatrix}3.15 & 4.41\\ 4.41 & 8.00\end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix}19.50\\ 32.73\end{bmatrix}, \quad \text{with}$$

$$\begin{bmatrix}\hat{a}\\ \hat{b}\end{bmatrix} = \begin{bmatrix}3.15 & 4.41\\ 4.41 & 8.00\end{bmatrix}^{-1}\begin{bmatrix}19.50\\ 32.73\end{bmatrix} = \begin{bmatrix}2.03\\ 2.97\end{bmatrix}.$$

The estimated linear regression model is

$$x(t_n) = 2.03t_n + 2.97.$$

For $t_n = 1.1$, we can predict $x(1.1) = 5.2$. The data and the results are shown in Fig. 7.9.

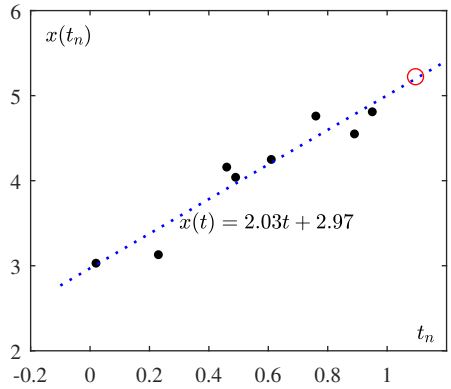


Figure 7.9 The data $x(t_n)$ measured at t_n for $n = 1,2,3,4,5,6,7,8$ (dots) and the linear model, $x(t_n) = 2.03t_n + 2.97$, obtained by the least squares approach (dotted line). The predicted signal value at $t_n = 1.1$ is marked by the circle.

The matrix form of (7.9) is $J(\mathbf{a}) = ||\mathbf{x} - \mathbf{T}\mathbf{a}||_2^2$, where

$$\mathbf{x} = [x(t_1), x(t_2), \ldots, x(t_N)]^T, \quad \mathbf{T} = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_N \end{bmatrix}, \quad \text{and} \quad \mathbf{a} = \begin{bmatrix} b \\ a \end{bmatrix}.$$

The solution to the minimization problem is obtained from $\partial J(\mathbf{a})/\partial \mathbf{a}^T = \mathbf{0}$ or $-2\mathbf{T}^T(\mathbf{x} - \mathbf{T}\hat{\mathbf{a}}) = \mathbf{0}$ as

$$\hat{\mathbf{a}} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{x}.$$

The regression analysis can be generalized to the cases with more than one independent variable. These regression forms will also be considered in Section 7.1.6, after the RANSAC method is presented in the next section.

7.1.5 Random Sample Consensus (RANSAC)

When the mean value of the data and the spread measure (commonly standard deviation) are known, we can define a criterion to identify the outliers in the data. The function used for this purpose is

$$z(n) = \frac{x(n) - \hat{\mu}_x(n)}{\hat{\sigma}_x(n)},$$

and it is called the z-score. It is common to assume the threshold value $T = 2.5$ and to declare the signal samples with $|z(n)| \leq T$ as inliers and $|z(n)| > T$ as outliers. The meaning of the threshold $T = 2.5$ will be explained later. Since the values of the average $\hat{\mu}_x(n)$ and the sample standard deviation $\hat{\sigma}_x(n)$ can be significantly compromised with possible outliers, it is recommended to use the median and the corresponding MAD in the z-score.

The random sample consensus (RANSAC) is used for linear regression when the outliers in the data are expected. Consider a set of data $x(t_n)$ sampled at random instants t_n and assume that the true data values fit a linear model. Since a large number of outliers is expected, the linear model can be far from most of the data samples. In the RANSAC approach we will:

1. Assume a small subset $\mathbb{S}$ with S randomly selected indices, t_n, of samples $x(t_n)$, $n \in \mathbb{S}$.

2. The samples with indices in $\mathbb{S}$ are used to estimate the linear regression model parameters,

$$\hat{a}\sum_{n\in\mathbb{S}} t_n^2 + \hat{b}\sum_{n\in\mathbb{S}} t_n = \sum_{n\in\mathbb{S}} t_n x(t_n)$$
$$\hat{a}\sum_{n\in\mathbb{S}} t_n + \hat{b}N = \sum_{n\in\mathbb{S}} x(t_n).$$

3. After the linear regression parameters a and b are estimated from

$$[\hat{a} \;\; \hat{b}]^T = \mathbf{A}_S^{-1}\mathbf{B}_S,$$

the line

$$x = \hat{a}t + \hat{b}$$

is defined. The distances d_n of all data points $(t_n, x(t_n))$, $n = 1, 2, \ldots, N$ from this line are calculated,

$$d_n = \frac{|\hat{a}t_n + \hat{b} - x(t_n)|}{\sqrt{1 + \hat{a}^2}}.$$

4. If a sufficient number of data points is such that their distance from the model line is lower than an assumed distance threshold d, then all these points are included into new set of data

$$\mathbb{D} = \{(t_n, x(t_n)) | \quad d_n \leq d\},$$

and the final estimation of the parameters a and b (for machine learning or prediction) is obtained with all data from $\mathbb{D}$.

5. If there was no sufficient number of data points within the distance d, a new random small set of data, $i \in \mathbb{S}$, is taken and the procedure is repeated from Step 2.

6. The procedure is stopped when the desired number of data points within $\mathbb{D}$ is achieved or the maximum number of trials is reached.

Example 7.7. Consider $N = 20$ random signal simples $x(t_n)$

$$[x(t_1), x(t_2), \ldots, x(t_N)]^T = [6.10, 3.09, 3.23, 6.90, 3.53, 3.67, 3.64, 3.97, 3.85, 4.08, 3.68, 4.05, 4.30, 4.27, 4.90, 4.56, 4.75, 1.80, 4.57, 2.9]$$

taken at the corresponding random instants t_n

$$[t_1, t_2, \ldots, t_N]^T = [0.022, 0.034, 0.200, 0.303, 0.307, 0.376, 0.429, 0.443, 0.519, 0.525, 0.538, 0.598, 0.704, 0.715, 0.837, 0.841, 0.899, 0.910, 0.953, 0.954].$$

Find the linear regression model parameters a and b for these data. Comment the results. Next, apply the RANSAC approach as follows: Use $S = 4$ randomly chosen samples with indices $n \in \mathbb{S} = \{8, 10, 18, 19\}$. Find the linear regression model parameters with this subset of data. How many data points are within the distance $d = 0.25$ from the obtained linear model line? If the number of the data points within these lines is below assumed $T = 10$, chose another random set $\mathbb{S}$. In that case use $\mathbb{S} = \{5, 11, 16, 19\}$ and repeat the procedure. If the number of data points between these lines is not below $T = 10$, use all the data points within these lines to find the linear regression model parameters.

★ For the given data set, the estimates of the linear regression model parameters a and b are obtained from

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} 7.8107 & 11.1070 \\ 11.1070 & 20.0000 \end{bmatrix}^{-1} \begin{bmatrix} 44.3483 \\ 81.8400 \end{bmatrix} = \begin{bmatrix} -0.6707 \\ 4.4645 \end{bmatrix}$$

The linear model

$$x(t_n) = -0.6707 t_n + 4.4645.$$

does not fit the data due to evident outliers at t_1, t_4, t_{18}, and t_{20} used in calculation, Fig. 7.10(a).

The RANSAC approach is started with the random selection of the data subset with $S = 4$ samples. This random subset turn out to be $\mathbb{S} = \{8, 10, 18, 19\}$. Since, one of the outliers, at t_{18}, is used in the calculation, the parameters estimation obtained from

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} 2.2082 & 2.8310 \\ 2.8310 & 4.0000 \end{bmatrix}^{-1} \begin{bmatrix} 9.8939 \\ 14.4200 \end{bmatrix} = \begin{bmatrix} -1.5245 \\ 4.6840 \end{bmatrix},$$

with the linear model $x(t_n) = -1.5245 t_n + 4.6840$ do not fit the data. This is confirmed by the fact that only $D = 8$ data points are within the lines at the distance $d = 0.25$, as can be seen in Fig. 7.10(b). Since the number of the data points within $\mathbb{D}$ is bellow $T = 10$, a new random set, $\mathbb{S} = \{5, 11, 16, 19\}$, is used. With the data corresponding to this subset, the estimation is obtained

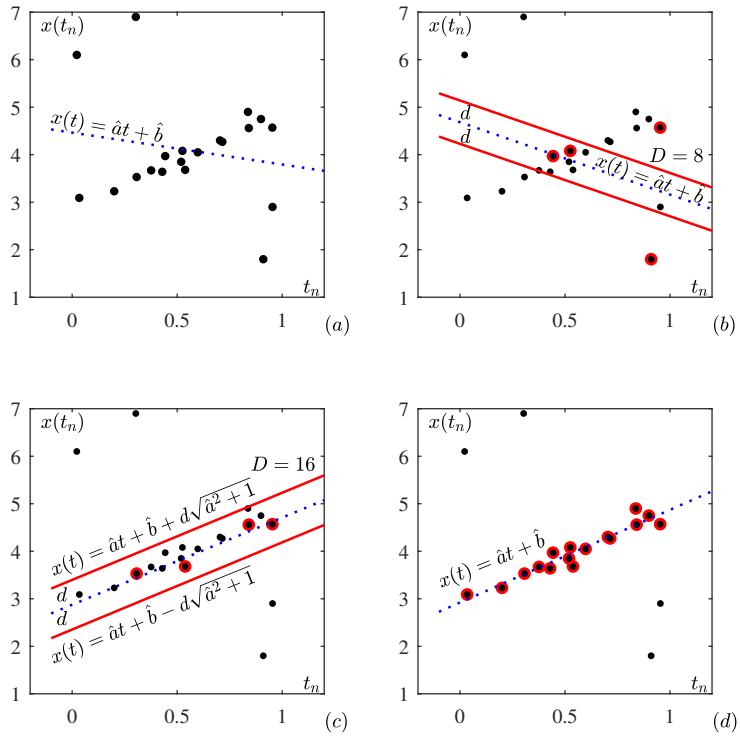


Figure 7.10 (a) The data $x(t_n)$ measured at t_n for $n = 1,2,\ldots,10$ (dots) and the linear model obtained by the least squares (dotted line). The RANSAC illustration: (b) The data (dots) and the linear model obtained by the least squares using a random subset of 4 marked samples at $\mathrm{S} = \{8,10,18,19\}$ (dotted line). (c) The data (dots) and the linear model obtained by the least squares using another random subset of 4 marked samples at $\mathrm{S} = \{5,11,16,19\}$ (dotted line). (d) The data (dots) and the linear model obtained by the least squares using all data at marked samples in $\mathbb{D}$ (dotted line).

from

$$\begin{bmatrix}\hat{a}\\ \hat{b}\end{bmatrix} = \begin{bmatrix}1.9992 & 2.6390\\ 2.6390 & 4.0000\end{bmatrix}^{-1}\begin{bmatrix}11.2537\\ 16.3400\end{bmatrix} = \begin{bmatrix}1.8342\\ 2.8749\end{bmatrix}$$

with the linear model $x(t_n) = 1.8342t_n + 2.8749$ that fits the data, since $D = 16$ data points are within $\mathbb{D}$. Since this number is above the threshold, $T = 10$, the algorithm is stopped and the linear regression model is re-estimated using all $D = 16$ data from the set $\mathbb{D}$, producing

$$\begin{bmatrix}\hat{a}\\ \hat{b}\end{bmatrix} = \begin{bmatrix}5.9802 & 8.9180\\ 8.9180 & 16.0000\end{bmatrix}^{-1}\begin{bmatrix}37.7188\\ 64.1400\end{bmatrix} = \begin{bmatrix}1.9502\\ 2.9218\end{bmatrix}$$

and the final estimated linear regression model $x(t_n) = 1.9502t_n + 2.9218$.

The probability that a subset of $S = M$ data points is free from the I outliers in a data sequence with N samples, is calculated in Example 7.10. This probability can be used to estimate the expected number of iterations in the RANSAC approach.

7.1.6 Ridge Regression

The regression can be generalized to more than one independent variable (multivariable) cases, when the considered sample, $x(n)$, is a function of the random variables $t_1(n)$, $t_2(n)$, ..., $t_M(n)$, that is, $x(n) = x(t_1(n), t_2(n), \ldots, t_M(n))$. The regression can be written in the form of a multidimensional linear model,

$$x(n) = a_1 t_1(n) + a_2 t_2(n) + \cdots + a_M t_M(n) + \varepsilon(n), \quad n = 1, 2, \ldots, N$$

where $t_i(n)$, $i = 1, 2, \ldots, M$, are independent variables for the sample $x(n)$. This system of equations can be written in the matrix/vector form as

$$\mathbf{x} = \mathbf{T}\mathbf{a} + \boldsymbol{\Xi}, \quad \text{where}$$

$$\mathbf{x} = \begin{bmatrix} x(1) \\ x(2) \\ \vdots \\ x(N) \end{bmatrix}, \quad \mathbf{T} = \begin{bmatrix} t_1(1) & t_2(1) & \ldots & t_M(1) \\ t_1(2) & t_2(2) & \ldots & t_M(2) \\ \vdots & \vdots & \ddots & \vdots \\ t_1(N) & t_2(N) & \ldots & t_M(N) \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_M \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\Xi} = \begin{bmatrix} \varepsilon(1) \\ \varepsilon(2) \\ \vdots \\ \varepsilon(N) \end{bmatrix}.$$

Common goal is to minimize the squared error between the data, $\mathbf{x}$, and the model, $\mathbf{T}\mathbf{a}$,

$$J(\mathbf{a}) = ||\mathbf{x} - \mathbf{T}\mathbf{a}||_2^2 = (\mathbf{x} - \mathbf{T}\mathbf{a})^T(\mathbf{x} - \mathbf{T}\mathbf{a}) = (\mathbf{x}^T - \mathbf{a}^T\mathbf{T}^T)(\mathbf{x} - \mathbf{T}\mathbf{a}).$$

The solution to this (least squares) minimization problem follows from $\partial J(\mathbf{a})/\partial \mathbf{a}^T = \mathbf{0}$, as

$$-2\mathbf{T}^T(\mathbf{x} - \mathbf{T}\hat{\mathbf{a}}) = \mathbf{0} \quad \text{or} \quad \mathbf{T}^T\mathbf{x} = \mathbf{T}^T\mathbf{T}\hat{\mathbf{a}}.$$

The estimation of the regression parameters $\mathbf{a} = [a_1, a_2, \ldots, a_N]^T$ that minimize the cost function $J(\mathbf{a})$, are obtained from

$$\hat{\mathbf{a}} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{x} = \text{pinv}\{\mathbf{T}\}\mathbf{x}, \tag{7.12}$$

where $\text{pinv}\{\mathbf{T}\} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T$ is the co called pseudo-inverse of the matrix $\mathbf{T}$.

The sensitivity of the reconstructed coefficients to the random variations in data $x(n)$, caused by the noise $\varepsilon(n)$, highly depends on the condition number of the matrix $\mathbf{T}^T\mathbf{T}$, whose inverse is to be calculated. For a high condition number of this matrix, corresponding to a relatively small value of the determinant $\det\{\mathbf{T}^T\mathbf{T}\}$, a small noise $\varepsilon(n)$ in the input data causes very high variations of the resulting parameters $\hat{\mathbf{a}} = [\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_M]^T$ in the model (ill-posed problem). In order to regularize the inversion (and to limit possible extremely large elements in the inverse matrix) a small value λ is added before the inversion, and the vector of parameters is calculated using

$$\hat{\mathbf{a}} = (\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\mathbf{x}. \tag{7.13}$$

It can easily be shown that this form of $\mathbf{a}$ is the solution to the minimization of the cost function $J(\mathbf{a}) = ||\mathbf{x} - \mathbf{T}\mathbf{a}||_2^2$, when the energy of the coefficients, $\mathbf{a} = [a_1, a_2, \ldots, a_M]^T$, constraint is added. The energy constraint in the minimization keeps the energy of $\mathbf{a} = [a_1, a_2, \ldots, a_M]^T$ as low as possible (in order to avoid high values of its elements, due to the possible instability of the inversion of $\mathbf{T}^T\mathbf{T}$). This is the reason why the regression estimation is called *shrinkage estimation* as well. The constrained cost function is of the form

$$J(\mathbf{a}) = ||\mathbf{x} - \mathbf{T}\mathbf{a}||_2^2 + \lambda||\mathbf{a}||_2^2,$$

where $||\mathbf{a}||_2^2$ is the L_2-norm of $\mathbf{a}$. The minimum value of this cost function is obtained from

$$\frac{\partial J(\mathbf{a})}{\partial \mathbf{a}^T} = -2\mathbf{T}^T(\mathbf{x} - \mathbf{T}\mathbf{a}) + 2\lambda\mathbf{a} = \mathbf{0}$$

in the form given by (7.13).

If the constraint is defined as the L_1-norm of the coefficients $\mathbf{a} = [a_1, a_2, \ldots, a_N]^T$, as in (7.5), then the cost function is

$$J(\mathbf{a}) = ||\mathbf{x} - \mathbf{T}\mathbf{a}||_2^2 + \lambda ||\mathbf{a}||_1.$$

In contrast to the L_2-norm constraint, which enforces small values (energy) of $\mathbf{a} = [a_1, a_2, \ldots, a_M]^T$, the L_1-norm constraint enforces sparse solution for $\mathbf{a} = [a_1, a_2, \ldots, a_M]^T$ (solution with as many zero-valued elements in $\mathbf{a} = [a_1, a_2, \ldots, a_M]^T$ as possible). The minimization of this cost function is known as *the least absolute shrinkage and selection operator* (LASSO). The LASSO solution cannot be obtained in a closed form.

This minimization problem is of great importance in compressive sensing and machine learning and will be considered in detail in Part VI. For the Bayesian framework interpretation see Example 7.28. The bias and variance of the estimated model parameters using the ridge regression and the least-squares are considered in Problem 7.3.

Example 7.8. The random variable $\mathbf{x} = [x(1), x(2), \ldots, x(N)]^T$ is measured at $N = 10$ instants. It is known that the random variable $x(n)$ is a linear function of $M = 7$ independent variables $\mathbf{t}(n) = [t_1(n), t_2(n), \ldots, t_M(n)]$, The values of the independent variables are given in the matrix $\mathbf{T}$, where $\mathbf{t}(n)$ are its rows, and the observed random variable $x(n)$ is given in the vector $\mathbf{x}$. Find the linear model parameters $\mathbf{a} = [a_1, a_2, \ldots, a_M]^T$ using the ridge regression and $\lambda = 0.01$.

$$\mathbf{T} = \begin{bmatrix} 1.1661 & 1.1607 & 0.5404 & 0.1656 & -0.8154 & -0.7506 & 0.4005 \\ 0.0884 & 0.3245 & 0.4657 & 0.0397 & 1.1129 & -0.1554 & 1.1253 \\ 0.2168 & 0.6992 & -0.2677 & 0.9177 & -0.4409 & -0.5107 & -0.4384 \\ -0.5292 & 0.9924 & 0.8458 & -0.7220 & -0.0592 & 0.9896 & 1.1560 \\ 0.2099 & 0.8470 & -0.5224 & 0.6903 & 0.2075 & 0.1795 & 0.5974 \\ -0.3515 & -1.2374 & 0.1422 & 1.0570 & -0.5864 & 0.2041 & 0.2922 \\ 0.6215 & 1.4229 & 0.5253 & 0.6104 & 0.3214 & -1.1383 & 0.3013 \\ 0.7764 & 0.9572 & 0.5978 & -0.3710 & 1.0122 & -1.2106 & -0.7862 \\ -1.3638 & -0.6115 & -0.1246 & 0.1535 & -0.0534 & 0.6406 & 0.3441 \\ -0.4296 & -0.7374 & 0.6755 & 0.3165 & -0.3022 & 0.4979 & -0.9710 \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} 3.1399 \\ -0.9308 \\ 0.8620 \\ -0.9898 \\ 0.2084 \\ -0.1228 \\ 0.9128 \\ 0.5337 \\ -2.6691 \\ -0.5552 \end{bmatrix}.$$

★ The estimation of the linear model parameters, obtained from the ridge regression (7.13), are

$$\hat{\mathbf{a}} = (\mathbf{T}^T\mathbf{T} + 0.01\mathbf{I})^{-1}\mathbf{T}^T\mathbf{x} = [1.9787, -0.0023, -0.0047, -0.0186, -1.0095, -0.0208, 0.006]^T.$$

The LASSO minimization, with the same penalty factor $\lambda = 0.01$, would produce the estimation

$$\hat{\mathbf{a}} = \text{lasso}(\mathbf{x}, \mathbf{T}, 0.01) = [1.9767,\ 0,\ 0,\ 0, -0.9776,\ 0,\ 0]^T,$$

enforcing as many zero-valued elements in $\mathbf{a}$ as possible. Due to this property (crucial in *sparse signal processing and compressive sensing*), the LASSO minimization would be able to produce the solution even in the case when the number of observations is smaller than the number of elements in $\mathbf{a}$. If we keep just the first $N = 6 < M = 7$ rows in $\mathbf{T}$ and $\mathbf{x}$ we would get

$$\hat{\mathbf{a}} = \text{lasso}(\mathbf{x}, \mathbf{T}, 0.01) = [1.9931,\ 0,\ 0.0004,\ 0, -0.9772,\ 0,\ 0]^T.$$

In this case $\det\{\mathbf{T}^T\mathbf{T}\} = 0$, since $N < M$.

A specific form of linear regression is the *polynomial fitting*

$$x(t_n) = a_0 + a_1 t_n + a_2 t_n^2 + \cdots + a_M t_n^M + \varepsilon(t_n), \;\; n = 1,2,\ldots,N.$$

This regression is still linear since linearity holds with respect to the model parameters $a_1, a_2, \ldots, a_N$. The independent variables matrix is given by

$$\mathbf{T} = \begin{bmatrix} 1 & t_1 & t_1^2 & \ldots & t_1^M \\ 1 & t_2 & t_2^2 & \ldots & t_2^M \\ \vdots & & & & \\ 1 & t_N & t_N^2 & \ldots & t_N^M \end{bmatrix}, \tag{7.14}$$

with all other vectors, results, and comments as in the previous multivariable case. Within the polynomial fitting framework, the *regularization* constraints on the solution, prevent over-fitting the model and keep the parameters low (see Problem 7.3).

7.2 BASIC PROBABILITY DEFINITIONS

Probability theory is a scientific discipline dealing with the analysis of random phenomena through a set of axioms. The outcomes of a random event are determined by chance. Probability is a measure of the likelihood of an event to occur.

For the calculation of the parameters of the first-order statistics, it is sufficient to know the probability or the probability density function of a random variable, as its basic probabilistic description.

7.2.1 Probability

Assumes that a random signal, $x(n)$, may take one of discrete values (amplitudes), ξ_i, from the set $\mathbb{A} = \{\xi_1, \xi_2, \ldots, \xi_N\}$. Then, we deal with probabilities that the random signal, $x(n)$, at an instant n takes a specific value ξ_i from the set of all possible values,

$$\text{Probability}\{x(n) = \xi_i\} = P_{x(n)}(\xi_i). \tag{7.15}$$

The probability function $P_{x(n)}(\xi)$ satisfies the following conditions (*axioms* of probability theory):
(1) $0 \leq P_{x(n)}(\xi) \leq 1$ for any ξ.
(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 ξ_i form the set $\mathbb{A}$ of all possible values of ξ is a *certain event*. Its probability is equal to 1, that is

$$\sum_{\xi \in \mathbb{A}} P_{x(n)}(\xi) = 1.$$

An *impossible event* has zero probability.

An example of a signal when the probabilities are estimated after the experiment (*a posteriori*) is already presented within Example 7.2. A posteriori probability that the signal $x(n)$ takes a value ξ_i is defined as a ratio of the number N_{ξ_i} of appearances of the event $x(n) = \xi_i$ and the total number of the performed experiments N

$$P_{x(n)}(\xi_i) = \frac{N_{\xi_i}}{N}$$

for a sufficiently large N and N_{ξ_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 number appearing in die tossing (assuming a fair balanced die), then the signal may take one of the values from the set $\xi_i \in \mathbb{A} = \{1,2,3,4,5,6\}$. In this case, the probability of each event is known in advance (*a priori*), and it is equal to $P(\xi_i) = 1/6$.

Independence. Two events (random signal samples) are independent of each other, if the probability that one event occurs (one signal sample takes a specific value) does not affects the probability of the other event occurring (does not affect the value of the other signal sample). If the signal samples $x(n)$ and $x(m)$ are statistically independent random variables then

$$\text{Probability}\left\{x(n) = \xi_i \text{ and } x(m) = \xi_j\right\} = P_{x(n)}(\xi_i)P_{x(m)}(\xi_j).$$

Exclusiveness (disjointness). Two random events (random signal sample values) are mutually exclusive or disjoint if they cannot both occur at the same time. If the signal sample values $x(n) = \xi_i$ and $x(n) = \xi_j$ are mutually exclusive events then (Property (2))

$$\text{Probability}\left\{x(n) = \xi_i \text{ or } x(n) = \xi_j\right\} = P_{x(n)}(\xi_i) + P_{x(n)}(\xi_j).$$

Example 7.9. Consider a random signal whose values are equal to the numbers appearing in a die tossing. The set of possible signal values is $\xi_i \in \mathbb{A} = \{1,2,3,4,5,6\}$. Find the probability that the signal sample takes the value $x(n) = 2$ or the value $x(n) = 5$, that is

$$\text{Probability}\{x(n) = 2 \text{ or } x(n) = 5\}.$$

Find the probability that the signal sample at an instant n takes the value $x(n) = 2$ and that in the next tossing the signal takes the value $x(n+1) = 5$, that is

$$\text{Probability}\{x(n) = 2 \text{ and } x(n+1) = 5\}.$$

★Events that $x(n) = 2$ and $x(n) = 5$ are obviously mutually exclusive. Thus, the probability of two mutual exclusive events is equal to the sum of their individual probabilities,

$$\text{Probability}\{x(n) = 2 \text{ or } x(n) = 5\} = P_{x(n)}(2) + P_{x(n)}(5) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

The events that $x(n) = 2$ and $x(n+1) = 5$ are statistically independent. In this case

$$\text{Probability}\{x(n) = 2 \text{ and } x(n+1) = 5\} = P_{x(n)}(2)P_{x(n)}(5) = \frac{1}{6}\frac{1}{6} = \frac{1}{36}.$$

Conditional probability. Conditional probability is the probability that an event A occurs, given that another event B has already occurred. The conditional probability of A, given B, is written in the form $P(A|B)$. The probability that both events A and B occur is

$$\text{Probability}\{A \text{ and } B\} = P(A|B)P(B),$$

where $P(B)$ is the probability that the event B has occurred, while $P(A|B)$ denotes the probability that the event A occurs subject to the condition that the event B already occurred.

Example 7.10. Assume that the length of random signal $x(n)$ 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 subset 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 any sample corrupted by the 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,

$$P(1) = P(B) = \frac{N-I}{N}$$

since there are N samples in total and $N - I$ of them are noise-free. After the first noise-free sample is chosen, in the remaining $(N-1)$ signal samples there are $(N-1-I)$ noise-free sample. The probability of choosing a noise-free sample is now $P(A|B) = (N-1-I)/(N-1)$. The probability that the second randomly chosen sample is not affected by the high noise, given that the first randomly chosen sample is not affected, is equal to the product of the probabilities,

$$P(2) = P(A) = P(A|B)P(B) = \frac{N-1-I}{N-1}\,\frac{N-I}{N}.$$

Here we used the conditional probability property.

Then, we continue the process of random sample 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

$$P(M) = \frac{N-I}{N}\frac{N-1-I}{N-1}\cdots\frac{N-(M-1)-I}{N-(M-1)} = \prod_{i=0}^{M-1}\frac{N-I-i}{N-i}.$$

For $N = 128$ signal samples, with $I = 16$ samples affected with high noise, the probability that $M = 32$ randomly selected samples are noise-free is equal to $P(32) = 0.0071$. If we repeat the whole procedure 1000 times, by selecting $M = 32$ samples, we can expect

$$P(32) \times 1000 = 7.1,$$

that is about 7 realizations where none of M signal samples is disturbed by the high noise. One high noise-free realization is expected in 140 realizations.

In literature, it is common to use the following calculation for the expected number of the iterations to get a high-noise free realization. The probability that one randomly selected sample is high noise-free (inlier) is $(N-I)/N)$. It is then assumed that this probability can be used for M samples (that the sample is returned and may be chosen again). The probability that there is at least one high-noise sample in M samples is $[1-((N-I)/N))^M]$. Finally, the probability of a high noise-free realization in N_{it} such trials is

$$P = 1 - [1-((N-I)/N))^M]^{N_{it}} \quad \text{where} \quad N_{it} = \frac{\ln(1-P)}{\ln(1-((N-I)/N))^M)}.$$

This calculation is correct if $(N-I-M)/(N-M) \sim (N-I)/N$, otherwise instead of $((N-I)/N))^M$ we should use $P(M) = \prod_{i=0}^{M-1}(N-I-i)/(N-i)$.

Bayes' theorem. Consider two events A and B. From the probability that both of these events happen

$$\text{Probability}\{A \text{ and } B\} = P(A|B)P(B) = \text{Probability}\{B \text{ and } A\} = P(B|A)P(A),$$

Bayes' relation follows

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{7.16}$$

Assume now that there are N possible events A_i, then $i = 1,2,\ldots,N$

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}.$$

Assume also that the events A_i, $i = 1,2,\ldots,N$ are independent and *exhaustive*. It means that two events A_i and A_j, $i \neq j$, cannot happen at the same time (independence) and that one of the events A_i, $i = 1,2,\ldots,N$, must happen (exhaustiveness). In that case, we may write

$$\begin{aligned} P(B) &= \text{Probability}\{B \text{ and (certain event)}\} = \text{Probability}\{B \text{ and } (A_1 \text{ or } A_2 \text{ or } \ldots A_N)\} \\ &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_N)P(A_N). \end{aligned}$$

Bayes' theorem is given by

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_N)P(A_N)}.$$

Since the Bayesian approach is of great importance in modern data processing we will comment on the terms in more detail.

- The event B is assumed (the evidence has happened). The probability $P(B)$ shows how probable is the assumed event (evidence) under all possible events (hypotheses) A_i. It can be written in the form $P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_N)P(A_N)$.
- The specific event A_i is called the hypothesis, assuming that the event B has occurred. The probability $P(A_i)$ shows how probable is the event A_i (hypothesis), independent of the event B occurrence.
- Probability $P(B|A_i)$ indicates how probable is the event B (evidence), given that the event A_i (hypothesis) is true.
- The result $P(A_i|B)$ is the probability of the event A_i (how the hypothesis A_i is probable) given the fact that the evidence B occurred.

Example 7.11. Consider four images, denoted by A_1, A_2, A_3, and A_4. In two images (A_1 and A_2) there are 20% of red pixels, in the third image (A_3) there are 30% of red pixels, while in the fourth image (A_4) there are 50% of red pixels. One image is chosen randomly and one of its pixels is observed. The chosen pixel is red (evidence B). What is the probability that the image A_4 was chosen?

★The probability that the image A_4 was chosen (hypothesis A_4) when the red pixel is obtained as the evidence (denoted by B) is equal to

$$P(A_4|B) = \frac{P(B|A_4)P(A_4)}{P(B|A_1)P(A_1)+P(B|A_2)P(A_2)+P(B|A_3)P(A_3)+P(B|A_4)P(A_4)},$$

where:

- The probability of the red pixel $P(B)$ being obtained, under all possible events (hypotheses) A_i, $i = 1,2,3,4$, is

$$P(B)=P(B|A_1)P(A_1)+P(B|A_2)P(A_2)+P(B|A_3)P(A_3)+P(B|A_4)P(A_4)$$
$$= \frac{20}{100}\frac{1}{4}+\frac{20}{100}\frac{1}{4}+\frac{30}{100}\frac{1}{4}+\frac{50}{100}\frac{1}{4} = 0.3.$$

  This probability is obtained as a sum over all events A_i, and is called the *marginal probability*.

- The probability that the hypothesis A_4 occurred, independently of the pixel color (independent of the event B) is

$$P(A_4) = 1/4 = 0.25,$$

  since there are four images whose choice is equally probable, $P(A_1) = P(A_2) = P(A_3) = P(A_4) = 1/4$.

- Probability $P(B|A_4)$ indicates how probable is the event B (red pixel evidence) given that the event A_4 (hypothesis that the fourth image is already chosen) is true. Its value is

$$P(B|A_4) = \frac{50}{100} = 0.5,$$

  since there are 50% red pixels in image A_4.

- The resulting probability of the image A_4 being chosen, given the evidence B that the red pixel occurred, is equal to

$$P(A_4|B) = \frac{0.5 \times 0.25}{0.3} = 0.4167.$$

Example 7.12. A system used for virus testing has reported its expected reliability. The probability of correct, positive or negative, results for a tested person is very high and it is equal to 0.978. The probability of a false-positive result (the tested person does not have the specified virus, but the test is positive) is $P_{F+} = 0.018$. The probability of false-negative results (the tested person does have the specified virus, but the test is negative) is $P_{F-} = 0.004$. What is the expected number of positive results in 1000 randomly tested persons from a country if the expected number of the contaminated people by the virus is: (a) 1 per 1000 people ($p = P(V) = 10^{-3}$); (b) 1 per 10,000 people ($p = P(V) = 10^{-4}$); (c) 1.75 per 100 people ($p = P(V) = 0.0175$); and (d) 25 per 100 of the selected population set for testing (formed by a prior evaluation of other symptoms)?

A randomly selected person is tested for the virus and the result is positive. Find the probability that this person is contaminated by the virus in all four previous cases.

★ The probability of a positive result is equal to the sum of the probability that the tested person does not have the virus and that the result is positive, $(1 - P(V))P_{F+}$, and the probability that the person does have the virus and the result is positive, $P(V)(1 - P_{F-})$. This means that the test of a randomly selected person is positive with the probability

$$P(+) = (1 - P(V))P_{F+} + P(V)(1 - P_{F-}) = (1 - p)P_{F+} + p(1 - P_{F-}).$$

For the given expected number of the virus contaminated people, $p = P(V)$, we get: (a) $P(+) = 0.019$, meaning that there will be 19 positive results in 1000 randomly tested persons, although the expected rate is 1 per 1000, in this case. Therefore, most of the test results are

false-positive. (b) $P(+) = 0.0181$, confirming that the false-positive dominates the test again. (c) In this case, $P(+) = 0.035$, meaning that half of the positive tested are indeed the people contaminated with the virus, since this probability, with an ideal test, would indicate that there are 3.5 contaminated people per 100. (d) For the selected symptomatic set of people, with a relatively high probability of the virus occurrence, we get $P(+) = 0.2625$, meaning that the agreement with the expected number of the contaminated people in this set is high.

The conclusion is that a *random screening* of the population would not produce a satisfactory result if the occurrence rate of the virus (disease) is low and the test is not an ideal one with the zero false-positive and false-negative rates. Testing should be done on a preselected set of people. This conclusion will be even more obvious from the next Bayes' analysis.

When a randomly selected person is tested and the result is positive, then the a posterior probability of the event that the person is virus contaminated, given the positive test, $P(V|+)$, is

$$P(V|+)=\frac{P(+|V)P(V)}{P(+)}=\frac{(1-P_{F-})P(V)}{(1-P(V))P_{F+}+P(V)(1-P_{F-})}=\frac{(1-P_{F-})p}{(1-p)P_{F+}+p(1-P_{F-})},$$

where $P(+|V)$ is the probability of the positive test, given that the person is virus contaminated, $p = P(V)$ is the probability that a random person has the virus, and $P(+)$ is the probability that the test is positive, including both cases that the person does and does not have the virus.

For the three considered cases, the values of the probability $P(V|+)$ values are: (a) $P(V|+) = 0.0525$, (b) $P(V|+) = 0.0055$, and (c) $P(V|+) = 0.4964$. (d) For the selected set of people, with a significant probability of having the virus, we get $P(V|+) = 0.9486$, meaning a high reliability of the test results (if the test is positive the person is contaminated).

These results confirm the conclusion that the *random test should not be done*, unless the probability of the virus (disease) in the tested people is increased using other symptoms, meaning that the set of the tested people will contain the virus (disease) with a significant probability.

In the previous example, we assumed the probability of $p = P(V)$. Commonly it is not known and should be estimated based on the posterior evidence that k out of N test were positive, with the given testing system. This problem will be considered in Section 7.4.3.

7.2.2 Expected Value and Variance

The *mean (average) value* is calculated over a set of available samples, resulting from an experiment and it is also of random nature. If the probabilistic description of a random signal (variable) is known, then we can predict the mean (average) value of this signal without using its specific random realization or performing experiments with random trials. This analytically obtained value is called the *expected value* and it represents the true value of the mean that would be obtained with a large number of experiments. The expected value is deterministic.

Expected value. The expected value of the signal sample $x(n)$ is calculated as a sum over the set of possible amplitudes, $\xi \in \mathbb{A} = \{\xi_1, \xi_2, \ldots, \xi_N\}$, weighted by the corresponding probabilities,

$$\mu_x(n) = \mathrm{E}\{x(n)\} = \sum_{\xi \in \mathbb{A}} \xi P_{x(n)}(\xi). \tag{7.17}$$

Variance. The variance of a random signal sample $x(n)$ which takes the values ξ from the discrete set $\mathbb{A} = \{\xi_1, \xi_2, \ldots, \xi_N\}$, with the known probabilities, $P_{x(n)}(\xi_i)$, is defined as

$$\sigma_x^2(n) = \mathrm{E}\{|x(n) - \mu_x(n)|^2\} = \sum_{\xi \in \mathbb{A}} |\xi - \mu_x(n)|^2 P_{x(n)}(\xi).$$

Example 7.13. A random signal $x(n)$ can take values from the set $\xi \in \mathbb{A} = \{0,1,2,3,4,5\}$. It is known that for $k = 0,1,2,3,4$ the probability of $x(n) = k$ is twice higher than the probability of $x(n) = k+1$. Find the probabilities $P_{x(n)}(\xi_k) = P\{x(n) = k\}$. Find the expected value and the variance of this random signal.

★Assume that $P\{x(n) = 5\} = A$ for $k = 5$. Then the probabilities that $x(n)$ takes a value k are

$\xi_k = k$	0	1	2	3	4	5
$P_{x(n)}(\xi_k) = P\{x(n) = k\}$	$32A$	$16A$	$8A$	$4A$	$2A$	A

Constant A can be found from

$$\sum_{k=0}^{5} P_{x(n)}(\xi_k) = 1.$$

Its value is $A = 1/63$. Now we have

$$\mu_x(n) = \mu_x = \sum_{k=0}^{5} kP_{x(n)}(\xi_k) = \frac{19}{21}$$

$$\sigma_x^2(n) = \sigma_x^2 = \sum_{k=0}^{5} \left(k - \frac{19}{21}\right)^2 P_{x(n)}(\xi_k) = \frac{626}{441}.$$

7.2.3 Probability Density Function

If a random signal can take continuous values in amplitude then we cannot define the probability that one exact signal amplitude value ξ is taken by the signal sample $x(n)$. In this case, the probability density function $p_{x(n)}(\xi)$ should be used. This function defines the probability that the nth signal sample $x(n)$ takes a value within an infinitesimally small interval $d\xi$ around ξ,

$$\text{Probability}\{\xi \le x(n) < \xi + d\xi)\} = p_{x(n)}(\xi)d\xi. \tag{7.18}$$

Properties of the probability density function $p_{x(n)}(\xi)$ are:
(1) It is nonnegative, $p_{x(n)}(\xi) \ge 0$ for any ξ/
(2) Since $\text{Probability}\{-\infty < x(n) < \infty\} = 1$, then

$$\int_{-\infty}^{\infty} p_{x(n)}(\xi)d\xi = 1.$$

The probability of an event that the signal $x(n)$ value is within $a < x(n) \le b$ is

$$\text{Probability}\{a < x(n) \le b\} = \int_{a}^{b} p_{x(n)}(\xi)d\xi.$$

Cumulative *probability distribution* $F_x(\chi)$ is defined as the probability that the signal $x(n)$ value is lower than χ,

$$F_x(\chi) = \text{Probability}\{x(n) \leq \chi\} = \int_{-\infty}^{\chi} p_{x(n)}(\xi)d\xi.$$

Obviously, $\lim_{\chi\to-\infty} F_x(\chi) = 0$, $\lim_{\chi\to+\infty} F_x(\chi) = 1$,

$$\text{Probability}\{a < x(n) \leq b\} = \int_{a}^{b} p_{x(n)}(\xi)d\xi = F_x(b) - F_x(a),$$

and $F_x(b) \geq F_x(a)$ if $b > a$. The probability distribution is a nondecreasing function.

The probability density function $p_{x(n)}(\xi)$ is equal to the derivative of the probability distribution $F_x(\xi)$,

$$p_{x(n)}(\xi) = \frac{dF_x(\xi)}{d\xi}.$$

The expected value of the random variable $x(n)$, in terms of the probability density function, is

$$\mu_x(n) = \text{E}\{x(n)\} = \int_{-\infty}^{\infty} \xi p_{x(n)}(\xi)d\xi. \tag{7.19}$$

For the case of random signals whose samples take continuous amplitude values, the variance is defined by

$$\sigma_x^2(n) = \int_{-\infty}^{\infty} \left|\xi - \mu_{x(n)}\right|^2 p_{x(n)}(\xi)d\xi,$$

where $p_{x(n)}(\xi)$ is the probability density function.

Example 7.14. Consider a real-valued random signal $x(n)$ with samples whose values are uniformly distributed over the interval $-1 \leq x(n) < 1$.

(a) Find the expected value and the variance of the signal $x(n)$ samples.

(b) The signal $y(n)$ is obtained as $y(n) = x^2(n)$. Find the expected value and the variance of signal $y(n)$.

★Since the random signal $x(n)$ is uniformly distributed within the interval $-1 \leq x(n) < 1$, its probability density function is of the form

$$p_{x(n)}(\xi) = \begin{cases} A & \text{for } -1 \leq \xi < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

The value of constant A, $A = 1/2$, is obtained from $\int_{-\infty}^{\infty} p_{x(n)}(\xi)d\xi = 1$. The expected value and the variance are given by

$$\mu_x(n) = \int_{-\infty}^{\infty} \xi p_{x(n)}(\xi)d\xi = \int_{-1}^{1} \frac{1}{2}\xi d\xi = 0, \quad \sigma_x^2(n) = \int_{-\infty}^{\infty} (\xi - \mu_x(n))^2 p_{x(n)}(\xi)d\xi = \int_{-1}^{1} \frac{1}{2}\xi^2 d\xi = \frac{1}{3}.$$

The probability that the amplitude of signal $y(n)$ is not larger than an assumed χ is, by definition, the probability distribution of $y(n)$. Its form is

$$F_y(\chi) = P\{y(n) \le \chi\} = P\{x^2(n) \le \chi\} = P\{-\sqrt{\chi} \le x(n) \le \sqrt{\chi}\}$$
$$= \begin{cases} 0 & \text{for } \chi \le 0 \\ \int_{-\sqrt{\chi}}^{\sqrt{\chi}} p_{x(n)}(\xi)d\xi & \text{for } 0 < \chi \le 1 \\ 1 & \text{for } \chi > 1 \end{cases} = \begin{cases} 0 & \text{for } \chi \le 0 \\ \sqrt{\chi} & \text{for } 0 < \chi \le 1 \\ 1 & \text{for } \chi > 1, \end{cases}$$

since $y(n) \le \chi$, when $x^2(n) \le \chi$, as illustrated in Fig. 7.11.

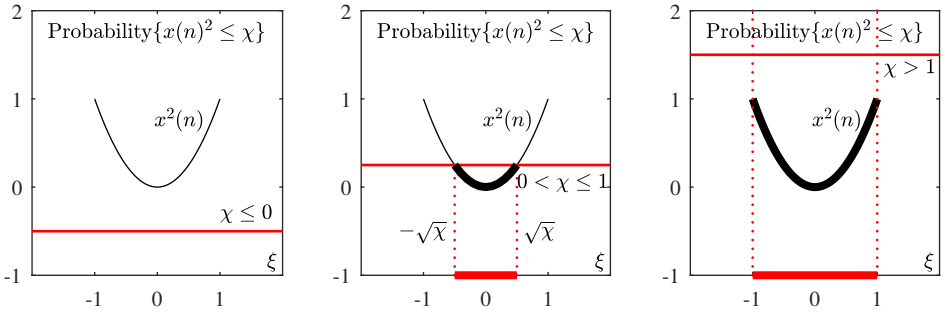


Figure 7.11 Illustration of the probability distribution $F_y(\chi)$ calculation for $y(n) = x^2(n)$, when $-1 \le x(n) \le 1$.

The probability density function is obtained as the derivative of the probability distribution $F_y(\chi)$, that is

$$p_{y(n)}(\xi) = \frac{dF(\xi)}{d\xi} = \begin{cases} \frac{1}{2\sqrt{\xi}} & \text{for } 0 < \xi \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

The expected value and the variance of the signal $y(n)$ are

$$\mu_y(n) = \mu_y = \int_0^1 \xi \frac{1}{2\sqrt{\xi}} d\xi = \frac{1}{3}, \quad \sigma_y^2(n) = \sigma_y^2 = \int_0^1 (\xi - \frac{1}{3})^2 \frac{1}{2\sqrt{\xi}} d\xi = \frac{4}{45}.$$

Example 7.15. Find the probability density function of $y(n)$ for an arbitrary monotonous function $y(n) = f(x(n))$, with inverse $f^{-1}(y(n)) = x(n)$, if the probability density function of $x(n)$ is $p_{x(n)}(\xi)$.

What is the form of $p_{y(n)}(\xi)$ if $x(n)$ is a random variable with the uniform probability density function, within the interval $[-\pi/2, \pi/2)$ and $y(n) = \tan(x(n))$, with the inverse function $x(n) = \arctan(y(n))$?

★ The probability distribution of the random signal $y(n)$ is given by

$$F_y(\chi) = P\{y(n) \le \chi\} = P\{f(x(n)) \le \chi\} = P\{x(n) \le f^{-1}(\chi)\} = \int_{-\infty}^{f^{-1}(\chi)} p_{x(n)}(\xi)d\xi$$

or

$$F_y(\chi) = F_x(f^{-1}(\chi)).$$

The probability density function is

$$p_{y(n)}(\xi) = \frac{dF_y(\xi)}{d\xi} = \frac{dF_x(f^{-1}(\xi))}{d\xi} = p_{x(n)}(f^{-1}(\xi))\Big|\frac{df^{-1}(\xi)}{d\xi}\Big|.$$

This relation can also be obtained from the fact that the probability contained in a differential area must be invariant under the change of variables, that is,

$$|p_{y(n)}(\xi)d\xi| = |p_{x(n)}(f^{-1}(\xi))df^{-1}(\xi)|.$$

For the random variable $x(n)$, with the uniform probability density function within the interval $[-\pi/2, \pi/2)$, the random variable $y(n) = \tan(x(n))$ is distributed from $-\infty$ to ∞. Its probability density function is

$$p_{y(n)}(\xi) = p_{x(n)}(\arctan(\xi))\Big|\frac{d(\arctan(\xi))}{d\xi}\Big| = \frac{1}{\pi}\frac{1}{1+\xi^2},$$

since $p_{x(n)}(\xi) = 1/\pi$, for $-\pi/2 \le \xi < \pi/2$.

The random signal $y(n)$ may take high values with a significant probability, since its probability distribution is $F_y(\chi) = (1/2 + \arctan(\chi)/\pi)$.

The random signals whose probability distribution is not exponentially bounded, that is, they have heavier (higher value) tails than the exponential distribution, are called the signals with a *heavy-tailed* probability density function. In this case, the condition for the heavy-tailed function, $\lim_{\chi\to\infty}(F_y(\infty) - F_y(\chi))e^{a\chi} = \infty$ for all $a > 0$, is satisfied (see Section 7.4.11).

As an introduction to the second-order statistics (that will be considered in the next section), consider two signals $x(n)$ and $y(n)$, with continuous amplitude values. The probability that the nth signal sample $x(n)$ takes a value within $\xi \le x(n) < \xi + d\xi$ and that $y(m)$ takes a value within $\zeta \le y(m) < \zeta + d\zeta$ is

$$\text{Probability}\{\xi \le x(n) < \xi + d\xi),\ \zeta \le y(m) < \zeta + d\zeta)\} = p_{x(n),y(m)}(\xi,\zeta)d\xi d\zeta,$$

where $p_{x(n),y(m)}(\xi,\zeta)$ is the joint probability density function. The probability of an event $a \le x(n) < b$ and $c \le y(m) < d$ is

$$\text{Probability}\{a \le x(n) < b,\ c \le y(m) < d\} = \int_a^b \int_c^d p_{x(n),y(m)}(\xi,\zeta)d\xi d\zeta.$$

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 7.16. The signal $x(n)$ is defined by $x(n) = a(n) + b(n) + c(n)$, where $a(n)$, $b(n)$, and $c(n)$ are mutually independent random signals with the uniform probability density function over the interval $[-1,1)$. Find the probability density function of the signal $x(n)$, its mean μ_x, and the variance σ_x^2.

★Consider a sum of two independent random signals $s(n) = a(n) + b(n)$. The probability that $s(n) = a(n) + b(n) \leq \chi$ can be calculated from the joint probability distribution of $a(n)$ and $b(n)$ as

$$F_s(\chi) = P\{s(n) \leq \chi\} = \text{Probability}\{-\infty < a(n) < \infty \text{ and } -\infty < a(n) + b(n) \leq \chi\}$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\chi-\zeta} p_{a(n),b(n)}(\xi,\zeta)\,d\xi\,d\zeta = \int_{-\infty}^{\infty} p_{b(n)}(\zeta) \int_{-\infty}^{\chi-\zeta} p_{a(n)}(\xi)\,d\xi\,d\zeta.$$

Now, we can calculate the probability density function of $s(n)$ as a derivative of $F_s(\chi)$, that is

$$p_{s(n)}(\chi) = \frac{dF_s(\chi)}{d\chi} = \int_{-\infty}^{\infty} p_{b(n)}(\zeta) \frac{d}{d\chi} \int_{-\infty}^{\chi-\zeta} p_{a(n)}(\xi)\,d\xi\,d\zeta$$
$$= \int_{-\infty}^{\infty} p_{b(n)}(\zeta) p_{a(n)}(\chi-\zeta)\,d\zeta = p_{b(n)}(\chi) *_\chi p_{a(n)}(\chi),$$

meaning that the probability density function of the 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

$$p_{x(n)}(\chi) = p_{c(n)}(\chi) *_\chi p_{b(n)}(\chi) *_\chi p_{a(n)}(\chi),$$

$$p_{x(n)}(\chi) = \begin{cases} \frac{(\chi+3)^2}{16} & \text{for } -3 \leq \chi \leq -1 \\ \frac{3-\chi^2}{8} & \text{for } -1 < \chi \leq 1 \\ \frac{(\chi-3)^2}{16} & \text{for } 1 < \chi \leq 3 \\ 0 & \text{for } |\chi| > 3. \end{cases}$$

The mean value and the variance can be calculated using $p_{x(n)}(\chi)$, or in a direct way using the linearity property, as

$$\mu_x = \mathrm{E}\{x(n)\} = \mathrm{E}\{a(n)\} + \mathrm{E}\{b(n)\} + \mathrm{E}\{c(n)\} = 0$$
$$\sigma_x^2 = \mathrm{E}\{(x(n) - \mu_x)^2\} = \mathrm{E}\{(a(n) + b(n) + c(n))^2\}$$
$$= \mathrm{E}\{a(n)^2\} + \mathrm{E}\{b(n)^2\} + \mathrm{E}\{c(n)^2\} + 2(\mu_a\mu_b + \mu_a\mu_c + \mu_b\mu_c)$$
$$= \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1.$$

Example 7.17. Consider two independent random signals $x(n)$ and $y(n)$, with probability density functions $p_{x(n)}(\xi)$ and $p_{y(n)}(\xi)$. A new random signal is defined is such a way that it takes the lower value of the signals $x(n)$ and $y(n)$ at every instant n,

$$z(n) = \min\{x(n), y(n)\}.$$

Find the probability distribution and the probability density function of the random signal $z(n)$. What is the probability density function of $z(n)$ if

$$p_{x(n)}(\xi) = \frac{1}{\beta_x} e^{-\xi/\beta_x} u(\xi) \quad \text{and} \quad p_{y(n)}(\xi) = \frac{1}{\beta_y} e^{-\xi/\beta_y} u(\xi)?$$

★ Since the random signal $z(n)$ takes the lower of the values $x(n)$ and $y(n)$, the probability that $z(n) = \min\{x(n), y(n)\}$ is lower than or equal to an assumed χ is equal to the probability that at least one of the random samples $x(n)$ and $y(n)$ is bellow this assumed χ, that is

$$P\{z(n) \le \chi\} = P\{\min\{x(n), y(n)\} \le \chi\}$$
$$= 1 - P\{x(n) > \chi \text{ and } y(n) > \chi\} = 1 - P\{x(n) > \chi\} P\{y(n) > \chi\}.$$

Since

$$P\{x(n) > \chi\} = 1 - F_x(\chi) \quad \text{and} \quad P\{y(n) > \chi\} = 1 - F_y(\chi),$$

we get the probability distribution of the random variable $z(n)$ in the form

$$F_z(\chi) = P\{z(n) \le \chi\} = 1 - (1 - F_x(\chi))(1 - F_y(\chi))$$
$$= F_x(\chi) + F_y(\chi) - F_x(\chi) F_y(\chi).$$

The probability density function follows as the derivative of the probability distribution,

$$p_{z(n)}(\xi) = \frac{dF_z(\xi)}{d\xi} = p_{x(n)}(\xi) + p_{y(n)}(\xi) - p_{x(n)}(\xi) F_y(\xi) - F_x(\xi) p_{y(n)}(\xi)$$
$$= p_{x(n)}(\xi)(1 - F_y(\xi)) + p_{y(n)}(1 - F_x(\xi)).$$

For the specific probability density functions,

$$p_{z(n)}(\xi) = \frac{1}{\beta_z} e^{-\xi/\beta_z} u(\xi)$$

with

$$\frac{1}{\beta_z} = \frac{1}{\beta_x} + \frac{1}{\beta_y},$$

since $F_x(\xi) = (1 - \exp(-\xi/\beta_x))u(\xi)$ and $F_y(\xi) = (1 - \exp(-\xi/\beta_y))u(\xi)$.

See also Problem 7.8 and its solution.

7.3 SECOND-ORDER STATISTICS

The correlations and covariances, as the most important parameters of the second-order statistics, will be analyzed in this section and related to the spectral power density of random signals.

7.3.1 Correlation and Covariance

Second-order statistics deals with two samples of random signals.

If the probability that a real-valued random signal $x(n)$ takes a value ξ_1 and that $x(m)$ takes ξ_2 is $P_{x(n),x(m)}(\xi_1,\xi_2)$, then the autocorrelation is defined by

$$r_{xx}(n,m) = \mathrm{E}\{x(n)x^*(m)\} = \sum_{\xi_1}\sum_{\xi_2}\xi_1\xi_2 P_{x(n),x(m)}(\xi_1,\xi_2). \tag{7.20}$$

For a real-valued random signal with continuous amplitudes of its samples and the second-order probability density function $p_{x(n),x(m)}(\xi_1,\xi_2)$, the autocorrelation is

$$r_{xx}(n,m) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\xi_1\xi_2 p_{x(n),x(m)}(\xi_1,\xi_2)d\xi_1 d\xi_2. \tag{7.21}$$

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

$$r_{xx}(n,m) = \mu_x(n)\mu_x(m).$$

For a signal $\{x_i(n)\}$, $n = 1,2,\ldots,N$ and $i = 1,2,\ldots,M$, being the index of realization of this signal, the *autocorrelation function* is estimated by

$$\hat{r}_{xx}(n,m) = \frac{1}{M}\sum_{i=1}^{M} x_i(n)x_i^*(m). \tag{7.22}$$

Autocorrelation matrix. For the signal $x_i(n)$, $i = 1,2,\ldots,M$, in the vector form

$$\mathbf{x}_i = [x_i(1),\ x_i(2),\ldots,x_i(N)]^H,$$

the autocorrelation matrix is estimated from

$$\hat{\mathbf{R}}_x = \frac{1}{M}\sum_{i=1}^{M}\mathbf{x}_i\mathbf{x}_i^H = \frac{1}{M}\sum_{i=1}^{M}\begin{bmatrix} x_i(1) \\ x_i(2) \\ \vdots \\ x_i(N)\end{bmatrix}\left[x_i^*(1),\ x_i^*(2),\ldots,x_i^*(N)\right]$$

$$= \begin{bmatrix} \hat{r}_{xx}(1,1) & \hat{r}_{xx}(1,2) & \ldots & \hat{r}_{xx}(1,N) \\ \hat{r}_{xx}(2,1) & \hat{r}_{xx}(2,2) & \ldots & \hat{r}_{xx}(2,N) \\ \vdots & & & \\ \hat{r}_{xx}(N,1) & \hat{r}_{xx}(N,2) & \ldots & \hat{r}_{xx}(N,N)\end{bmatrix}.$$

The *autocovariance* function is defined by

$$c_{xx}(n,m) = \mathrm{E}\{(x(n)-\mu_x(n))\,(x(m)-\mu_x(m))^*\}. \tag{7.23}$$

It may be easily shown that

$$c_{xx}(n,m) = \mathrm{E}\{(x(n) - \mu_x(n))\,(x(m) - \mu_x(m))^*\} = r_{xx}(n,m) - \mu_x(n)\mu_x^*(m).$$

For $m = n$, the value of the autocovariance is equal to the variance

$$\sigma_x^2(n) = c_{xx}(n,n) = \mathrm{E}\{|x(n) - \mu_x(n)|^2\} = r_{xx}(n,n) - |\mu_x(n)|^2. \tag{7.24}$$

Diagonal elements in the covariance matrix, $\mathbf{C}_x$, are the variances $\sigma_x^2(n)$.

The cross-correlation and the cross-covariance of two signals $x(n)$ and $y(n)$ are defined as

$$r_{xy}(n,m) = \mathrm{E}\{x(n)y^*(m)\}$$

and

$$c_{xy}(n,m) = \mathrm{E}\{(x(n) - \mu_x(n))\,(y(m) - \mu_y(m))^*\} = r_{xy}(n,m) - \mu_x(n)\mu_y^*(m). \tag{7.25}$$

For the signals whose samples are available, the autocovariance is estimated using the following relation,

$$\hat{c}_{xx}(n,m) = \frac{1}{M}\sum_{i=1}^{M}(x_i(n) - \hat{\mu}_x(n))\,(x_i(m) - \hat{\mu}_x(m))^*.$$

The covariance matrix can be written in the form

$$\hat{\mathbf{C}}_x = \mathrm{Cov}(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_x\})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_x\})^H = \hat{\mathbf{R}}_x - \hat{\boldsymbol{\mu}}_x\hat{\boldsymbol{\mu}}_x^H.$$

Example 7.18. Consider the nth set of experiments, where the independent variable in the experiment assume N random values $t_1(n), t_2(n), \ldots, t_N(n)$ and the result of experiment takes the random values $x_1(n), x_2(n), \ldots, x_N(n)$. If the linear model is assumed, $x_i(n) = at_i(n) + b$, show that the solution for the parameter a in this linear regression problem, can be written as

$$a = \frac{c_{tx}(n,n)}{c_{tt}(n,n)} = \frac{\mathrm{E}\{(x(n) - \mu_x(n))\,(t(n) - \mu_t(n))\}}{\mathrm{E}\{(t(n) - \mu_t(n))^2\}}$$

where the covariances are estimated using

$$\hat{c}_{tx}(n,n) = \hat{r}_{tx}(n,n) - \hat{\mu}_t(n)\hat{\mu}_x(n) = \frac{1}{N}\sum_{i=1}^{N}t_i(n)x_i(n) - \Big(\frac{1}{N}\sum_{i=1}^{N}t_i(n)\Big)\Big(\frac{1}{N}\sum_{i=1}^{N}x_i(n)\Big)$$

$$\hat{c}_{tt}(n,n) = \hat{r}_{tt}(n,n) - \hat{\mu}_t^2(n,n) = \frac{1}{N}\sum_{i=1}^{N}t_i^2(n) - \Big(\frac{1}{N}\sum_{i=1}^{N}t_i(n)\Big)^2.$$

★ The solution to the linear regression model is obtained from the system (see Section 7.1.4)

$$\frac{1}{N}\sum_{i=1}^{N}t_i(n)\Big(x_i(n) - at_i(n) - b\Big) = 0$$

$$\frac{1}{N}\sum_{i=1}^{N}\Big(x_i(n) - at_i(n) - b\Big) = 0$$

This system of equations can be written in the form

$$\begin{aligned} a r_{tt}(n,n) + b\mu_t(n) &= r_{tx}(n,n) \\ a\mu_t(n) + b &= \mu_x(n) \end{aligned}$$

with

$$a(\hat{r}_{tt}(n,n) - \hat{\mu}_t^2(n)) = \hat{r}_{tx}(n,n) - \hat{\mu}_t(n)\hat{\mu}_x(n)$$

producing the estimate of parameter a in the form $\hat{a} = \hat{c}_{tx}(n,n)/\hat{c}_{tt}(n,n)$.

7.3.2 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

$$\begin{aligned} \mu_x(n) &= \mathrm{E}\{x(n)\} = \mu_x \\ r_{xx}(n,m) &= \mathrm{E}\{x(n)x^*(m)\} = r_{xx}(n-m). \end{aligned} \tag{7.26}$$

A signal is stationary in the strict sense (SSS) if all order statistics are invariant to a shift in time.

Example 7.19. Show that for a stationary real-valued signal $x_i(n)$, $n = 1,2,\ldots,N$, $i = 1,2,\ldots,M$

$$\hat{r}_{xx}(0) \geq \hat{r}_{xx}(n) \tag{7.27}$$

holds for any $n \neq 0$.

★ For any two signal samples we can write

$$\frac{1}{M}\sum_{i=1}^{M}(x_i(n) - x_i(m))^2 \geq 0.$$

This mean that

$$\frac{1}{M}\sum_{i=1}^{M} x_i^2(n) + x_i^2(m) \geq \frac{1}{M}\sum_{i=1}^{M} 2x_i(n)x_i(m).$$

For a stationary real-valued signal, we get $\hat{r}_{xx}(0) + \hat{r}_{xx}(0) \geq 2\hat{r}_{xx}(n-m)$, what proves (7.27).

A random signal $x(n)$ is wide-sense cyclostationary (WSCS) if

$$\begin{aligned} &\mathrm{E}\{x(n)\} = \mathrm{E}\{x(n+N)\} \text{ and} \\ &\mathrm{E}\{x(n+N)x^*(m)\} = r_{xx}(N+n,m) = r_{xx}(n,m). \end{aligned}$$

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

$$M_{xxx}(n,m,l) = \mathrm{E}\{x(n)x^*(m)x^*(l)\}. \tag{7.28}$$

For stationary signals, the third-order moment is of the form

$$M_{xxx}(m,l) = \mathrm{E}\{x(n)x^*(n-m)x^*(n-l)\}.$$

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 $m = l = 0$, the third-order moment of the real-valued random variable $x(n)$, is defined by

$$M_3 = \mathrm{E}\{x^3(n)\} = \int_{-\infty}^{\infty} \xi^3 p_{x(n)} d\xi.$$

Higher-order statistics is commonly described through cumulants.

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 the averaging over realizations. The process is ergodic in parameter β if that particular parameter can be estimated by averaging over time instead of averaging over realizations. If a random signal $x(n)$ is a realization of a process ergodic in mean then

$$\mu_x(n) = \lim_{M\to\infty} \frac{1}{M}(x_1(n) + \cdots + x_M(n)) = \lim_{N\to\infty} \frac{1}{N}(x_i(n) + \cdots + x_i(n-N+1)).$$

7.3.3 Characteristic Function and Moments

The characteristic function of a random variable $x(n)$ is defined as the expected value of the random variable $y(n) = e^{j\theta x(n)}$, that is

$$\Phi_x(\theta) = \mathrm{E}\{e^{j\theta x(n)}\} = \int_{-\infty}^{\infty} p_{x(n)}(\xi) e^{j\theta\xi} d\xi.$$

It can be interpreted as the Fourier transform of the probability density function $p_{x(n)}(\xi)$ (with sign $+$ in the exponent instead of $-$). The characteristic function can be related to the moments of the random variable $x(n)$, using the Taylor series expansion of $e^{j\theta\xi}$ around zero ($\xi = 0$),

$$e^{j\theta\xi} = 1 + j\theta\xi - \frac{1}{2!}\theta^2\xi^2 - j\frac{1}{3!}\theta^3\xi^3 + \ldots.$$

The characteristic function can be written in the form

$$\begin{aligned}\Phi_x(\theta) &= \int_{-\infty}^{\infty} p_{x(n)} d\xi + j\theta \int_{-\infty}^{\infty} \xi p_{x(n)} d\xi - \frac{1}{2!}\theta^2 \int_{-\infty}^{\infty} \xi^2 p_{x(n)} d\xi - j\frac{1}{3!}\theta^3 \int_{-\infty}^{\infty} \xi^3 p_{x(n)} d\xi + \ldots \\ &= 1 + j\theta M_1 - \frac{1}{2!}\theta^2 M_2 - j\frac{1}{3!}\theta^3 M_3 + \ldots, \end{aligned} \tag{7.29}$$

where the moments M_i are defined by

$$M_i = \int_{-\infty}^{\infty} \xi^i p_{x(n)} d\xi.$$

From the series (7.29) expansion, we can conclude that

$$\Phi_x(0)=1, \quad \frac{d\Phi_x(\theta)}{jd\theta}\Big|_{\theta=0}=M_1, \quad \frac{d^2\Phi_x(\theta)}{-d\theta^2}\Big|_{\theta=0}=M_2, \quad \frac{d^3\Phi_x(\theta)}{-jd\theta^3}\Big|_{\theta=0}=M_3, \ \ldots$$

For the sum of random variables, $z(n)=x(n)+y(n)$, whose probability density function is equal to the convolution of the corresponding probability density functions , $p_{z(n)}(\xi)=p_{x(n)}(\xi)*p_{y(n)}(\xi)$ (see Example 7.16), the characteristic function is equal to the product of their individual characteristic functions,

$$\Phi_z(\theta)=\Phi_x(\theta)\Phi_y(\theta). \tag{7.30}$$

From this relation, we can easily find the moments of the sum of random variables (see also Problem 7.21 and Example 7.24).

7.3.4 Power Spectral Density

For stationary signals, the autocorrelation function is a function of the difference of time arguments,

$$r_{xx}(n)=\mathrm{E}\{x(n+m)x^*(m)\}.$$

The Fourier transform of the autocorrelation function of a WSS signal is the power spectral density

$$S_{xx}(e^{j\omega})=\sum_{n=-\infty}^{\infty} r_{xx}(n)e^{-j\omega n} \tag{7.31}$$

$$r_{xx}(n)=\frac{1}{2\pi}\int_{-\pi}^{\pi} S_{xx}(e^{j\omega})e^{j\omega n}d\omega. \tag{7.32}$$

Integral of $S_{xx}(e^{j\omega})$ over frequency,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} S_{xx}(e^{j\omega})d\omega=r_{xx}(0)=\mathrm{E}\{|x(n)|^2\}, \tag{7.33}$$

is equal to the expected power of the random signal.

Example 7.20. Find the expected value, the autocorrelation, and the power spectral density of the random signal

$$x(n)=\sum_{k=1}^{K} a_k e^{j(\omega_k n+\theta_k)},$$

where θ_k are random variables uniformly distributed over $-\pi<\theta_k\le\pi$. All random variables are statistically independent. Frequencies ω_k are $-\pi<\omega_k\le\pi$ for every k.

★The expected value is

$$\mu_x=\sum_{k=1}^{K} a_k\mathrm{E}\{e^{j(\omega_k n+\theta_k)}\}=\sum_{k=1}^{K} a_k\int_{-\pi}^{\pi}\frac{1}{2\pi}e^{j(\omega_k n+\theta_k)}d\theta_k=0.$$

The autocorrelation is

$$r_{xx}(n) = \mathrm{E}\{\sum_{k=1}^{K} a_k e^{j(\omega_k(n+m)+\theta_k)} \sum_{k=1}^{K} a_k e^{-j(\omega_k m+\theta_k)}\} = \sum_{k=1}^{K} a_k^2 e^{j\omega_k n},$$

while the power spectral density for $-\pi < \omega \leq \pi$ is

$$S_{xx}(e^{j\omega}) = \mathrm{FT}\{r_{xx}(n)\} = 2\pi \sum_{k=1}^{K} a_k^2 \delta(\omega - \omega_k).$$

The average signal power of a signal $x(n)$ has been defined as (2.9)

$$P_{AV} = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} |x(n)|^2 = \left\langle |x(n)|^2 \right\rangle.$$

This relation leads to another definition of the power spectral density of random discrete-time signals, given by

$$P_{xx}(e^{j\omega}) = \lim_{N\to\infty} \frac{1}{2N+1} \mathrm{E}\{\left|X_N(e^{j\omega})\right|^2\} = \lim_{N\to\infty} \frac{1}{2N+1} \mathrm{E}\{\left|\sum_{n=-N}^{N} x(n)e^{-j\omega n}\right|^2\}. \tag{7.34}$$

Different notation is used since the definitions of power spectral density (7.31) and (7.34), in general, will not produce the same result. We can write

$$P_{xx}(e^{j\omega}) = \lim_{N\to\infty} \frac{1}{2N+1} \mathrm{E}\{\sum_{m=-N}^{N} \sum_{n=-N}^{N} x(m)x^*(n)e^{-j\omega(m-n)}\}.$$

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 $-N \leq m \leq N,\ -N \leq n \leq N$, Fig. 7.12. 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 the area $-N \leq m \leq N,\ -N \leq n \leq N$. Along this diagonal there are $2N+1$ 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 $2N$ 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 $2N+1-|k|$ terms with $r_{xx}(m-n)e^{-j\omega(m-n)} = r_{xx}(\pm k)e^{\mp jk\omega}$. This means that we can write

$$P_{xx}(e^{j\omega}) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{k=-2N}^{2N} (2N+1-|k|)r_{xx}(k)e^{-j\omega k}$$

$$= \lim_{N\to\infty} \sum_{k=-2N}^{2N} (1 - \frac{|k|}{2N+1})r_{xx}(k)e^{-j\omega k} = \lim_{N\to\infty} \sum_{k=-2N}^{2N} w_B(k)r_{xx}(k)e^{-j\omega k}.$$

The function $w_B(k)$ corresponds to the Bartlett (triangular) window over the calculation interval.

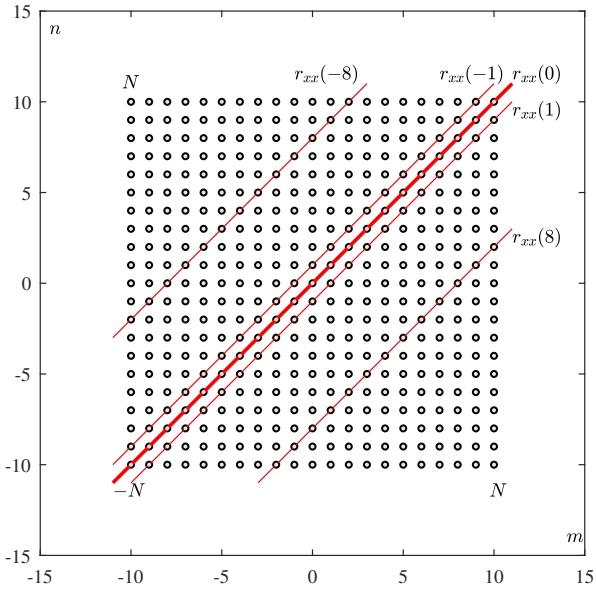


Figure 7.12 Illustration of the power spectral density domain and the autocorrelation function $r_{xx}(m-n)$.

If the values of the 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

$$P_{xx}(e^{j\omega}) = \lim_{N\to\infty} \sum_{k=-2N}^{2N} r_{xx}(k) e^{-j\omega k} = \mathrm{FT}\{r_{xx}(n)\} = S_{xx}(e^{j\omega}).$$

This is true for $r_{xx}(k) = C\delta(k)$ or $r_{xx}(k)$ being nonzero within the region $|k| < k_0$, such that $k_0/(2N+1)$ is negligible. 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 7.52). Estimation of the power spectral density will be revisited in Section 7.5.6.

Periodically extended signals. Another way to estimate the power spectral density is to assume that a WSS signal $x(n)$, $n = 0,1,2,\ldots,N-1$, is *periodically extended*. Then

$$P_{xx}(k) = \frac{1}{N}\mathrm{E}\{|X(k)|^2\} = \frac{1}{N}\mathrm{E}\{\sum_{m=0}^{N-1}\sum_{n=0}^{N-1} x(m)x^*(n)e^{-j2\pi k(m-n)/N}\}$$

$$= \frac{1}{N}\sum_{m=0}^{N-1}\sum_{n=0}^{N-1}\mathrm{E}\{x(m)x^*(n)\}e^{-j2\pi k(m-n)/N} = \frac{1}{N}\sum_{m=0}^{N-1}\sum_{n=0}^{N-1} r_{xx}(m-n)e^{-j2\pi k(m-n)/N}$$

$$= \frac{1}{N}\sum_{m=0}^{N-1}\sum_{i=m}^{N-1+m} r_{xx}(i)e^{-j2\pi ki/N} = \sum_{i=0}^{N-1} r_{xx}(i)e^{-j2\pi ki/N} = S_{xx}(k).$$

Since the signal, $x(n)$, is periodically extended, the autocorrelation, $r_{xx}(n)$, is periodically extended as well. This means that $r_{xx}(N) = \mathrm{E}\{x(m+N)x^*(m)\} = \mathrm{E}\{x(m)x^*(m)\} = r_{xx}(0)$, $r_{xx}(N+1) = \mathrm{E}\{x(m+N+1)x^*(m)\} = \mathrm{E}\{x(m+1)x^*(m)\} = r_{xx}(1)$, and so on, producing the last equality in the previous derivation.

The power spectrum matrix

$$\mathbf{P}_x = \frac{1}{N}\mathrm{E}\{\mathbf{X}\mathbf{X}^H\}$$

with elements

$$P_{xx}(k,l) = \frac{1}{N}\mathrm{E}\{X(k)X^*(l)\},$$

is a diagonal matrix for the WSS signal $x(n)$, with the elements on the diagonal equal to $S_{xx}(k)$. To show this property of the WSS stationary signals consider

$$P_{xx}(k,l) = \frac{1}{N}\mathrm{E}\{X(k)X^*(l)\} = \frac{1}{N}\mathrm{E}\{\sum_{m=0}^{N-1}\sum_{n=0}^{N-1} x(m)x^*(n)e^{-j2\pi(km-ln)/N}\}$$

$$= \frac{1}{N}\sum_{m=0}^{N-1}\sum_{n=0}^{N-1} r_{xx}(m-n)e^{-j2\pi(km-ln)/N} = \frac{1}{N}\sum_{m=0}^{N-1}\sum_{n=0}^{N-1} r_{xx}(m-n)e^{-j2\pi k(m-n)/N}e^{j2\pi(l-k)n/N}$$

$$= \frac{1}{N}\sum_{n=0}^{N-1}\sum_{i=n}^{N-1+n} r_{xx}(i)e^{-j2\pi ki/N}e^{j2\pi(l-k)n/N} = \sum_{i=0}^{N-1} r_{xx}(i)e^{-j2\pi ki/N}\delta(l-k) = S_{xx}(k)\delta(k-l).$$

This means that we can write

$$\mathbf{P}_x = \frac{1}{N}\mathrm{E}\{\mathbf{X}\mathbf{X}^H\} = \frac{1}{N}\mathrm{E}\{(\mathbf{W}\mathbf{x})(\mathbf{W}\mathbf{x})^H\} = \frac{1}{N}\mathbf{W}\mathrm{E}\{\mathbf{x}\mathbf{x}^H\}\mathbf{W}^H = \frac{1}{N}\mathbf{W}\mathbf{R}_x\mathbf{W}^H.$$

The autocorrelation matrix, $\mathbf{R}_x$, of a WSS signal is diagonalizable, since there is a diagonal matrix (the power spectrum matrix), $\mathbf{P}_x$, such that

$$\mathbf{R}_x = N\mathbf{W}^H\mathbf{P}_x\mathbf{W}.$$

Example 7.21. Consider the random signal

$$x(n) = \varepsilon(n) + 0.7\varepsilon(n-1) + 0.5\varepsilon(n-1) + 0.7\varepsilon(n-3) + 0.9\varepsilon(n-4),$$

where the WSS random signal $\varepsilon(n)$, $n = 0,1,2,\ldots,N-1$, $N = 16$, with $\mathrm{E}\{\varepsilon(n)\} = 0$ and $r_{\varepsilon\varepsilon}(n) = \delta(n)$, is periodically extended in such a way that $\varepsilon(n+Nk) = \varepsilon(n)$, where k is an integer. Find the autocorrelation $r_{xx}(n)$ within the basic period, $n = 0,1,2,\ldots,N-1$, and the power spectral density $S_{xx}(k)$. Use 10000 realizations of $\varepsilon(n)$ to calculate $X_i(k) = \mathrm{DFT}_N\{x_i(n)\}$ and plot $\mathrm{mean}_i\{X_i(k)X_i^*(l)\}/N$, for $k = 0,1,2,\ldots,N-1$ and $l = 0,1,2,\ldots,N-1$.

★ The elements of the autocorrelation function,

$$r_{xx}(n) = \mathrm{E}\{x(m+n)x^*(m)\},$$

are obtained using $r_{\varepsilon\varepsilon}(n) = \delta(n)$ as

$$\begin{aligned}
r_{xx}(0) &= 1 + 0.7^2 + 0.5^2 + 0.7^2 + 0.9^2\\
r_{xx}(\pm 1) &= 0.7 + 0.5\cdot 0.7 + 0.7\cdot 0.5 + 0.9\cdot 0.7\\
r_{xx}(\pm 2) &= 0.5 + 0.7\cdot 0.7 + 0.9\cdot 0.5\\
r_{xx}(\pm 3) &= 0.7 + 0.9\cdot 0.7\\
r_{xx}(\pm 4) &= 0.9,
\end{aligned}$$

$r_{xx}(\pm n) = 0$, for $4 < n < 16 - 4$ and $r_{xx}(n + 16k) = r_{xx}(n)$. The exact value of $r_{xx}(n)$ and its estimation using 10000 realizations

$$\hat{r}_{xx}(n) = \frac{1}{10000N} \sum_{i=1}^{10000} \sum_{m=0}^{N-1} x_i(n+m)x_i(m), \quad \text{with } x(N+n) = x(n),$$

are shown in Fig. 7.13(c). The value of $\text{mean}_i\{X_i(k)X_i^*(l)\}$, averaged over 10000 realizations, is given in Fig. 7.13(a). For comparison, the DFT of $r_{xx}(n)$ is presented on the diagonal of Fig. 7.13(b), with its exact value $\text{mean}\{X(k)X^*(k)\}$ shown in Fig. 7.13(d).

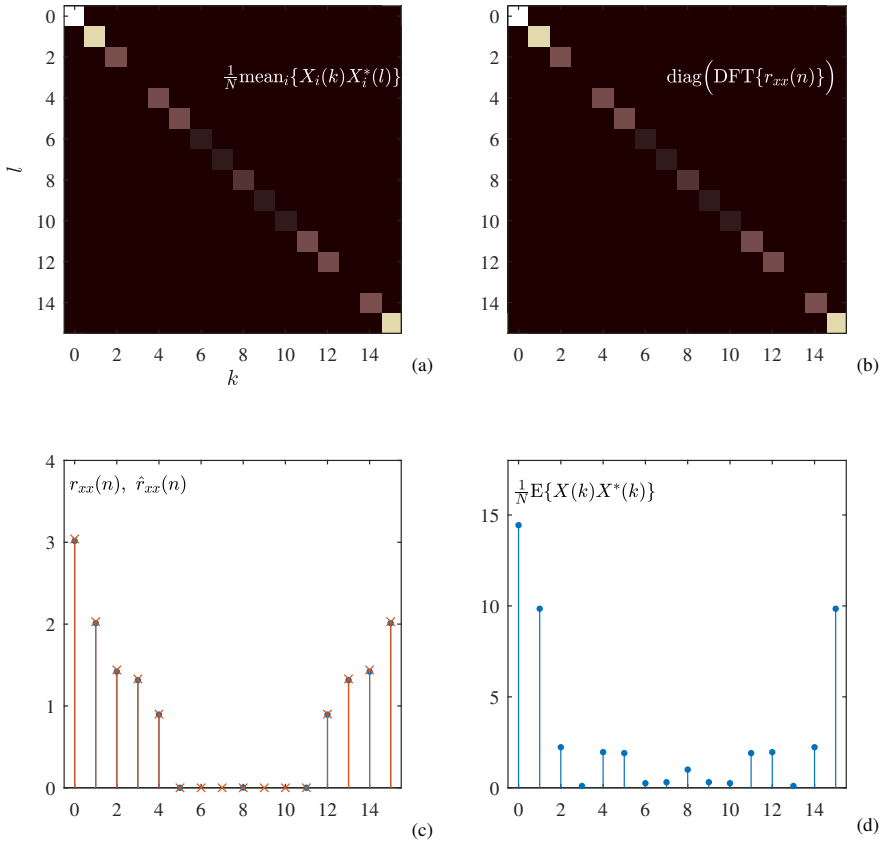


Figure 7.13 The power spectral matrix illustration. (a) The value of $\text{mean}_i\{X_i(k)X_i^*(l)\}/N$ averaged over 10000 realizations. (b) The DFT of $r_{xx}(n)$ as a diagonal matrix. (c) The exact value of $r_{xx}(n)$ and its estimation, $\hat{r}_{xx}(n)$, using 10000 realizations. (d) The exact value of $\text{E}\{X(k)X^*(k)\}/N$ from (b).

7.4 NOISE AND RANDOM SIGNAL EXAMPLES

In many applications, the desired signal is disturbed by various forms of random signals, caused by numerous factors in 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 random, commonly multiple sources, disturbances.

A noise is said to be *white* if its values are uncorrelated

$$r_{\varepsilon\varepsilon}(n,m) = \sigma_\varepsilon^2 \delta(n-m) \tag{7.35}$$

Spectral density of this kind of noise is constant (like it is the case in the white light),

$$S_{\varepsilon\varepsilon}(e^{j\omega}) = \mathrm{FT}\{r_{\varepsilon\varepsilon}(n)\} = \sigma_\varepsilon^2.$$

If this property is not satisfied, then the power spectral density is not constant. Such a noise is referred to as **colored**.

Regarding the distribution of noise $\varepsilon(n)$ amplitudes the most common types of noise in signal processing are: uniform, binary, Gaussian, Rayleigh, Laplacian, Cauchy, and Poison noise.

7.4.1 Uniform Noise

The uniform noise is a discrete-time signal with the probability density function defined by

$$p_{\varepsilon(n)}(\xi) = \frac{1}{\Delta}, \quad \text{for } -\Delta/2 \le \xi < \Delta/2 \tag{7.36}$$

and $p_{\varepsilon(n)}(\xi) = 0$ elsewhere, Fig. 7.14. This noise takes values within the interval $[-\Delta/2, \Delta/2)$ with equal probability. The variance of uniform noise is

$$\sigma_\varepsilon^2 = \int_{-\Delta/2}^{\Delta/2} \xi^2 p_{\varepsilon(n)}(\xi) d\xi = \frac{\Delta^2}{12}.$$

This kind of noise is used to model rounding errors in the amplitude quantization of discrete-time signals. Its distribution indicates that all errors within $-\Delta/2 \le \xi < \Delta/2$ are equally probable.

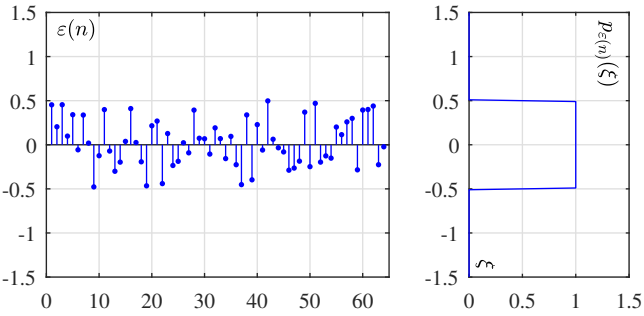


Figure 7.14 A realization of the uniform noise (left) with its probability density function (right), when $\Delta = 1$.

7.4.2 Binary, Bernoulli, and Binomial Random Signal

Random *binary* sequence, or binary noise, is a stochastic signal which randomly takes one of the two fixed signal values. Assume that the noise $\varepsilon(n)$ values are, for example, $\{-1,1\}$ and that the probability that $\varepsilon(n)$ takes value 1 is $P_\varepsilon(1)=p$, while $P_\varepsilon(-1)=1-p$. The expected value 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

$$\sigma_\varepsilon^2=\sum_{\xi=-1,1}(\xi-\mu_\varepsilon)^2P_\varepsilon(\xi)=4p(1-p).$$

A special case is obtained when the values from the set $\{-1,1\}$ are equally probable, that is, when $p=1/2$. Then, we get $\mu_\varepsilon=0$ and $\sigma_\varepsilon^2=1$.

When the random signal $\varepsilon(n)$ values are from the set $\{0,1\}$, then this form of binary signal is referred to as the *Bernoulli random signal* or Bernoulli noise. This signal takes the value $\varepsilon(n)=1$ with the probability p, while the probability of $\varepsilon(n)=0$ is equal to $(1-p)$. The probability that the signal sample $\varepsilon(n)$ takes one specific value, can be written as

$$P(\varepsilon(n)|p)=p^{\varepsilon(n)}(1-p)^{1-\varepsilon(n)},$$

since $P(\varepsilon(n)=1|p)=p$ and $P(\varepsilon(n)=0|p)=1-p$. The expected value of the Bernoulli noise is $\mu_\varepsilon=p$, while the variance is

$$\sigma_\varepsilon^2=\sum_{\xi=0,1}(\xi-\mu_\varepsilon)^2P_\varepsilon(\xi)=(0-p)^2(1-p)+(1-p)^2p=p(1-p).$$

Example 7.22. Consider a set of $N\to\infty$ balls. An equal number of balls is marked with 1 (or white) and 0 (or black). A random signal $x(n)$, $n=0,1,2,3$, corresponds to drawing of four balls in a row. This signal has four samples $x(0)$, $x(1)$, $x(2)$, and $x(3)$. The 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, $x_m(n)$, with the number k being equal to the number of appearances of digit 1 in every signal realization, are given in the next table.

$x_m(0)$	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
$x_m(1)$	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
$x_m(2)$	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
$x_m(3)$	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

$$y(m)=\sum_{n=0}^{3}x_m(n)=k \quad 0\ 1\ 1\ 2\ 1\ 2\ 2\ 3\ 1\ 2\ 2\ 3\ 2\ 3\ 3\ 4$$

Possible values of k are $0,1,2,3,4$ with the corresponding probabilities

$$P(0)=1\cdot\left(\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}\right),\quad P(1)=4\cdot\left(\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}\right)\tfrac{1}{2},\quad P(2)=6\cdot\left(\tfrac{1}{2}\tfrac{1}{2}\right)\tfrac{1}{2}\tfrac{1}{2},$$
$$P(3)=4\cdot\left(\tfrac{1}{2}\right)\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2},\ \text{ and }\quad P(4)=1\cdot\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}.$$

These probabilities can be considered as the terms of the binomial expression

$$(p+q)^4 = \binom{4}{0}p^4 + \binom{4}{1}p^3q + \binom{4}{2}p^2q^2 + \binom{4}{3}pq^3 + \binom{4}{4}q^4$$

with $p = 1/2$ and $q = 1-p = 1/2$. For the case when N is a finite number see Problem 7.9.

An interesting form of the random variable, resulting from the experiment when the random varaible can take only two possible values $\{-1,1\}$ or $\{0,1\}$ or $\{\text{No},\text{Yes}\}$ or $\{A,B\}$, is the **binomial random variable**. Binomial random variable is equal to the number, k, of successes (1, or Yes or B) in a sequence of N independent binary experiments, each of which yields success with probability p. This random varaible obeys the binomial distribution which is the basis for the popular binomial test of statistical significance.

The binomial random variable, k, has been introduced through the previous simple example. In general, if the signal $x(n)$ takes the value B from the set $\{A,B\}$ with the probability p, then the probability that there is exactly k values of B in a specific order, within a sequence of N samples of $x(n)$, is $p^k(1-p)^{N-k}$. For $k=1$, it is possible to achieve this result in N specific orders (combinations, see Example 7.22). When $k=2$, then there are $N(N-1) = \binom{N}{2}$ such combinations. In general, for any k, there are $\binom{N}{k}$ orders (combinations) that $x(n)$ takes k times value B, that is

$$P(k) = \binom{N}{k}p^k(1-p)^{N-k} = \frac{N!}{k!(N-k)!}p^k(1-p)^{N-k}. \tag{7.37}$$

This is a binomial coefficients form of $(p+q)^N = (p+(1-p))^N$. The expected value of the number of appearances $y(m) = k$ of the event B or "Yes" in N samples is

$$\mu_y = \mathrm{E}\{y\} = \sum_{k=0}^{N} kP(k) = \sum_{k=0}^{N} k\frac{N!}{k!(N-k)!}p^k(1-p)^{N-k}.$$

Since the first term in the summation is 0, we will shift the summation for one and reindex it to

$$\begin{aligned}\mu_y = \mathrm{E}\{y\} &= \sum_{k=0}^{N-1}(k+1)\frac{N(N-1)!}{(k+1)!((N-(k+1))!}p^{k+1}(1-p)^{N-(k+1)} \\ &= Np\sum_{k=0}^{N-1}\frac{(N-1)!}{k!((N-1)-k)!}p^k(1-p)^{(N-1)-k}.\end{aligned}$$

The sum in the last expression is equal to 1 since

$$1 = (p+(1-p))^{N-1} = \sum_{k=0}^{N-1}\binom{N-1}{k}p^k(1-p)^{(N-1)-k} = \sum_{k=0}^{N-1}\frac{(N-1)!}{k!(N-1-k)!}p^k(1-p)^{N-1-k}$$

resulting, with $p+(1-p)=1$, into

$$\mu_y = \mathrm{E}\{y\} = Np.$$

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 $\mathrm{E}\{y\} = Np$. 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 $y-1$,

$$\mathrm{E}\{y(y-1)\} = \sum_{k=0}^{N} k(k-1)P(k) = \sum_{k=0}^{N} k(k-1)\frac{N!}{k!(N-k)!}p^k(1-p)^{N-k}.$$

Since the first two terms are 0, we can reindex the summation as

$$\begin{aligned}\mathrm{E}\{y(y-1)\} &= \sum_{k=0}^{N-2}(k+2)(k+1)\frac{N!}{(k+2)!(N-2-k)!}p^{k+2}(1-p)^{N-2-k} \\ &= N(N-1)p^2\sum_{k=0}^{N-2}\frac{(N-2)!}{k!(N-2-k)!}p^k(1-p)^{N-2-k}.\end{aligned}$$

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

$$\mathrm{E}\{y(y-1)\} = N(N-1)p^2.$$

The variance of y follows from

$$\sigma_y^2 = \mathrm{E}\{y^2\} - (\mathrm{E}\{y\})^2 = \mathrm{E}\{y(y-1)\} + \mathrm{E}\{y\} - (\mathrm{E}\{y\})^2 = Np(1-p).$$

Therefore, in a sequence of N values of signal $x(n)$ that can take values $\{0,1\}$, the expected value of the number of appearances of 1, $y = \sum_{n=1}^{N} x(n)$, divided by N, is

$$\mathrm{E}\{z\} = \mathrm{E}\{\frac{1}{N}\sum_{n=1}^{N} x(n)\} = \mathrm{E}\{\frac{y}{N}\} = \frac{Np}{N} = p. \tag{7.38}$$

The variance of the normalized number of appearances of the value 1 is

$$\sigma_z^2 = \frac{1}{N^2}\sigma_y^2 = \frac{Np(1-p)}{N^2} = \frac{p(1-p)}{N}$$

By 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 (see Example 7.33).

Notice that the random variable z is the mean of the independent random variables $x(n)$.

7.4.3 Bayesian Inference for Binary and Binomial Random Signal

Here we will reconsider Bayes' theorem and use Bayesian analysis to test hypotheses about the probabilistic model parameters in the case of binary random signals. *Bayesian inference* is a method in which Bayes' theorem is used to update the probability for a hypothesis as more evidence samples (events) become available.

Assume that the random signal $x(n)$ takes the values $+1$ and 0 (positive or negative test result; 'head' or 'tail' in the coin-tossing; value 1 or -1) with probabilities p and $1-p$, respectively, that are not known.

The event B consists of N observed samples $x(n)$, $n = 1,2,\ldots,N$, and the number k of $x(n) = 1$ occurrences in this sequence. The aim is to estimate the probability p based on the results obtained in the observed event B (that is, based on k and N).

The classical frequentist approach to the estimation of the probability p is based on (7.38) and

$$\hat{p} = \frac{1}{N}\sum_{n=1}^{N} x(n) = \frac{k}{N}$$

assuming that $x(n)$ takes the value of 1 with the probability p and the value of 0 with the probability $q = 1 - p$. This problem is elaborated in detail in Example 7.33.

Here, we will consider the problem of the probability p estimation within the Bayes framework. In this case, Bayes' relation can be written in the form

$$P(p|B) = \frac{P(B|p)P(p)}{P(B)}, \tag{7.39}$$

where the hypothesis is that the probability p takes a particular value from the given set of all possible values and that the event B occurred assuming this probability. The terms in this expression are:

- *Prior* $P(p)$ for the hypothesis p. It has to be assumed based on our possible knowledge about the resulting p:
  1. If all values of p are equally probable, then the uniform prior is assumed, $P(p) = C$, for a discrete set of possible values
     $$p \in \{0,\ \Delta p,\ 2\Delta p, \ldots, 1\}$$
     with the step, for example, $\Delta p = 0.01$ and C is a constant (as it will be shown, not relevant).
  2. If we expect that the value of p is close to 0.5, then we can assume, for example, the prior $P(p) = 2C(1 - 2|p - 0.5|)$ for $0 \leq p \leq 1$, with the step Δp for p, or
  3. The Gaussian function prior form
     $$P(p) = C\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(p-0.5)^2}{2\sigma^2}}, \quad \text{calculated at the discrete values of } p, \text{ with the step } \Delta p.$$
- The *likelihood* factor $P(B|p)$ is equal to the probability that the event B occurred for the assumed value of p. For the random binary signal, the event B denotes a realization which consists of k samples $x(n) = 1$ and $N - k$ samples when $x(n) = 0$. The probability of this event B, for the assumed p, is
  $$P(B|p) = \binom{N}{k} p^k (1-p)^{N-k}. \tag{7.40}$$
- The probability of observing the data specified by the event B (k times $x(n) = 1$ and $N - k$ times $x(n) = 0$), summed over all hypotheses (all possible values of p) is
  $$P(B) = \sum_{p} \binom{N}{k} p^k (1-p)^{N-k} P(p).$$

It is common to avoid the last (*marginal*) probability, $P(B)$, in (7.39) which is not dependent on p (plays the role of a normalization factor), and to consider only the so called *posterior*

$$P(p|B) \propto P(B|p)P(p).$$

In order to give probabilistic interpretation, the values $P(p|B)$ can be normalized so that

$$\sum_{p} P(p|B) = 1$$

for all considered discrete values of p.

Example 7.23. Here, will calculate the posterior of the hypothesis p, $P(p|B) \propto P(B|p)P(p)$, for the binary signal $x(n)$ when the evidence B consists of N signal samples, with $x(n) = 1$ appearing k times. The posterior $P(p|B)$ is updated as N increases. The following events are analyzed:

(a) The event B of $N = 6$ samples $x(n)$, with $k = 2$ samples taking the value $x(n) = 1$.

(b) The event B when the number of available samples (observations) is increased to $N = 50$ and $k = 9$ times $x(n) = 1$ is obtained.

(c) The event B with a large number of available samples, $N = 1000$, when $k = 220$ nonzero samples, $x(n) = 1$, are observed.

For the hypothesis p use:
(i) The uniform prior $P(p) = C$ and
(ii) the Gaussian prior $P(p) = C\exp(-(p-0.5)^2/(2\sigma^2))/(\sigma\sqrt{2\pi})$, with $\sigma = 0.05$,
and the set of values $0 \leq p \leq 1$ with the step $\Delta p = 0.01$. The value of constant C is not relevant, since the results are normalized.

★ The results for the posterior, $P(p|B) \propto P(B|p)P(p)$, with $P(B|p)$ defined in (7.40), are shown in Fig. 7.15, for the uniform prior, $P(p)$ (left) and foe the Gaussian prior $P(p)$ (right), for various p and given N and k in (a), (b), and (c), respectively.

We can see that the hypothesis' probability is influenced by the prior distribution. When the evidence is large (with large N, as in (c)) both cases produce highly concentrated probability descriptions (denoted by (c)) close to the expected value of the parameter p.

For the probabilistic interpretation, the values of $P(p|B)$ in Fig. 7.15 should be normalized so that $\sum_p P(p|B) = 1$ for every considered case (bottom panels).

The maximum position of the likelihood factor $P(B|p)$ (*maximum likelihood estimation*) can ne found in an analytic way. From

$$\frac{dP(B|p)}{dp} = 0$$

follows $k(1-p) = (N-k)p$ or $p = k/N$. This solution holds for the maximization of the posterior $P(p|B)$ for the uniform prior $P(p)$. However, for other prior functions, the maximum of the likelihood factor depends on the prior function, especially for low N. The influence of the prior will be analyzed next.

Log-Likelihood Function. A specific form of the binary random signal, called the Bernoulli random signal, will be used to introduce few more important concepts in Bayesian analysis.

For the Bernoulli random signal, the probability that the signal sample $x(n)$ takes one specific value, with the assumed parameter value p (hypothesis), can be written in a compact form as

$$P(x(n)|p) = p^{x(n)}(1-p)^{1-x(n)},$$

since $P(x(n) = 1|p) = p$ and $P(x(n) = 0|p) = 1 - p$.

For N statistically independent samples, given in the vector $\mathbf{x} = [x(1), x(2), \ldots, x(N)]^T$, the likelihood function is the joint probability of N random samples $x(n)$, and it is equal to the product of

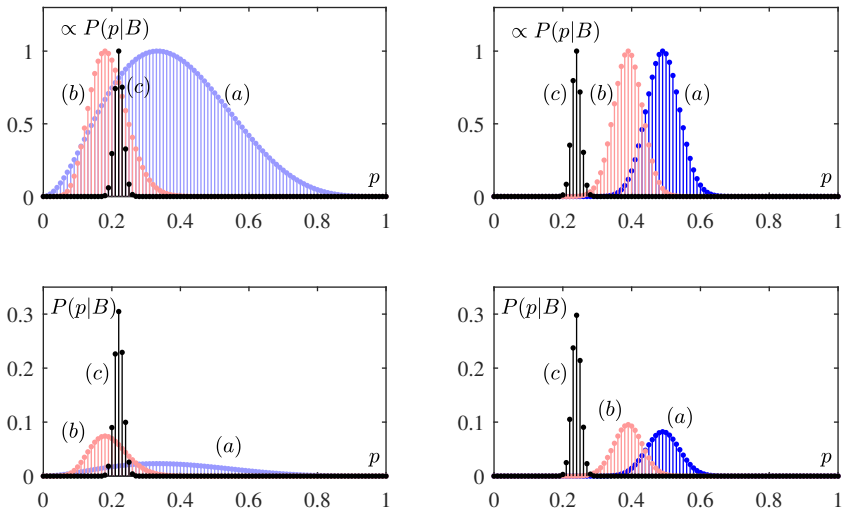


Figure 7.15 Bayesian approach based estimation of the probability p that a nonzero sample, $x(n) = 1$, is obtained in the binary random signal for the different number of available samples (realizations) N and the number of nonzero samples k: (a) $N = 6$, $k = 2$, (b) $N = 50$, $k = 9$, and (c) $N = 1000$, $k = 220$. The uniform prior $P(p)$ is used for the left panels and the Gaussian prior centered at $p = 0.5$ for the right panels. The step in p was $\Delta p = 0.01$. All shown probabilities are normalized to one (top panels). For the probabilistic interpretation, the values of $P(p|B)$ should be normalized so that $\sum_p P(p|B) = 1$ for each of the considered case (bottom panels).

their individual probabilities, that is

$$P(\mathbf{x}|p) = \prod_{n=1}^{N} p^{x(n)}(1-p)^{1-x(n)} = p^{\sum_{n=1}^{N} x(n)}(1-p)^{\sum_{n=1}^{N}(1-x(n))}.$$

The aim is to find the maximum of $P(\mathbf{x}|p)$. We may use the derivative of $P(\mathbf{x}|p)$, or the derivative of its logarithm, since the logarithm is a monotonous function and will not change the maximum position for the positive considered function. In general, we have

$$\frac{d\ln(f(x))}{dx} = \frac{f'(x)}{f(x)},$$

meaning that both $f(x)$ and $\ln(f(x))$ will produce the same result for the maximum position, $f'(x) = 0$. The *log-likelihood function* is of the form

$$\ln(P(\mathbf{x}|p)) = \Big(\sum_{n=1}^{N} x(n)\Big)\ln(p) + \Big(\sum_{n=1}^{N}(1-x(n))\Big)\ln(1-p).$$

The maximum of this function (*maximum likelihood estimation* (MLE) of the model parameter p) is obtained from

$$p_{MLE} = \arg\{\max\{\ln(P(\mathbf{x}|p))\}\}.$$

It follows by making the derivative of $\ln(P(B|p))$ with regard of p equal to 0,

$$\frac{d(\ln(P(\mathbf{x}|p)))}{dp} = 0.$$

After differentiation of $\ln(P(\mathbf{x}|p))$ over p, from

$$\frac{\sum_{n=1}^{N} x(n)}{p} - \frac{\sum_{n=1}^{N}(1-x(n))}{1-p} = 0$$

we get

$$p_{MLE} = \frac{\sum_{n=1}^{N} x(n)}{N} = \frac{k}{N}$$

if exactly k signal samples in $\mathbf{x} = (x(1), x(2), \ldots, x(N))$ take the value $x(n) = 1$. This event is denoted by B in the previous example, with the probability $P(B|p) = P(\mathbf{x}|p)$.

Frequentist versus Bayesian inference. In the previous analysis, we get a specific value for the hypothesis parameter p (this is the so-called *frequentist inference*). It does not give the probability of the hypothesis parameter p, as it was the case in Fig. 7.15. The Bayes analysis, presented in Fig. 7.15, is based on the posterior $P(p|B) \propto P(B|p)P(p)$ which results not in a single value of the parameter p, but in its probabilistic description (*Bayesian inference*). This probabilistic description includes our prior belief $P(p)$ and the current experiment outcome (evidence), $P(B|p)$.

The *maximum a posterior* (MAP) solution is obtained as the position of the maximum of the logarithm of the a posterior probability $P(p|B) \propto P(B|p)P(p)$, that is

$$p_{MAP} = \arg\{\max\{\ln(P(B|p)P(p))\}\} = \arg\{\max\{\ln(P(B|p)) + \ln(P(p))\}\}.$$

Note that the result of the maximum likelihood solution p_{MLE} is the special case of the maximum a posterior solution p_{MAP}, when the prior $P(p)$ is uniform.

In many optimization approaches, the negative value of the logarithm is used as a cost function. Then, instead of the maximum position, the position of the minimum is the problem solution

$$p_{MAP} = \arg\{\min\{-\ln(P(B|p)P(p))\}\}.$$

In solving the minimization problem, various gradient-based algorithms can also be used.

7.4.4 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 distribution.

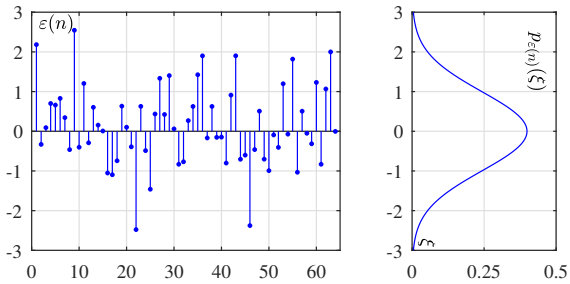


Figure 7.16 A realization of Gaussian noise (left) with its probability density function (right).

The Gaussian zero-mean noise has the probability density function

$$p_{\varepsilon(n)}(\xi) = \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} e^{-\xi^2/(2\sigma_\varepsilon^2)}. \tag{7.41}$$

The variance of this noise is σ_ε^2 (see Problem 7.10). For a Gaussian distributed random signal with the mean value μ and the variance σ_ε^2, whose probability density function is

$$p_{\varepsilon(n)}(\xi) = p_{\varepsilon(n)}(\xi|\mu,\sigma_\varepsilon) = \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} e^{-(\xi-\mu)^2/(2\sigma_\varepsilon^2)}, \tag{7.42}$$

we can use the notation $\mathcal{N}(\mu,\sigma_\varepsilon^2)$.

Example 7.24. Consider N random signals (variables), $x_i(n)$, $i = 1,2,\ldots,N$, that are independent, and identically distributed (i.i.d.), with the unit variance and zero-mean. The probability density functions of the random signals $x_i(n)$ are the same and equal to $p_x(\xi)$. A new random signal $y(n)$ is formed as a sum

$$y(n) = \frac{1}{\sqrt{N}}x_1(n) + \frac{1}{\sqrt{N}}x_2(n) + \cdots + \frac{1}{\sqrt{N}}x_N(n).$$

The factors $1/\sqrt{N}$ are introduced so that the variance of $y(n)$ is $\sigma_{y(n)}^2 = 1$.

Find the probability density function of the random signal $y(n)$ when $N \to \infty$.

★ The probability density function of the random signal $x_i(n)/\sqrt{N}$ is $p_x(\xi\sqrt{N})\sqrt{N}$. The characteristic functions of $x_i(n)$ and $x_i(n)/\sqrt{N}$ are, respectively, (7.29)

$$\Phi_x(\theta) = 1 + j\theta M_1 - \frac{1}{2!}\theta^2 M_2 - j\frac{1}{3!}\theta^3 M_3 + \ldots = 1 - \frac{1}{2!}\theta^2 - j\frac{1}{3!}\theta^3 M_3 + \ldots,$$

$$\Phi_{\frac{x}{\sqrt{N}}}(\theta) = 1 + j\frac{\theta}{\sqrt{N}}M_1 - \frac{1}{2!}\frac{\theta^2}{N}M_2 - j\frac{1}{3!}\frac{\theta^3}{N^{3/2}}M_3 + \cdots = 1 - \frac{1}{2!}\frac{\theta^2}{N} - j\frac{1}{3!}\frac{\theta^3}{N^{3/2}}M_3 + \ldots,$$

since $M_1 = 0$ (zero-mean) and $M_2 = 1$ (unit variance).

The random signal $y(n)$ is obtained as a sum of signals $x_i(n)/\sqrt{N}$, $i = 1,2,\ldots,N$, and its characteristic function is equal to the product of the individual characteristic functions, (7.30)

$$\Phi_y(\theta) = \Phi_{x/\sqrt{N}}^N(\theta) = \left(1 - \frac{1}{2!}\frac{\theta^2}{N} - j\frac{1}{3!}\frac{\theta^3}{N^{3/2}}M_3 + \ldots\right)^N.$$

For large N, we can write

$$\lim_{N\to\infty}\Phi_y(\theta) = \lim_{N\to\infty}\left(1 - \frac{1}{2!}\frac{\theta^2}{N} - j\frac{1}{3!}\frac{\theta^3}{N^{3/2}}M_3 + \ldots\right)^N = \lim_{N\to\infty}\left(1 - \frac{\theta^2}{2N}\right)^N = e^{-\theta^2/2},$$

since

$$\lim_{N\to\infty}(1 - x/N)^N = e^{-x}.$$

The inverse Fourier transform (with sign $-$) of $e^{-\theta^2/2}$ is equal to the probability density function of the unit-variance Gaussian random variable,

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\theta^2/2}e^{-j\theta\xi}d\xi = \frac{1}{\sqrt{2\pi}}e^{-\xi^2/2},$$

what proves the *central limit theorem* (CLT) for the sum of independent, and identically distributed (i.i.d.) random variables.

The probability that the amplitude of a zero-mean Gaussian random variable takes a value smaller than λ is

$$\text{Probability}\{|\varepsilon(n)| < \lambda\} = \frac{1}{\sigma_\varepsilon\sqrt{2\pi}}\int_{-\lambda}^{\lambda} e^{-\xi^2/(2\sigma_\varepsilon^2)}d\xi = \text{erf}\left(\frac{\lambda}{\sqrt{2}\sigma_\varepsilon}\right) \tag{7.43}$$

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 Gaussian random variable is within the standard deviation, two standard deviations (two-sigma rule), or three standard deviations are

$$\begin{aligned}
\text{Probability}\{-\sigma_\varepsilon < \varepsilon(n) < \sigma_\varepsilon\} &= \text{erf}(1/\sqrt{2}) = 0.6827, \\
\text{Probability}\{-2\sigma_\varepsilon < \varepsilon(n) < 2\sigma_\varepsilon\} &= \text{erf}(\sqrt{2}) = 0.9545, \\
\text{Probability}\{-3\sigma_\varepsilon < \varepsilon(n) < 3\sigma_\varepsilon\} &= \text{erf}(3/\sqrt{2}) = 0.9973.
\end{aligned} \tag{7.44}$$

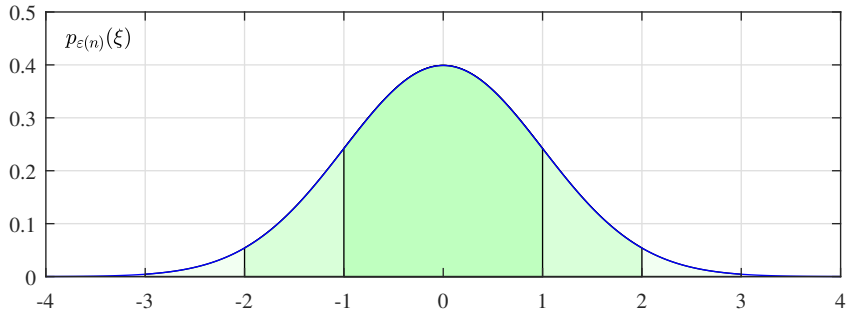


Figure 7.17 Probability density function with the intervals corresponding to $-\sigma_\varepsilon < \varepsilon(n) < \sigma_\varepsilon$, $-2\sigma_\varepsilon < \varepsilon(n) < 2\sigma_\varepsilon$, and $-3\sigma_\varepsilon < \varepsilon(n) < 3\sigma_\varepsilon$. Value of $\sigma_\varepsilon = 1$ is used.

Example 7.25. Given 12 measurements of a Gaussian zero-mean noise $\varepsilon(n) \in \{-0.7519, 1.5163, -0.0326, -0.4251, 0.5894, -0.0628, -2.0220, -0.9821, 0.6125, -0.0549, -1.1187, 1.6360\}$. Estimate the sample standard deviation of this data and use it to 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 using (7.7) with $\mu = 0$ and $M = 12$ (see also Section 7.4.5). Its values is $\sigma = 1.031$. Thus, the absolute value of this noise will be smaller than 2.5 with the probability of

$$P\{|\varepsilon(n)| < 2.5\} = \frac{1}{1.031\sqrt{2\pi}} \int_{-2.5}^{2.5} e^{-\xi^2/(2\cdot 1.031^2)} d\xi = \text{erf}(2.5/(\sqrt{2}\cdot 1.031)) = 0.9847.$$

Example 7.26. The random signal $x(n)$ is a Gaussian noise with the mean value $\mu_x = 1$ and the variance $\sigma_x^2 = 1$. The random sequence $y(n)$ is obtained by omitting samples from the signal $x(n)$ that are either negative or higher than 1. Find the probability density function of sequence $y(n)$. Find its mean and variance, μ_y and σ_y.

★The probability density function for the sequence $y(n)$ is

$$p_{y(n)}(\zeta) = \begin{cases} B\frac{1}{\sqrt{2\pi}} e^{-\frac{(\xi-1)^2}{2}} & \text{for } 0 < \xi \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The constant B can be calculated from $\int_{-\infty}^{\infty} p_{y(n)}(\xi)d\xi = 1$, resulting in $B = 2/\text{erf}(\frac{1}{\sqrt{2}})$. Now we have

$$\mu_y = \int_0^1 \xi \frac{2}{\text{erf}(\frac{1}{\sqrt{2}})} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\xi-1)^2}{2}} d\xi = 1 - \frac{\sqrt{2}(1-e^{-1/2})}{\sqrt{\pi}\,\text{erf}(\frac{1}{\sqrt{2}})} \approx 0.54$$

$$\sigma_y^2 = \int_0^1 (\xi - \mu_{y(n)})^2 \frac{2}{\text{erf}(\sqrt{2})} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\xi-1)^2}{2}} d\xi \approx 0.08.$$

7.4.5 Estimation of the Gaussian Distribution Parameters

Estimation of the Gaussian distribution parameters based on the observed set of the signal values will be presented in this section, using the maximum likelihood approach.

Stationary signal. Consider a stationary random Gaussian distributed signal $x(n)$ whose N samples are available. The probability density function of a signal sample $x(n)$ is defined by

$$p_{x(n)}(\xi|\sigma,\mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\xi-\mu)^2}{2\sigma^2}\right),$$

where σ and μ are the assumed (unknown) parameters of the Gaussian distribution.

A stationary random signal with N independent samples may be considered as an N-dimensional random variable, with the joint probability density function

$$p_{x(1),\ldots,x(n)}(\xi_1,\ldots,\xi_N|\sigma,\mu) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(\xi_1-\mu)^2}{2\sigma^2}) \times \cdots \times \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(\xi_N-\mu)^2}{2\sigma^2}). \tag{7.45}$$

Vector form relation of this probability density function is

$$p_{\mathbf{x}}(\boldsymbol{\xi}|\sigma,\mu) = \frac{1}{\sigma^N\sqrt{(2\pi)^N}}\exp(-\frac{||\boldsymbol{\xi}-\mu||_2^2}{2\sigma^2}),$$

where $\mathbf{x}=[x(1),x(2),\ldots,x(N)]^T$, $\boldsymbol{\xi}=[\xi_1,\xi_2,\ldots,\xi_N]^T$, and $||\boldsymbol{\xi}-\mu||_2^2=(\xi_1-\mu)^2+\cdots+(\xi_N-\mu)^2$.

Within the maximum likelihood estimation (MLE) framework, the goal is to find the unknown (prior) parameters σ and μ so that the distribution fits best the observed data $\mathbf{x} = [x(1),x(2),\ldots,x(N)]^T$. The probability of σ and μ, given the observed random signal samples, $P(\sigma,\mu|\mathbf{x})$, can be written using Bayes' relation for a posterior distribution as in (7.16), (7.39)

$$P(\sigma,\mu|\mathbf{x}) = \frac{P(\mathbf{x}|\sigma,\mu)P(\sigma,\mu)}{P(\mathbf{x})}.$$

Since $P(\mathbf{x})$ does not depend on parameters σ and μ, this (marginal) probability does not influence the optimization with regard to the parameters σ and μ, and is commonly omitted from the analysis. Furthermore, using the uniform priors, $P(\sigma,\mu)=c$, we can write

$$P(\sigma,\mu|\mathbf{x}) \propto P_{\mathbf{x}}(\mathbf{x}|\sigma,\mu).$$

The probability that the random signal $x(n)$ takes specific values, given by

$$\mathbf{x} = [x(1),x(2),\ldots,x(N)]^T,$$

is related to the probability density function as

$$P(\mathbf{x}|\sigma,\mu) = p_{\mathbf{x}}(\mathbf{x}|\sigma,\mu)d\mathbf{x} = \frac{1}{\sigma^N\sqrt{(2\pi)^N}}\exp(-\frac{||\mathbf{x}-\mu||_2^2}{2\sigma^2})d\mathbf{x}.$$

Therefore, the best fitting parameter (σ,μ) values can be obtained by maximizing

$$P(\sigma,\mu|\mathbf{x}) \propto p_{\mathbf{x}}(\mathbf{x}|\sigma,\mu).$$

For the Gaussian distributed random signal, the maximization is performed straightforwardly, by differentiating the probability density (likelihood) function $p(\mathbf{x}|\sigma,\mu)$ or its logarithm (log-likelihood) function.

We will use the negative logarithm function, when the likelihood maximization problem is equivalent to the log-likelihood minimization problem stated as

$$(\sigma,\mu)_{MLE} = \arg\Big\{\min\{-\ln\Big(p_{\mathbf{x}}(\mathbf{x}|\sigma,\mu)\Big)\}\Big\}. \tag{7.46}$$

This means that we have to minimize the cost function $-\ln\Big(p_{\mathbf{x}}(\mathbf{x}|\sigma,\mu)\Big)$, defined by

$$J(\sigma,\mu) = -\ln\Big(p_{\mathbf{x}}(\mathbf{x}|\sigma,\mu)\Big) = \frac{N}{2}\ln(2\pi) + N\ln(\sigma) + \frac{||\mathbf{x}-\mu||_2^2}{2\sigma^2}, \tag{7.47}$$

where $||\mathbf{x}-\mu||_2^2$ is the squared two-norm (L_2-norm) of the vector $\mathbf{x}-\mu$,

$$||\mathbf{x}-\mu||_2^2 = (x(1)-\mu)^2 + (x(2)-\mu)^2 + \cdots + (x(N)-\mu)^2.$$

Using $\partial J(\sigma,\mu)/\partial\mu = 0$, the parameter μ estimate follows from

$$2(x(1)-\mu) + 2(x(2)-\mu) + \cdots + 2(x(N)-\mu) = 0, \quad \text{as}$$
$$\hat{\mu} = \frac{1}{N}(x(0) + x(1) + \cdots + x(N)),$$

while using $\partial J(\sigma,\mu)/\partial\sigma = 0$, an estimate of the parameter σ is obtained from

$$\frac{N}{\sigma} - ||\mathbf{x}-\mu||_2^2 \frac{1}{\sigma^3} = 0, \quad \text{as} \quad \hat{\sigma}^2 = \frac{1}{N}||\mathbf{x}-\hat{\mu}||_2^2.$$

These are the well-known statistical relations for the mean value and the variance introduced intuitively (frequentist inference) in Section 7.1. In Bayesian inference, we should provide $P(\sigma,\mu|\mathbf{x}) \propto p_{\mathbf{x}}(\mathbf{x}|\sigma,\mu)P(\sigma,\mu)$, for an assumed prior $P(\sigma,\mu)$ and a set of possible values for μ and σ, rather than their specific values (as in the next example).

Example 7.27. The concept of finding the parameters μ and σ of the Gaussian distribution, to fit data, is illustrated on a simple data set. Assume that four observations of a Gaussian stationary signal $x(n)$ are available, and given by $x(1) = 0.2$, $x(2) = -0.3$, $x(3) = -0.4$, and $x(4) = 0.5$. Estimate the expected value, μ, and the variance, σ^2, of the Gaussian distribution from the observed data. The data set is then increased to $N = 20$ available samples, whose values are given in Fig. 7.18(right).

Find the posterior distribution $P(\sigma,\mu|\mathbf{x})$ for the discrete sets $-1 \leq \mu \leq 1$ and $0.1 \leq \sigma \leq 1$, with the step 0.05, and the uniform prior $P(\sigma,\mu) = C$.

★ The log-likelihood function of the joint distribution of the observed data is given by (7.47)

$$J(\sigma,\mu) = \frac{4}{2}\ln(2\pi) + 4\ln(\sigma) + \frac{(x(1)-\mu)^2 + (x(2)-\mu)^2 + (x(3)-\mu)^2 + (x(4)-\mu)^2}{2\sigma^2}.$$

Differentiation of this expression with respect to μ, with $\partial J(\sigma,\mu)/\partial\mu = 0$, produces the estimate

$$\hat{\mu} = \frac{1}{4}(x(1) + x(2) + x(3) + x(4)) = 0,$$

while the differentiation of $J(\sigma,\mu)$ with respect to σ results in

$$\frac{4}{\sigma} - \Big((x(1)-\mu)^2 + (x(2)-\mu)^2 + (x(3)-\mu)^2 + (x(4)-\mu)^2\Big)\frac{1}{\sigma^3} = 0$$
$$\hat{\sigma} = \frac{1}{2}\sqrt{0.2^2 + 0.3^2 + 0.4^2 + 0.5^2} = 0.3674.$$

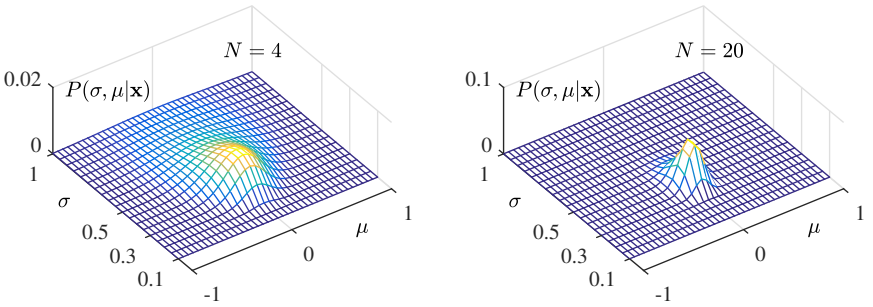
The Bayesian inference approach, for the uniform prior $P(\sigma,\mu) = C$, would produce the probability

$$P(\sigma,\mu|\mathbf{x}) \propto p_{\mathbf{x}}(\mathbf{x}|\sigma,\mu)P(\sigma,\mu) \propto p_{\mathbf{x}}(\mathbf{x}|\sigma,\mu) = \frac{1}{\sigma^N\sqrt{(2\pi)^N}}\exp\left(-\frac{||\mathbf{x}-\mu||_2^2}{2\sigma^2}\right)$$

as shown in Fig. 7.18(left) for given $\mathbf{x} = [2,-3,-4,\ 5]^T/10$ and variable μ and σ.

In order to show the influence of the number of samples on the reliability of the result for μ and σ we have also calculated $P(\sigma,\mu|\mathbf{x}) \propto p_{\mathbf{x}}(\mathbf{x}|\sigma,\mu)$ for $N = 20$ available signal samples and discrete sets $-1 \leq \mu \leq 1$ and $0.1 \leq \sigma \leq 1$, with the step 0.05, Fig. 7.18(right).

Both of these sets of the available samples $x(n)$, with $N = 4$ and $N = 20$, produce almost the same result in the frequentist inference approach, $\mu \approx 0.00$ and $\sigma \approx 0.37$, while their posterior distributions $P(\sigma,\mu|\mathbf{x})$ are quite different.



$\mathbf{x} = [2,-3,-4,\ 5]^T/10,$ $\mathbf{x} = [2,-3,-4,\ 5,\ 8,\ 1,-1,\ 3,\ 0,-7,\ 0,-5,\ 2,\ 4,-1,\ 4,-6,-1,\ 1,-1]^T/10.$

Figure 7.18 Bayesian inference approach based estimation of the parameters μ and σ in the Gaussian random signal for different numbers of the available samples (realizations) N. All shown probabilities $P(\sigma,\mu|\mathbf{x})$ are normalized so that $\sum_\sigma \sum_\mu P(\sigma,\mu|\mathbf{x}) = 1$ holds for considered cases. The uniform prior, $P(\sigma,\mu) = C$, is used.

Example 7.28. A noisy random variable $x(n)$ is a function of M independent variables $t_i(n)$, $i = 1,2,\ldots,M$,

$$x(n) = a_1 t_1(n) + a_2 t_2(n) + \cdots + a_M t_M(n) + \varepsilon(n),\ \ n = 1,2,\ldots,N.$$

The values of $x(n)$, $n = 1,2,\ldots,N$, are obtained with the available $t_i(n)$, $i = 1,2,\ldots,M$, $n = 1,2,\ldots,N$, and the assumed parameters $a_1,a_2,\ldots,a_M$, while $\varepsilon(n)$ is a zero-mean Gaussian noise.

Since $\varepsilon(n)$ is assumed to be a zero-mean Gaussian variable, then as in (7.45), we have

$$p_{\varepsilon(1),\ldots,\varepsilon(n)}(\xi_1,\ldots,\xi_N|\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{\xi_1^2}{2\sigma^2}\right) \times \cdots \times \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{\xi_N^2}{2\sigma^2}\right),$$

Having in mind that

$$\varepsilon(n) = x(n) - a_1 t_1(n) + a_2 t_2(n) + \cdots + a_M t_M(n) = x(n) - \mathbf{t}(n)\mathbf{a},$$

where $\mathbf{a} = [a_1,a_2,\ldots,a_M]^T$ and $\mathbf{t}(n) = [t_1(n),t_2(n),\ldots,t_M(n)]$, we get

$$p_{\mathbf{x}}(\mathbf{x}|\sigma,\mathbf{a},\mathbf{t}(n)) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x(1)-\mathbf{t}(1)\mathbf{a})^2}{2\sigma^2}\right) \times \cdots \times \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x(N)-\mathbf{t}(N)\mathbf{a})}{2\sigma^2}\right).$$

As explained in this section, the best fitting parameters are obtained by maximizing

$$P(\sigma,\mathbf{a},\mathbf{t}(n)|\mathbf{x}) \propto p_{\mathbf{x}}(\mathbf{x}|\sigma,\mathbf{a},\mathbf{t}(n)).$$

The log-likelihood cost function $J(\sigma,\mathbf{a},\mathbf{T}) = -\ln\Big(P(\sigma,\mathbf{a},\mathbf{t}(n)|\mathbf{x})\Big)$, is equal to

$$J(\sigma,\mathbf{a},\mathbf{T}) = \frac{N}{2}\ln(2\pi) + N\ln(\sigma) + \frac{||\mathbf{x}-\mathbf{T}\mathbf{a}||_2^2}{2\sigma^2}, \tag{7.48}$$

where $\mathbf{T}$ is the matrix whose rows are $\mathbf{t}(n)$.

The minimization with respect to $\mathbf{a}$ produces the MAP result $\mathbf{a} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{x}$, as in (7.12). This maximum a posterior (MAP) solution corresponds to the uniform prior $P(\sigma,\mathbf{a})$ and its is equal to the MLE solution. If a nonuniform prior $P(\sigma,\mathbf{a})$ to σ and $\mathbf{a}$ is added, then the posterior probability is

$$P(\sigma,\mathbf{a},\mathbf{T}(n)|\mathbf{x}) \propto p_{\mathbf{x}}(\mathbf{x}|\sigma,\mathbf{a},\mathbf{t}(n))P(\sigma,\mathbf{a}).$$

For the Gaussian prior to σ and $\mathbf{a}$

$$P(\sigma,\mathbf{a}) = \exp(-\lambda||\mathbf{a}||_2^2/(2\sigma^2)),$$

where λ is the parameter, the cost function is

$$J(\sigma,\mathbf{a},\mathbf{T}) = \frac{N}{2}\ln(2\pi) + N\ln(\sigma) + \frac{||\mathbf{x}-\mathbf{T}\mathbf{a}||_2^2 + \lambda||\mathbf{a}||_2^2}{2\sigma^2}. \tag{7.49}$$

Its MAP solution is given by (7.13).

For the Laplacian prior to σ and $\mathbf{a}$,

$$P(\sigma,\mathbf{a}) = \exp(-\lambda||\mathbf{a}||_1/(\sigma^2)),$$

which penalizes high values of elements and enforces the solution with the maximum possible number of the zero-valued elements in vector of coefficients $\mathbf{a}$, we get

$$J(\sigma,\mathbf{a},\mathbf{T}) = \frac{N}{2}\ln(2\pi) + N\ln(\sigma) + \frac{||\mathbf{x}-\mathbf{T}\mathbf{a}||_2^2 + \lambda||\mathbf{a}||_1}{2\sigma^2}, \tag{7.50}$$

whose MAP solution for $\mathbf{a}$ corresponds to the LASSO minimization of $||\mathbf{x}-\mathbf{T}\mathbf{a}||_2^2 + \lambda||\mathbf{a}||_1$.

For the Bayesian interpretation, the posterior probability $P(\sigma,\mathbf{a},\mathbf{T}(n)|\mathbf{x})$ should be calculated for discrete values of parameters $\mathbf{a}$ and σ, and the probabilistic interpretation of the results should be given.

The case of nonstationary Gaussian random signal, when we cannot assume either that the expected value and the variance of the samples are time-invariant or that the samples are statistically independent is more complex. This case will be considered in Part VI.

7.4.6 Cramer-Rao Bound

Consider the signal $\mathbf{x} = [x(1),x(2),\ldots,x(n)]^T$ and its true parameter θ whose unbiased estimate, obtained using the data in $\mathbf{x}$, is $\hat{\theta}(\mathbf{x})$. For the unbiased estimate, holds

$$\mathrm{E}\{\hat{\theta}(\mathbf{x}) - \theta\} = 0,$$

by definition. This means that

$$\frac{\partial}{\partial\theta}\mathrm{E}\{\hat{\theta}(\mathbf{x})-\theta\}=0 \quad \text{for all} \quad \theta.$$

Assuming that the probability density function of $\mathbf{x}$, with an assumed θ, is $p_{\mathbf{x}}(\mathbf{x}|\theta)$, we can write

$$\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}(\hat{\theta}(\mathbf{x})-\theta)p_{\mathbf{x}}(\mathbf{x}|\theta)d\mathbf{x}=0 \quad \text{for all} \quad \theta.$$

After the differentiation is performed, this equation can be rewritten in the form

$$-\int_{-\infty}^{\infty}p_{\mathbf{x}}(\mathbf{x}|\theta)d\mathbf{x}+\int_{-\infty}^{\infty}(\hat{\theta}(\mathbf{x})-\theta)\frac{\partial p_{\mathbf{x}}(\mathbf{x}|\theta)}{\partial\theta}d\mathbf{x}=0 \quad \text{or}$$

$$\int_{-\infty}^{\infty}(\hat{\theta}(\mathbf{x})-\theta)\frac{\partial p_{\mathbf{x}}(\mathbf{x}|\theta)}{\partial\theta}d\mathbf{x}=1, \tag{7.51}$$

since the first integral is equal to 1, by definition. We know that the derivative of the logarithm of a function $p_{\mathbf{x}}(\mathbf{x}|\theta))$ is given by

$$\frac{\partial\ln(p_{\mathbf{x}}(\mathbf{x}|\theta))}{\partial\theta}=\frac{1}{p_{\mathbf{x}}(\mathbf{x}|\theta))}\frac{\partial p_{\mathbf{x}}(\mathbf{x}|\theta)}{\partial\theta}.$$

This means that relation (7.51) can be written in the form

$$\int_{-\infty}^{\infty}(\hat{\theta}(\mathbf{x})-\theta)\frac{\partial\ln(p_{\mathbf{x}}(\mathbf{x}|\theta))}{\partial\theta}p_{\mathbf{x}}(\mathbf{x}|\theta)d\mathbf{x}=1.$$

Now, we will adjusted the form of this relation for the Schwartz inequality application,

$$\Big(\int_{-\infty}^{\infty}\Big[(\hat{\theta}(\mathbf{x})-\theta)\sqrt{p_{\mathbf{x}}(\mathbf{x}|\theta)}\Big]\Big[\frac{\partial\ln(p_{\mathbf{x}}(\mathbf{x}|\theta))}{\partial\theta}\sqrt{p_{\mathbf{x}}(\mathbf{x}|\theta)}\Big]d\mathbf{x}\Big)^2=1 \tag{7.52}$$

According to the Schwartz inequality, $(\int f(x)g(x)dx)^2\leq\int f^2(x)dx\int g^2(x)dx$,

$$\begin{aligned}1&=\Big(\int_{-\infty}^{\infty}\Big[(\hat{\theta}(\mathbf{x})-\theta)\sqrt{p_{\mathbf{x}}(\mathbf{x}|\theta)}\Big]\Big[\frac{\partial\ln(p_{\mathbf{x}}(\mathbf{x}|\theta))}{\partial\theta}\sqrt{p_{\mathbf{x}}(\mathbf{x}|\theta)}\Big]d\mathbf{x}\Big)^2\\&\leq\int_{-\infty}^{\infty}(\hat{\theta}(\mathbf{x})-\theta)^2p_{\mathbf{x}}(\mathbf{x}|\theta)d\mathbf{x}\int_{-\infty}^{\infty}\Big(\frac{\partial\ln(p_{\mathbf{x}}(\mathbf{x}|\theta))}{\partial\theta}\Big)^2p_{\mathbf{x}}(\mathbf{x}|\theta)d\mathbf{x}.\end{aligned}$$

The inequality terms can be recognized as

$$\mathrm{Var}(\hat{\theta}(\mathbf{x}))=\int_{-\infty}^{\infty}(\hat{\theta}(\mathbf{x})-\theta)^2p_{\mathbf{x}}(\mathbf{x}|\theta)d\mathbf{x} \quad \text{and}$$

$$\int_{-\infty}^{\infty}\Big(\frac{\partial\ln(p_{\mathbf{x}}(\mathbf{x}|\theta))}{\partial\theta}\Big)^2p_{\mathbf{x}}(\mathbf{x}|\theta)d\mathbf{x}=\mathrm{E}\{\Big(\frac{\partial\ln(p_{\mathbf{x}}(\mathbf{x}|\theta))}{\partial\theta}\Big)^2\}.$$

Applying this notation, we finally get the Cramer-Rao bound for the variance of the estimated parameter

$$\mathrm{Var}(\hat{\theta}(\mathbf{x})) \geq \frac{1}{\mathrm{E}\{\left(\frac{\partial \ln(p_{\mathbf{x}}(\mathbf{x}|\theta))}{\partial\theta}\right)^2\}} = \frac{1}{I(\theta)}$$

where $I(\theta)$ is used to denote

$$I(\theta) = \mathrm{E}\{\left(\frac{\partial \ln(p_{\mathbf{x}}(\mathbf{x}|\theta))}{\partial\theta}\right)^2\}. \tag{7.53}$$

The equality in the Schwartz inequality holds *if and only if* the two functions in the integral in (7.52) are proportional to each other

$$\hat{\theta}(\mathbf{x}) - \theta = k\frac{\partial \ln(p_{\mathbf{x}}(\mathbf{x}|\theta))}{\partial\theta},$$

where $\sqrt{p_{\mathbf{x}}(\mathbf{x}|\theta)}$ on both sides is omitted. The constant k is obtained as $k = 1/I(\theta)$ from the condition that the integral in (7.52) is equal to 1 for $\hat{\theta}(\mathbf{x}) - \theta = k\partial \ln(p_{\mathbf{x}}(\mathbf{x}|\theta))/\partial\theta$. Therefore, for the optimal estimator and the minimal variance, the following equality

$$\frac{\partial \ln(p_{\mathbf{x}}(\mathbf{x}|\theta))}{\partial\theta} = I(\theta)(\hat{\theta}(\mathbf{x}) - \theta) \tag{7.54}$$

holds. This relation can be used to find the optimal estimator, $\hat{\theta}(\mathbf{x})$, and the minimal variance, $1/I(\theta)$, without the evaluation of the second-order derivative

$$\frac{\partial^2 \ln(p_{\mathbf{x}}(\mathbf{x}|\theta))}{\partial\theta^2} = -I(\theta) \tag{7.55}$$

Example 7.29. Consider the signal $x(n) = s(n) + \varepsilon(n)$, where $\varepsilon(n)$ is a zero-mean Gaussian noise. The aim is to estimate a parameter a of the sinusoidal signal $s(n)$, for example, its amplitude, frequency, or phase, from N samples of the signal, $\mathbf{x} = [x(1), x(2), \ldots, x(N)]^T$. Find the minimum variance estimator and the Cramer-Rao bound.

★The random variable $x(n) - s(n) = \varepsilon(n)$ is Gaussian distributed. For N statistically independent values of the error $\varepsilon(n)$, with the assumed parameter a value, holds

$$p_{\mathbf{x}}(x(1), \ldots, x(n)|a) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x(1)-s(1|a))^2}{2\sigma^2}} \times \cdots \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x(N)-s(N|a))^2}{2\sigma^2}}, \tag{7.56}$$

or in vector form

$$p_{\mathbf{x}}(\mathbf{x}|a) = \frac{1}{\sigma^N\sqrt{(2\pi)^N}} \exp\left(-\frac{||\mathbf{x} - \mathbf{s}|a||_2^2}{2\sigma^2}\right),$$

where $s(n|a)$ is the considered signal with the assumed parameter a, and $\mathbf{s}|a$ is its vector form.

The log-likelihood function for this random signal is

$$J(\mathbf{x}|a) = -\ln\left(p_{\mathbf{x}}(\mathbf{x}|a)\right) = \frac{N}{2}\ln(2\pi) + N\ln(\sigma) + \frac{||\mathbf{x} - \mathbf{s}|a||_2^2}{2\sigma^2}. \tag{7.57}$$

The first derivative of the expected value is given by

$$\mathrm{E}\Big\{\frac{\partial \ln\big(p_{\mathbf{x}}(\mathbf{x}|a)\big)}{\partial a}\Big\} = -\mathrm{E}\Big\{\frac{\partial J(\mathbf{x}|a)}{\partial a}\Big\} = \frac{2}{2\sigma^2}\mathrm{E}\Big\{\Big(\sum_{n=1}^{N}(x(n)-s(n|a))\frac{\partial s(n|a)}{\partial a}\Big)\Big\}. \tag{7.58}$$

(a) In the case when we want to estimate the amplitude a of the sinusoidal signal

$$s(n) = s(n|a) = a\cos(2\pi nk_0/N),$$

then $\partial s(n|a)/\partial a = \cos(2\pi nk_0/N)$ and

$$\begin{aligned} -\frac{\partial J(\mathbf{x}|a)}{\partial a} &= \frac{1}{\sigma^2}\Big(\sum_{n=1}^{N}(x(n)-a\cos(2\pi nk_0/N))\cos(2\pi nk_0/N)\Big) \\ &= \frac{N}{2\sigma^2}\Big(\sum_{n=1}^{N}(\frac{2}{N}x(n)\cos(2\pi nk_0/N)) - a\Big), \end{aligned} \tag{7.59}$$

since $\sum_{n=1}^{N} a\cos^2(2\pi nk_0/N) = aN/2$. Comparing relation (7.59) with (7.54), we can conclude that the optimal estimator and the minimum variance are, respectively, the cosine transform

$$\hat{a} = g(\mathbf{x}) = 2\sum_{n=1}^{N} x(n)\cos(2\pi nk_0/N)$$

and its variance

$$\mathrm{Var}\{\hat{a}\} = \sigma_a^2 = \frac{1}{I(a)} = \frac{2\sigma^2}{N}.$$

The minimum variance can also be obtained from the second derivative of the log-likelihood function, as in (7.55),

$$\frac{\partial^2 \ln(p_{\mathbf{x}}(\mathbf{x}|a))}{\partial a^2} = -\frac{\partial^2 J(\mathbf{x}|a)}{\partial a^2} = -I(a) \tag{7.60}$$

to produce

$$-\mathrm{E}\Big\{\frac{\partial^2 J(\mathbf{x}|a)}{\partial a^2}\Big\} = \frac{1}{\sigma^2}\mathrm{E}\Big\{\frac{\partial}{\partial a}\Big(\sum_{n=1}^{N}(x(n)-s(n|a))\frac{\partial s(n|a)}{\partial a}\Big)\Big\}$$

$$= \frac{1}{\sigma^2}\mathrm{E}\Big\{\Big(\sum_{n=1}^{N}(x(n)-s(n|a))\frac{\partial^2 s(n|a)}{\partial a^2} - \Big(\frac{\partial s(n|a)}{\partial a}\Big)^2\Big)\Big\} = -\frac{1}{\sigma^2}\sum_{n=1}^{N}\Big(\frac{\partial s(n|a)}{\partial a}\Big)^2,$$

since $\mathrm{E}\Big\{\Big(\sum_{n=1}^{N}(x(n)-s(n|a))\Big\} = 0$. The Cramer-Rao bound for the variance of the amplitude estimate is

$$\mathrm{Var}\{\hat{a}\} \geq \frac{1}{I(a)} = \frac{1}{\mathrm{E}\{\frac{\partial^2 J(\mathbf{x}|a)}{\partial a^2}\}} = \frac{\sigma^2}{\sum_{n=1}^{N}\Big(\frac{\partial s(n|a)}{\partial a}\Big)^2}.$$

For the sinusoidal signal $s(n) = a\cos(2\pi nk_0/N)$, in this way we confirm the previous result

$$\mathrm{Var}\{\hat{a}\} \geq \frac{2\sigma^2}{N},$$

with $\partial s(n|a)/\partial a = \cos(2\pi nk_0/N)$ and $\sum_{n=1}^{N}\cos^2(2\pi nk_0/N) = N/2$.

(b) Consider now the frequency estimation of the signal

$$x(n) = \sin(an) + \varepsilon(n).$$

For this signal $\partial s(n|a)/\partial a = n\cos(an)$ and the bound for the variance of the frequency estimation is

$$\text{Var}\{\hat{a}\} \geq \frac{\sigma^2}{\sum_{n=1}^{N}\left(\frac{\partial s(n|a)}{\partial a}\right)^2} = \frac{\sigma^2}{\sum_{n=1}^{N} n^2\cos^2(an)}.$$

The Cramer-Rao bound is shown in Fig. 7.19 for $N = 10$ and $N = 50$, and various values of a, with $\sigma^2 = 1$.

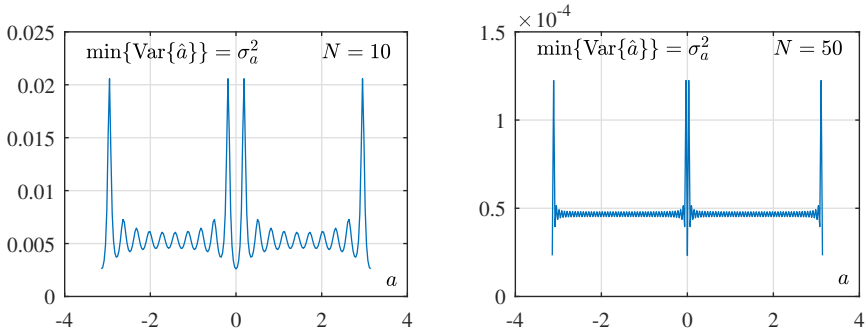


Figure 7.19 Cramer-Rao bound for the variance of the frequency estimation.

Example 7.30. Consider the signal $x(t_n) = at_n + \varepsilon(n)$, $n = 1,2,\ldots,N$, where $\varepsilon(n)$ is a zero-mean Gaussian noise. The gaol is to revisit the linear regression model and the estimation of the parameter a and its variance. What is the optimal estimator for a from the available data $x(t_n)$, given in the vector $\mathbf{x}$ for instants being elements of the vector $\mathbf{t}$? What is the variance of the optimal estimator of a (the Cramer-Rao bound)?

★The cost function for this random signal, with zero-mean Gaussian noise $\varepsilon(n)$, is

$$J(\mathbf{x}|a) = -\ln\left(p_{\mathbf{x}}(\mathbf{x}|a)\right) = \frac{N}{2}\ln(2\pi) + N\ln(\sigma) + \frac{||\mathbf{x} - a\mathbf{t}||_2^2}{2\sigma^2}. \tag{7.61}$$

The first derivative is given by

$$\frac{\partial J(\mathbf{x}|a)}{\partial a} = -\frac{1}{\sigma^2}\sum_{n=1}^{N}(x(t_n) - at_n)t_n = -\frac{\sum_{n=1}^{N} t_n^2}{\sigma^2}\left(\frac{\sum_{n=1}^{N} x(t_n)t_n}{\sum_{n=1}^{N} t_n^2} - a\right).$$

When this expression is compared to $I(\theta)(g(\mathbf{x}) - \theta))$ in (7.54), we get the optimal estimator form

$$\hat{a}(\mathbf{x}) = \frac{\sum_{n=1}^{N} x(t_n)t_n}{\sum_{n=1}^{N} t_n^2}$$

and its variance is

$$\mathrm{Var}(\hat{a}(\mathbf{x})) = \sigma_a^2 = \frac{\sigma^2}{\sum_{n=1}^{N} t_n^2}.$$

Example 7.31. We can come to the Cramer-Rao relations in an inductive way, analyzing the mean value estimation in the Gaussian distributed random variable, presented in Section 7.4.5,

$$p_{\mathbf{x}}(\mathbf{x}|\mu) = \frac{1}{\sigma^N \sqrt{(2\pi)^N}} \exp\left(-\frac{||\mathbf{x}-\mu||_2^2}{2\sigma^2}\right).$$

with the log-likelihood function (7.47) used for the estimation of the Gaussian distribution parameters

$$J(\mu) = -\ln\Big(p_{\mathbf{x}}(\mathbf{x}|\mu)\Big) = \frac{N}{2}\ln(2\pi) + N\ln(\sigma) + \frac{||\mathbf{x}-\mu||_2^2}{2\sigma^2}. \tag{7.62}$$

The estimated mean value follows from $\partial J(\mu)/\partial\mu = 0$ with

$$\frac{\partial J(\mu)}{\partial \mu} = -\frac{\partial \ln\Big(p_{\mathbf{x}}(\mathbf{x}|\mu)\Big)}{\partial \mu} = -\frac{N}{\sigma^2}\Big(\frac{1}{N}\Big(\sum_{n=1}^{N}(x(n)) - \mu\Big). \tag{7.63}$$

The second derivative, with respect to the mean value, is

$$\frac{\partial^2 J(\mu)}{\partial \mu^2} = \frac{N}{\sigma^2}. \tag{7.64}$$

The minimum variance of the mean value estimation is

$$\sigma_\mu^2 = \frac{1}{\frac{\partial^2 J(\mu)}{\partial \mu^2}} = \frac{\sigma^2}{N}. \tag{7.65}$$

This is quite specific formulation. However, this relation holds for a general unbiased estimator of the parameter θ and the cost function $J(\mathbf{x}|\theta) = -\ln\Big(p_{\mathbf{x}}(\mathbf{x}|\theta)\Big)$. The minimum variance of the estimated parameter θ is

$$\mathrm{Var}\{\hat{\theta}\} = \sigma_\theta^2 \geq \frac{1}{\mathrm{E}\{\frac{\partial^2 J(\mathbf{x}|\theta)}{\partial \theta^2}\}} = \frac{1}{-\mathrm{E}\{\Big(\frac{\partial \ln(p_{\mathbf{x}}(\mathbf{x}|\theta))}{\partial \theta}\Big)^2\}}.$$

In addition, it has been shown that an unbiased estimator attains the bound for all θ if and only if the first derivative of the log-likelihood function can be written in the form (7.54)

$$\frac{\partial \ln\Big(p_{\mathbf{x}}(\mathbf{x}|\theta)\Big)}{\partial \theta} = I(\theta)(g(\mathbf{x}-\theta)), \tag{7.66}$$

where the estimator with minimum variance is defined by $\hat{\theta}(\mathbf{x}) = g(\mathbf{x})$ and the minimum variance value is $\mathrm{Var}\{\theta\} = \frac{1}{I(\theta)}$.

Notice that the relation in (7.63) is of this form, with $I(\mu) = N/\sigma^2$ and $g(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N} x(n)$.

Cramer-Rao bound for the minimum variances in simultaneous estimation of more than one parameter, for example, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$, from the data in $\mathbf{x}$ can also be derived following similar concepts.

7.4.7 Confidence Intervals

The result of an experiment or calculation is commonly a random variable. When the estimate $x(n)$ is provided, the main question is the confidence in this specific value and how far the true (expected) value of the considered physical or mathematical value could be? The confidence interval provides a range within which the true value is estimated to lie. This interval provides the reliability of the presented estimate.

Consider a Gaussian distributed random variable as the most common case in practice. Assume that the experiment (or calculation) results are Gaussian distributed. The aim of the experiment is to estimate the unknown true value μ_x. For the Gaussian variable, it is known that all results of an experiment, $x(n)$, will be within the interval (7.44)

$$[\mu_x - 2\sigma_x,\ \ \mu_x + 2\sigma_x],$$

with the probability

$$\text{Probability}\{-2\sigma_x < x(n) - \mu_x < 2\sigma_x\} = \text{erf}(\sqrt{2}) = 0.9545. \tag{7.67}$$

This probability is sufficient for most of the experiments. If required, the probability can be increased using wider intervals. Here, the unknown true value μ_x and the interval bounds are fixed values, without any randomness.

The confidence intervals are calculated for the specific outcome of the experiment, $x(n)$, and the a priory estimated spread measure (here the standard deviation σ_x). The confidence intervals are defined as

$$[x(n) - 2\sigma_x,\ \ x(n) + 2\sigma_x] \tag{7.68}$$

Obviously, the confidence interval is not the same as the interval in (7.67), meaning that a 0.95 probability of (7.67) does not mean that any results of the experiment will be within the confidence interval with the same probability. However, if we know that the obtained result $x(n)$ is within the interval in (7.67), with the probability of 0.95, then it means that the true value μ_x is within the confidence interval, $[x(n) - 2\sigma_x, x(n) + 2\sigma_x]$, with the same probability, Fig. 7.20.

Example 7.32. A deterministic signal $s(n)$, with an additive Gaussian noise $\varepsilon(n)$, is observed at two instants n_1 and n_2. The standard deviation of the measurement method at the instant n_1 was $\sigma_x(n_1) = 0.5$, while the standard deviation of the measurement method at n_2 was $\sigma_x(n_2) = 0.2$ (different estimation approaches were used, for example, different windows for averaging; for the same measurement method, $\sigma_x(n_1) = \sigma_x(n_2)$ would hold). The observed values in these two measurements are denoted by $x(n)$, and they are equal to:

(a) $x(n_1) = 1.1$ and $x(n_2) = -0.2$;

(b) $x(n_1) = -0.6$ and $x(n_2) = 1.8$.

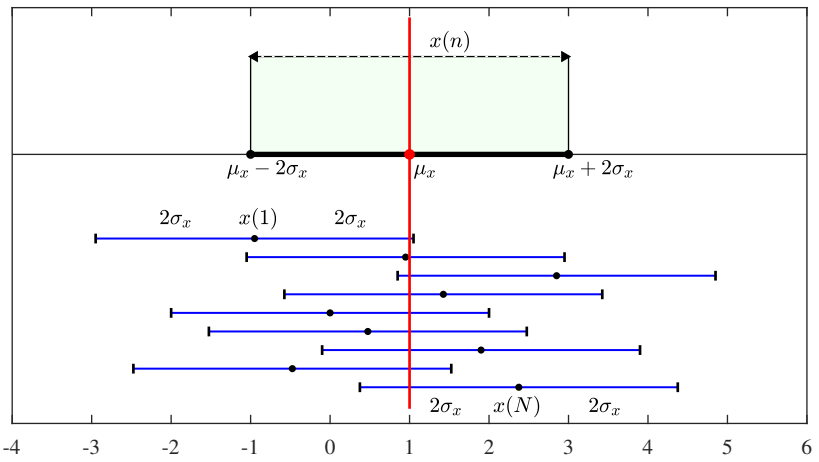


Figure 7.20 The interval $[\mu_x - 2\sigma_x,\ \mu_x + 2\sigma_x]$ where the Gaussian random variable $x(n)$ lies with the probability of 0.95, along with the confidence intervals for various $x(n)$ from the defined interval. The common point for all these confidence intervals is the true mean value μ_x (vertical line).

Could we conclude that the true signal $s(n)$ has changed, that is $s(n_1) \neq s(n_2)$, for these two cases (for an experiment, this is the question how can we be confident that a difference in the true result is obtained under different experiment conditions at n_1 and n_2). The common probability of 0.95 is assumed for the confidence interval definition.

★ (a) For the signal values (experiment outcomes) $x(n_1) = 0.6$ and $x(n_2) = 0.1$, the corresponding confidence intervals are

$$[x(n_1) - 2\sigma_x(n_1), x(n_1) + 2\sigma_x(n_1)] = [1.1 - 1, 1.1 + 1] = [0.1, 2.2] \tag{7.69}$$

and

$$[x(n_2) - 2\sigma_x(n_2), x(n_2) + 2\sigma_x(n_2)] = [-0.2 - 0.4, -0.2 + 0.4] = [-0.6, 0.2]. \tag{7.70}$$

These two confidence intervals overlap, meaning that we can not exclude the case that both true values, $s(n_1)$ and $s(n_2)$, are within the overlapping interval $[0.1, 0.2]$ and that they can take the same value within this overlapping interval.

(b) When the obtained signal values are $x(n_1) = -0.6$ and $x(n_2) = 1.8$, the corresponding confidence intervals are

$$[x(n_1) - 2\sigma_x(n_1), x(n_1) + 2\sigma_x(n_1)] = [-0.6 - 1, -0.6 + 1] = [-1.6, -0.4] \tag{7.71}$$

and

$$[x(n_2) - 2\sigma_x(n_2), x(n_2) + 2\sigma_x(n_2)] = [1.8 - 0.4, 1.8 + 0.4] = [1.4, 2.2]. \tag{7.72}$$

These two confidence intervals are clearly separated, meaning that the true signal values, $s(n_1)$ and $s(n_2)$, are different with a sufficiently high probability.

Example 7.33. Consider a random signal $x(n)$ that can take values {No, Yes} or $\{0,1\}$ 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 $k = 555$ times, find the interval where the true p will be with the probability of 0.95.

Notes: The mean value of samples $x(n)$ is defined by

$$\hat{p} = \frac{1}{N}\Big(x(1) + x(2) + \cdots + x(N)\Big) = k/N.$$

For the binomial distribution, $\binom{N}{k}p^k(1-p)^{N-k}$ with $x(n) \in \{0,1\}$, the expected value of $\hat{p}$ is

$$\mathrm{E}\{\hat{p}\} = \mathrm{E}\{k\}/N = \frac{pN}{N} = p.$$

The variance of $\hat{p}$ is given by

$$\sigma_{\hat{p}}^2 = \mathrm{Var}\{\hat{p}\} = \mathrm{Var}\{k\}/N^2 = p(1-p)/N.$$

The number $k = x(1) + x(2) + \cdots + x(N)$, as a sum of random variables, $x(n)$, can be considered as the Gaussian distributed variable (according to the CLT) with the expected value pN and the variance $\mathrm{Var}\{k\} = p(1-p)N$, that is

$$\binom{N}{k}p^k(1-p)^{N-k} \simeq \frac{1}{\sqrt{2\pi p(1-p)N}}e^{-\frac{(k-pN)^2}{2p(1-p)N}},$$

with a good rule of thumb $Np \geq 10$ and $Np(1-p) \geq 10$, Fig. 7.21.

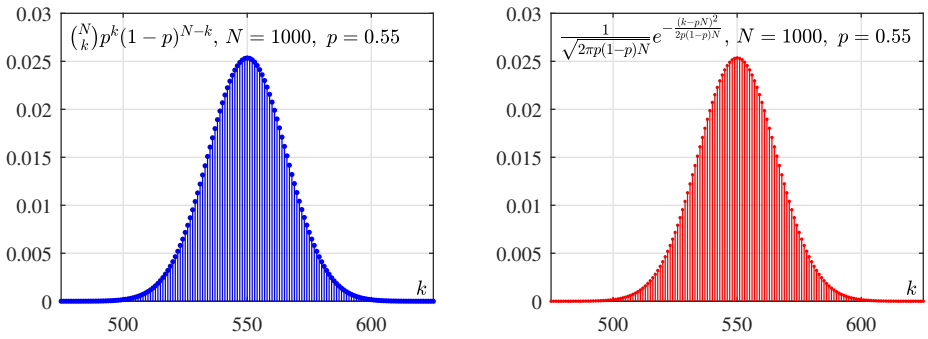


Figure 7.21 Binomial distribution for $N = 1000$ and $p = 0.55$ as a function of k (left) and the Gaussian distribution with the mean value pN and the variance $\sigma^2 = p(1-p)N$.

★For the given observation, with $k = 555$ responses $x(n) = 1$, the expected value p of the binomial distributed random variable is estimated as

$$\hat{p} = \frac{k}{N} = \frac{555}{1000} = 0.555.$$

The variance of this random variable is

$$\sigma_{\hat{p}}^2 = \text{Var}\{k\}/N^2 = p(1-p)/N.$$

For the variance estimation we should know the exact value of p, which is not the case. With the assumption that $\hat{p}$ is not far from the exact p, we can use the value of $\hat{p}$ in the variance calculation

$$\sigma_{\hat{p}}^2 = \frac{p(1-p)}{N} = \frac{\hat{p}(1-\hat{p})}{N}\frac{p(1-p)}{\hat{p}(1-\hat{p})} \simeq \frac{\hat{p}(1-\hat{p})}{N} = \frac{\frac{555}{1000}(1-\frac{555}{1000})}{1000} = \frac{0.2470}{1000},$$

and $\hat{\sigma}_{\hat{p}} = 0.0157$. Therefore, the estimated value $\hat{p} = 0.555$ is within the interval

$$\hat{p} = 0.555 \in \left[p - 2\hat{\sigma}_{\hat{p}}, p + 2\hat{\sigma}_{\hat{p}}\right] = [p - 0.0314, p + 0.0314]$$

with the probability of 0.95, meaning that the true value p is within

$$-0.0314 \leq 0.555 - p \leq 0.0314 \quad \text{or} \quad |0.555 - p| \leq 0.0314.$$

with the same probability. The true value is around 55.5%, within ±3.14% range (from 52.36% to 58.64%) with the probability of 0.95.

By increasing the value of N we can reduce the margin of the estimation error ($\hat{\sigma}_{\hat{p}} \propto 1/\sqrt{N}$). However, about 1000 values are commonly used for various *opinion poll estimations*.

Bayesian analysis. Within the Bayes' framework, the probability of the event B (k times $x(n) = 1$ (Yes) and $N-k$ times $x(n) = 0$ (No)), with an assumed p is equal to

$$P(B|p) = \binom{N}{k} p^k (1-p)^{N-k}. \tag{7.73}$$

The posterior is

$$P(p|B) = P(B|p)P(p)/P(B) \propto P(B|p)P(p) \propto P(B|p),$$

for the uniform prior $P(p) = C$.

For the given event B, when $k = 555$ and $N = 1000$, we get

$$P(p|B) \propto P(B|p) = \binom{1000}{555} p^{555}(1-p)^{445}.$$

The value of the posterior $P(p|B)$ is shown in Fig. 7.22 for $0 \leq p \leq 1$ with a step of $\Delta p = 0.005$.

From Fig. 7.22 we can concluded that the posterior, $P(p|B)$, is maximum at $p = 0.555$, while the region of significant $P(p|B)$ values is about 7 steps $\Delta p = 0.005$ left and right from the maximum position, corresponding to $0.555 \pm 7 \cdot 0.005 = 0.555 \pm 0.035$.

Student's t-distribution. In the previous analysis, we assumed that the standard deviation is known. When the true value (mean) estimation is done based on a small number of samples, then the standard deviation has to be estimated as well. For the set of samples $x(n)$, we have the mean and the variance estimations

$$\hat{\mu}_x(n) = \frac{1}{N}\Big(x_1(n) + x_2(n) + \cdots + x_N(n)\Big) \tag{7.74}$$

$$\hat{\sigma}_x(n) = \sqrt{\frac{1}{N-1}\Big(|x_1(n) - \hat{\mu}_x(n)|^2 + |x_2(n) - \hat{\mu}_x(n)|^2 + \cdots + |x_N(n) - \hat{\mu}_x(n)|^2\Big)}. \tag{7.75}$$

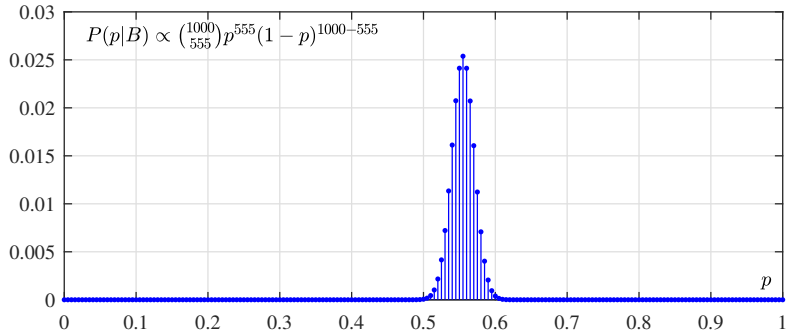


Figure 7.22 The posterior $P(p|B)$ proportional to the binomial distribution for the event B when $N = 1000$ and $k = 0.55$, as a function of the probability p.

The new random variable, where both $\hat{\mu}_x(n)$ and $\hat{\sigma}_x(n)$ are random, is

$$z(n) = \frac{\hat{\mu}_x(n) - \mu}{\hat{\sigma}_x(n)/\sqrt{N}}$$

is t-distributed (student distribution). The t-distribution is defined for a given degree of freedom $\nu = N - 1$ using the gamma functions. For large ν it approaches to the Gaussian distribution, while for $\nu = 1$ (just two samples) it is quite heavy-tailed and equal to the Cauchy distribution (see Section 7.4.11). The interval $-t_\nu < z(n) < t_\nu$, where a t-distributed random variable $z(n)$ takes its value with the probability 0.95, is

$$\text{Probability}\{-t_\nu < z(n) < t_\nu\} = 0.95 \tag{7.76}$$

for the value of t_ν given in Table 7.4 and for some values of $\nu = N - 1$. We can conclude that the confidence intervals are very wide for small N (for example, six times wider for $N = 2$ than for $N = 60$), while they are almost the same as in the Gaussian distributed random variable for large N, for example, $N \geq 12$.

$\nu = N - 1$	1	2	3	5	12	20	60	120
t_ν	12.076	4.303	3.182	2.571	2.179	2.086	2.000	1.980

Table 7.4

Values of the interval bounds for the t-distribution for $\text{Probability}\{-t_\nu < z(n) < t_\nu\} = 0.95$.

Example 7.34. The available samples of the random signal $\mathbf{x}$, with the elements $x(n)$, are given by
(a) $\mathbf{x} = (0.93,\ 0.17, -0.69, -0.72)$,
(b) $\mathbf{x} = (0.93,\ 0.17, -0.69, -0.72, -0.57, -0.31,\ 0.27,\ 1.33. - 1.33,\ 1.40, -0.57, -0.35, -0.64)$.
Find the confidence intervals, where the true mean of this random signal is expected with the probability of 0.95.

★In this experiment, both the mean value and the variance are not known and should be estimated based on the available data.

(a) For $N = 4$ and $\nu = N - 1 = 3$, and the available realizations of $x(n)$, we have

$$\hat{\mu}_x = \frac{1}{4}\Big(x(0) + x(1) + x(2) + x(3)\Big) = -0.08,$$

$$\hat{\sigma}_x = \sqrt{\frac{1}{3}\Big(|x(0) - \hat{\mu}_x|^2 + |x(1) - \hat{\mu}_x|^2 + |x(2) - \hat{\mu}_x|^2 + |x(3) - \hat{\mu}_x|^2\Big)} = 0.79.$$

The confidence interval of the normalized and centered random signal

$$z(n) = \frac{\hat{\mu}_x - \mu}{\hat{\sigma}_x/\sqrt{4}} = \frac{-0.08 - \mu}{0.39},$$

for the probability of 0.95, is defined by $-3.182 < z(n) < 3.182$ (see Table 7.4), or

$$[-0.08 - 0.39 \times 3.182, -0.08 + 0.39 \times 3.182] = [-0.08 - 1.24,\ -0.08 + 1.24]. \tag{7.77}$$

(b) When the number of realizations is increased to $N = 13$, we get

$$\hat{\mu}_x = \frac{1}{13}\sum_{n=0}^{12} x(n) = -0.08,$$

$$\hat{\sigma}_x = \sqrt{\frac{1}{12}\sum_{n=0}^{12} |x(n) - \hat{\mu}_x|^2} = 0.82.$$

For the random variable

$$z(n) = \frac{\hat{\mu}_x - \mu}{\hat{\sigma}_x/\sqrt{13}} = \frac{-0.08 - \mu}{0.23},$$

the confidence interval, for $\nu = N - 1 = 12$ and the probability of 0.95, is (see Table 7.4)

$$[-0.08 - 0.23 \times 2.179, -0.08 + 0.23 \times 2.179] = [-0.08 - 0.5,\ -0.08 + 0.5]. \tag{7.78}$$

Although the same value of the mean value is obtained in both cases, with similar standard deviations, the confidence interval in (b) shows that the experiment with $N = 13$ realizations produces a more reliable estimation of the true mean μ.

Repeat Example 7.27 with the data from this example and comment on the results within the frequentist and the Bayesian framework.

Variance stabilization – Delta method. Consider again the Bernoulli random variable from Example 7.33. The estimate of the expected value, p, of the probability that $x(n) = 1$ will appear in the Bernoulli trial, is given by

$$\hat{p} = \frac{k}{N} = \frac{1}{N}\sum_{n=1}^{N} x(n),$$

where k is the number of $x(n) = 1$ appearances in N samples. For large N, this estimate is approximately Gaussian distributed, Fig. 7.21, with the expected value $\mathrm{E}\{\hat{p}\} = \mathrm{E}\{k\}/N = pN/N = p$ and the variance

$$\sigma_{\hat{p}}^2 = \mathrm{Var}\{\hat{p}\} = \mathrm{Var}\{k\}/N^2 = p(1-p)/N.$$

The property that $\hat{p} - p$ tends to the Gaussian distributed random variable can be written as

$$\hat{p} - p \xrightarrow{D} \mathcal{N}(0, \sigma_{\hat{p}}^2).$$

The problem in the confidence interval definition for p was that the variance $\sigma_{\hat{p}}^2$ depends on the parameter p which is to be estimated. This has been solved in Example 7.33 using $p \simeq \hat{p}$. Another approach to this problem is based on the so called *Delta method*. This method states that for any differentiable function $g(x)$ holds

$$g(\hat{p}) - g(p) \xrightarrow{D} \mathcal{N}(0, (g'(p))^2 \sigma_{\hat{p}}^2), \tag{7.79}$$

for $g'(p) \neq 0$. The proof is simple since for large data set size, N, we can assume that $\hat{p} - p$ is small so that the linear Taylor series expansion for the function $g(\hat{p})$ holds around p, that is

$$g(\hat{p}) = g(p) + g'(p)(\hat{p} - p),$$

meaning that $g(\hat{p}) - g(p)$ behaves as $(\hat{p} - p)$, for sufficiently large N, with the deterministic proportionality factor of $g'(p)$. Since $\text{Var}\{a(\hat{p} - p)\} = a^2 \text{Var}\{\hat{p}\}$, from the previous relation we get

$$\text{Var}\{g(\hat{p})\} = (g'(p))^2 \text{Var}\{\hat{p}\} = (g'(p))^2 \sigma_{\hat{p}}^2,$$

proving the Delta method.

Now the Delta method will be used to avoid the variance dependence on parameter p (variance stabilization). The resulting variance in (7.79) is parameter p independent if

$$(g'(p))^2 \sigma_{\hat{p}}^2 = (g'(p))^2 p(1-p)/N = C$$

holds, where C is a constant, equal to the resulting variance.

The intuitive solution to this problem is

$$g(p) = \arcsin(\sqrt{p}) \quad \text{when} \quad g'(p) = \frac{1}{2\sqrt{p}} \frac{1}{\sqrt{1-p}} \tag{7.80}$$

and

$$(g'(p))^2 \sigma_{\hat{p}}^2 = (g'(p))^2 p(1-p)/N = \frac{1}{4N}.$$

This means that the random variable $\arcsin(\sqrt{\hat{p}})$ is Gaussian distributed with the expected value $\arcsin(\sqrt{p})$ and the variance $1/(4N)$. The confidence intervals for $\arcsin(\sqrt{p})$, with the probability of 0.95, are defined by the two-sigma rule

$$\arcsin(\sqrt{p}) \in [\arcsin(\sqrt{\hat{p}}) - 2\frac{1}{\sqrt{4N}}, \ \arcsin(\sqrt{\hat{p}}) + 2\frac{1}{\sqrt{4N}}]. \tag{7.81}$$

The confidence intervals for p are then obtained taking the sinus of both bounds and squaring the result

$$p \in [\sin^2\left(\arcsin(\sqrt{\hat{p}}) - \sqrt{\frac{1}{N}}\right), \ \sin^2\left(\arcsin(\sqrt{\hat{p}}) + \sqrt{\frac{1}{N}}\right)]. \tag{7.82}$$

In the case when $\sin\left(\arcsin(\hat{p}) - 1/\sqrt{N}\right)$ is negative, the zero value is used as the lower bound.

For the data from Example 7.33 we get

$$p \in [\sin^2\left(\arcsin(\sqrt{0.555}) - \sqrt{\frac{1}{1000}}\right), \ \sin^2\left(\arcsin(\sqrt{0.555}) + \sqrt{\frac{1}{1000}}\right)] = [0.5235, \ 0.5863].$$

This interval is almost the same as $[0.5236,\ 0.5864]$, obtained in Example 7.33, using $p \simeq \hat{p}$ in the variance estimation.

7.4.8 Bootstrap Method

The bootstrap is a simple method for statistical inference using remarkable modern computing power, without relying on many assumptions about the random variable. The main idea is to estimate a statistic of the considered signal by increasing the number of signal realizations using the existing data and resampling. Here, is the origin of the method name "pulling itself up by its own bootstrap." In producing many realizations, the bootstrap method relies on resampling the existing signal with replacement.

The bootstrap method can be summarized as follows:

1. Consider a signal (data set) $\{x(n),\ n = 1,2,\ldots,N\}$, being a part of much larger population $\{x(n),\ n = 1,2,\ldots,P\}$, $P \gg N$.

   The aim is to provide a statistic as an estimate of the corresponding large population parameter, using the available data set with N samples only.

2. The original data set $x(n)$, $n = 1,2,\ldots,N$ is resampled into new signals of the length M. We will consider cases with $M = N/2$ and $M = N$ and the inference is performed based on these resampled data.

   A new resampled realization of the signal is formed as follows: (a) A random signal sample from $x(n),\ n = 1,2,\ldots,N$, is picked up and assigned to $x_1(1)$. Then this sample is "returned" to the original data set (so that it can be picked up again, by chance – *resampling with replacement*). (b) A new signal sample is randomly picked up from the original set $x(n),\ n = 1,2,\ldots,N$ and assigned to $x_1(2)$. This sample is also "returned" to the original data set. This procedure is repeated M times to form new resampled signal (*Bootstrap Sample*) $\mathbf{x}_1$ with M elements.

3. The desired statistic (in our example, we will consider the mean value) is estimated using $\mathbf{x}_1$ as $\hat{\mu}(1) = \text{mean}\{\mathbf{x}_1\}$.

4. The Steps 2 and 3 are repeated for every $\mathbf{x}_b$, $b = 1,2,\ldots,B$, to get

$$\hat{\mu}(b) = \text{mean}\{\mathbf{x}_b\},\ \ b = 1,2,\ldots,B.$$

   In practice, B can be quite large.

5. Statistics of the data set and the sampling distribution is obtained and analyzed as if we had a large number of data subsets $\mathbf{x}_b$, $b = 1,2,\ldots,B$. This new distribution can be used to make a further statistical inference, such as, for example, to estimate the confidence intervals for the estimated parameter.

Example 7.35. In order to introduce the basic definitions and principles of the bootstrap method we will revisit the introductory Example 7.1 and the signal shown in Fig. 7.1, whose values are given in Table 7.1. Here, we will assume that this set of $N = 100$ signal values, $x(n)$, is a sample of a large population with $P \gg N$ elements. The aim is to estimate the mean value of a large population using the statistics of the available data set.

In order to perform the statistical analysis using the bootstrap, new realizations should be created by resampling the original data with replacement. An illustration of this resampling is given in Table 7.5 for $M = 20$ and $B = 15$. The new resampled signals, $\mathbf{x}_b$, are obtained by sampling the original data with replacement, as described in Step 2. Consider, for example, $\mathbf{x}_7$, given in the seventh column of this table. Note that the signal sample $x(n) = 48$ is repeated,

although there is only one sample $x(n) = 48$ in the original data set, while many other signal values do not appear at all in this realization.

A set of $B = 1000$ resampled realizations of the original signal, $\mathbf{x}_b$ is formed next. The bootstrap is applied to this data set with: (a) $M = N/2 = 50$ and (b) $M = N = 100$.

The results are shown in Fig. 7.23. We can see that the maximum of the normalized histogram of $B = 1000$ values of $\hat{\mu}(b) = \text{mean}\{\mathbf{x}_b\}$ is close to the sample mean value calculated as the sample average of all 100 available data values. We can also conclude that the confidence intervals can be estimated considering the probability distribution, $F_{\hat{\mu}}$, and its, for example, 0.05 and 0.95 levels.

Since the considered data in Example 7.1 are of the Gaussian nature (what is not an assumption required by the bootstrap method) we can compare this result with the one obtained from the variance of the mean value estimate in the Gaussian distributed variable (7.65), using the standard deviation calculated in Example 7.5, $\sigma_{\hat{\mu}} = \hat{\sigma}_x/\sqrt{M} = 17.73/\sqrt{50} = 2.5$. For the probability of 0.90 the confidence intervals for the Gaussian distribution would be $[55.76 - 1.65\sigma_{\hat{\mu}},\ 55.76 + 1.65\sigma_{\hat{\mu}}] = [51.63,\ 59.88]$. The confidence intervals of the mean value estimation obtained with the bootstrap method correspond with the theoretical ones for this distribution.

Table 7.5

Bootstrap resampling of the signal $x(n)$, $n = 1,2,\ldots,100$ from Fig. 7.1. $B = 15$ new signals $\mathbf{x}_b$, $b = 1,2,\ldots,B$, are formed. Every new signal is of $M = 20$ length. New resampled signals are formed by randomly picking up a sample form $x(n), n = 1,2,\ldots,N$, then "returning" this sample into the original set (so that it can be picked up again, by chance), randomly picking up second sample of $\mathbf{x}_b$, "returning" it, and so on M times. This procedure is repeated for every $\mathbf{x}_b$, $b = 1,2,\ldots,B$. In practice, B is commonly large.

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$	$\mathbf{x}_{10}$	$\mathbf{x}_{11}$	$\mathbf{x}_{12}$	$\mathbf{x}_{13}$	$\mathbf{x}_{14}$	$\mathbf{x}_{15}$
42	66	83	70	38	69	42	42	77	49	55	80	25	12	66
56	49	36	57	77	54	52	55	38	37	95	73	84	42	66
40	52	64	59	45	56	50	54	50	36	69	35	36	64	64
71	99	31	55	71	75	67	64	66	57	69	64	18	56	61
67	72	59	61	28	41	61	25	54	42	25	49	61	26	47
44	69	50	55	80	87	39	71	99	70	50	52	44	25	38
67	62	54	57	57	52	60	48	66	44	36	51	23	50	89
41	70	48	79	71	66	35	37	42	55	64	69	62	11	77
26	55	18	68	49	84	57	66	57	37	54	60	58	26	66
66	25	23	80	60	71	48	53	12	56	52	56	54	51	63
50	59	71	42	66	23	70	71	69	44	49	53	71	62	68
37	31	66	43	40	25	45	80	83	71	71	71	57	18	28
11	54	71	26	84	69	60	52	40	56	55	57	61	50	55
69	55	89	42	50	55	39	36	53	71	38	67	63	45	62
18	95	43	77	74	34	52	67	55	67	84	66	44	52	64
40	66	66	99	59	55	59	35	57	57	67	42	66	42	80
87	71	72	34	69	76	66	51	28	39	42	42	66	57	73
66	42	23	64	12	66	48	61	51	54	53	48	55	23	41
66	47	52	69	89	50	66	60	80	73	42	12	66	63	67
69	62	63	54	71	23	61	34	37	83	95	57	77	18	28

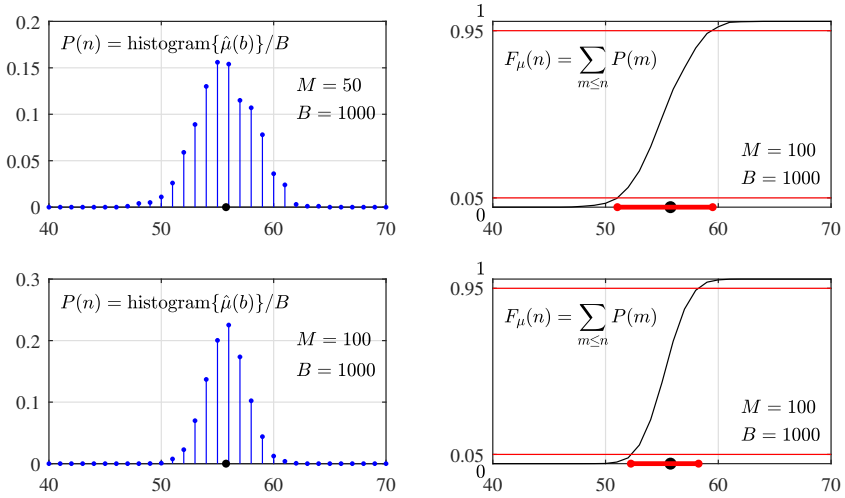


Figure 7.23 Bootstrap statistics of the mean value of a large population with the reduced set of available data shown in Fig. 7.1. A large point on the horizontal axis stands for the sample average of the considered data set.

7.4.9 Hypothesis Testing

The hypothesis testing was introduced in statistics since its foundations are established in the first part of the last century. The main goal of the hypothesis testing is to provide a statistical decision based on the experimental data (random signal values). Although the answer to this kind of question can be provided, in an indirect way, using the presented Bayes' inference or the confidence intervals, we will provide here its original analysis due to importance in signal processing and detection theory.

The basic concepts in the hypothesis testing are:

- *Null hypothesis*, H_0. It assumes that the tested event has not happened and that the experiment result is obtained by pure chance.
- *Alternative hypothesis* H_1 is contrary to the null hypothesis, meaning that the null hypothesis is rejected and the experiment result is not obtained by pure chance.
- *Level of significance* shows how we are confident in the decision made about accepting or rejecting the null hypothesis since the probability of 1 is not possible in this kind of testing. It is common to assume that the level of significance is equal to $\alpha = 0.05$ or $\alpha = 0.01$, corresponding to the probabilities of 0.95 or 0.99.
- *Type of error I* or false-negative result when the null hypothesis is rejected although this hypothesis was true.
- *Type of error II* or false-positive result when the null hypothesis is accepted, while this hypothesis was not true.

Example 7.36. Consider a multiple-choice test with 5 answers to each of $N = 20$ questions. Only one of these 5 answers is correct for every question. The null hypothesis, H_0, is the assumption that the person who answers the test does not have any knowledge of the test topic. Find the number of the correct answers when the null hypothesis can be accepted with the probability of 0.95.

★ The probability of a correct answer to a specific question, if the null hypothesis holds, is $p = 1/5$. The probability that the person will give k correct answers to $N = 20$ questions, with the null hypothesis, is already calculated, (7.37), and it is equal to

$$P(k|H_0) = \binom{20}{k} p^k (1-p)^{20-k}.$$

These probabilities are calculated for every k and shown in Fig. 7.24(a). The probability distribution is given in Fig. 7.24(b) and (c).

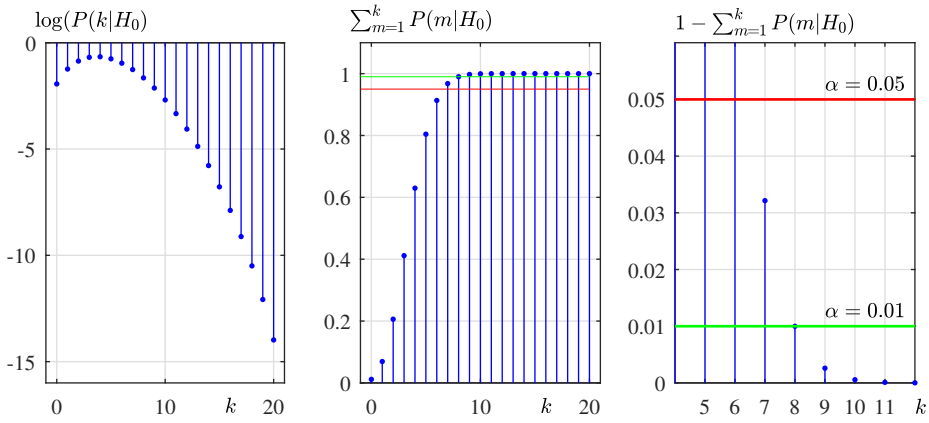


Figure 7.24 Hypothesis testing. (a) The logarithm of the probability of k correct answers with the null hypothesis, $P(k|H_0)$. (b) Cumulative probability distribution of $P(k|H_0)$. (c) Values of the complementary probability distribution, being equal to the probability that more than k correct answers will be given with the null hypothesis.

Now, for the given probability we can find the limit number, k, of the positive answers if the null hypothesis is true. Obviously, it is $k = 7$ for $\alpha = 0.05$. This means that if the person has given $k < 7$ correct answers, the decision should be that the null hypothesis (the person does have any knowledge of the tested subject) is true, with a significance level of 0.05. The hypothesis *rejection region* is $k \geq 7$.

For the significance level of $\alpha = 0.01$, the hypothesis rejection region would be $k \geq 8$.

For example, if the tested person provided $k = 10$ correct answers on this multiple choice test, the so-called p-value of this result of the experiment is equal to probability of the considered experiment producing such an outcome or anyone more extreme,

$$p = \sum_{k=10}^{20} P(k|H_0) = \sum_{k=10}^{20} \binom{20}{k} p^k (1-p)^{20-k} = 0.0026 < \alpha.$$

This means that for $k = 10$ correct answers, the null hypothesis should be rejected for both $\alpha = 0.05$ or $\alpha = 0.01$.

Finally, we will calculate the type of error I (false-negative result) for the case $k = 10$. It is equal to the probability that we have decided to reject the null hypothesis, since the p-value is $p = 0.0026 < \alpha$, but that the person, in reality, does not have any knowledge in this area and that null hypothesis was true. The type of error I is equal to the probability that the null hypothesis holds and that there were $k = 10$ correct answers, $P(k|H_0) = \binom{20}{k} p^k (1-p)^{20-k} = 0.002$, meaning that 1 person in 500 will achieve this kind of the result.

In many practical cases, we may assume that the random variables (random signal samples) in the considered experiment (in hypothesis testing, called population) are Gaussian distributed, under the null hypothesis. This assumption also holds if the particular random variable is not Gaussian, but the total number of samples is sufficiently large so that the distribution of the sample mean value is approximately Gaussian, for example, as it was the case in the poll analysis in Example 7.33, and proven in Example 7.24.

Consider a random signal, $x(n)$, whose probability density function is

$$p_x(\xi) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(\xi-\mu_x)/(2\sigma_x^2)}$$

under the null hypothesis. The result of the experiment is the signal sample $x(n) = A$. Here, we may consider three possible scenarios of practical interest for the null hypothesis rejection:

- The experiment result is not equal to the expected mean (two-sided test). This case corresponds to the case when we want to make the decision if any constant value (positive or negative) is added to the considered random variable under the null hypothesis. For the assumed level of significance, the region of rejection is obtained from

$$\text{Probability}\{|A-\mu_x| > \lambda\} = 1 - \text{erf}(\frac{\lambda}{\sqrt{2}\sigma_x}) < \alpha.$$

  For the significance level of $\alpha = 0.05$, the rejection region for the null hypothesis is

$$|A-\mu_x| > 2\sigma_x.$$

- The experiment result is greater than the expected mean (right-tailed test). This scenario appears when the aim is to establish if a certain action has increased the expected value in a positive direction. Here, we are not interested in a possible decrease in the mean. For the assumed level of significance, the region of rejection follows from

$$\text{Probability}\{A-\mu_x > \lambda\} = \frac{1}{2}(1 - \text{erf}(\frac{\lambda}{\sqrt{2}\sigma_x})) < \alpha.$$

  For the significance level of $\alpha = 0.05$, the rejection region for the null hypothesis is

$$A-\mu_x > 1.645\sigma_x. \tag{7.83}$$

- The experiment result is lower than the expected mean (left-tailed test). This is opposite to the previous one.

Example 7.37. An author was selling 34 books on average per week. To improve the sales of his book, the author designed and implemented an advertisement campaign. The following week, he sold 41 books. Can the author reject the null hypothesis (meaning that the advertisement campaign had no impact on book sales) with a significance level of $\alpha = 0.05$?

The number of sold books with the null hypothesis obeys the Poisson distribution (Section 7.4.12 and Problem 7.23)

$$P(x(n) = k|H_0) = \frac{\lambda^k e^{-\lambda}}{k!}$$

with $\lambda = 34$, that can be approximated (for large $\lambda \geq 20$) by the Gaussian distribution, with

$$\mu = \lambda = 34,$$

$$\sigma^2 = \lambda = 34,$$

as illustrated in Fig. 7.27,

$$p(\xi|H_0) = \frac{1}{\sqrt{68\pi}} e^{-(\xi-34)^2/68}.$$

★Since we are looking for a possible influence of the advertisement campaign to the increase in the number of sold books, we are interested in the right-tailed test, when the criterion for the hypothesis rejection is (7.83)

$$41 - 34 > 1.645 \times \sqrt{34},$$

with the observed value $A = 41$. Since $1.645 \times \sqrt{34} = 9.5919$, the author cannot reject the null hypothesis, meaning that the hypothesis that the advertisement campaign does not have any influence on the number of sold books cannot be rejected.

Example 7.38. The Fourier transform of a signal is presented in Fig. 7.25(left). The Fourier transform elements of the noise only (null hypothesis) are zero-mean Gaussian random variables with the variance $\sigma_X^2 = 1$. For every element of the Fourier transform, $X(k)$, test the null hypothesis and indicate the elements for which this hypothesis can be rejected with significance level $\alpha = 0.001$, meaning that we can reject the hypothesis that there is no signal component at the considered frequency index.

★ For the significance level of $\alpha = 0.001$ the rejection region of the null hypothesis, for the Gaussian random variable with the mean μ_X and the variance σ_X^2, is

$$\text{Probability}\{|X(k) - \mu_X| > \lambda\} = 1 - \text{erf}(\frac{\lambda}{\sqrt{2}\sigma_X}) < 0.001,$$
$$|X(k) - \mu_X| > 3.2905\sigma_X \quad \text{or} \quad |X(k)| > 3.2905$$

Therefore, we cannot reject the null hypothesis for all Fourier transform elements ($\mu_X = 0$), except those at $k \in \{4, 10, 66, 71, 88\}$, as shown in Fig. 7.25(right), where the rejection region, for the significance level $\alpha = 0.001$, is shaded.

When the mean value and the variance of the random variable in the experiment are not known in advance, then the t-distribution (see Example 7.34) should be used.

7.4.10 Complex Gaussian Noise and Rayleigh Distribution

In many application 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

$$\mu_\varepsilon = \text{E}\{\varepsilon(n)\} = \text{E}\{\varepsilon_r(n)\} + j\text{E}\{\varepsilon_i(n)\} = 0 + j0.$$

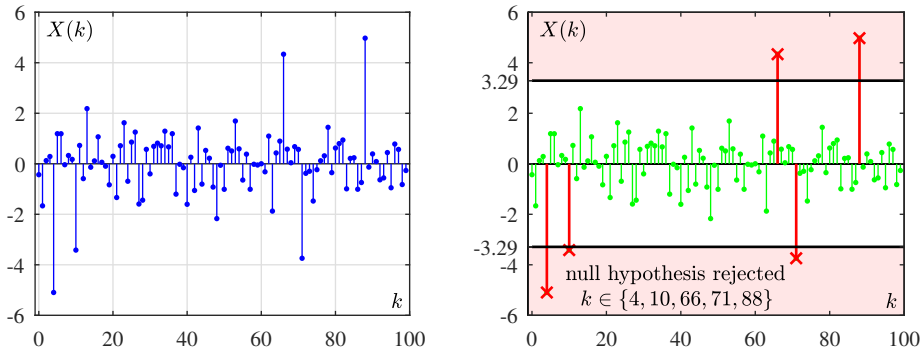


Figure 7.25 The null hypothesis testing for the Fourier transform of a signal with zero-mean Gaussian noise with the variance $\sigma_X^2 = 1$ (left). The null hypothesis rejection regions (shaded) for the random variable $X(k)$ with the significance level of $\alpha = 0.001$, corresponding to $|X(k)| > 3.2905$.

The variance is

$$\sigma_\varepsilon^2 = \mathrm{E}\{\varepsilon(n)\varepsilon^*(n)\} = \mathrm{E}\{\varepsilon_r(n)\varepsilon_r(n)\} + \mathrm{E}\{\varepsilon_i(n)\varepsilon_i(n)\} + j(\mathrm{E}\{\varepsilon_i(n)\varepsilon_r(n)\} - \mathrm{E}\{\varepsilon_r(n)\varepsilon_i(n)\})$$
$$= \mathrm{E}\{\varepsilon_r(n)\varepsilon_r(n)\} + \mathrm{E}\{\varepsilon_i(n)\varepsilon_i(n)\} = \sigma^2.$$

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 of the form

$$p_{|\varepsilon(n)|}(\xi) = \frac{2\xi}{\sigma^2} e^{-\xi^2/\sigma^2} u(\xi).$$

The probability density function $p_{|\varepsilon(n)|}(\xi)$ is called the Rayleigh distribution.

In order to prove the previous relation, consider the probability density functions of $\varepsilon_r(n)$ and $\varepsilon_i(n)$. Since they are independent and equally distributed, we get

$$p_{\varepsilon_r\varepsilon_i}(\xi,\zeta) = p_{\varepsilon_r}(\xi)p_{\varepsilon_i}(\zeta) = \frac{1}{\sigma^2\pi} e^{-(\xi^2+\zeta^2)/\sigma^2}.$$

The probability that $|\varepsilon(n)| = \sqrt{\varepsilon_r^2(n) + \varepsilon_i^2(n)} < \chi$ is

$$P\{\sqrt{\varepsilon_r^2(n) + \varepsilon_i^2(n)} < \chi\} = \iint\limits_{\xi^2+\zeta^2<\chi^2} p_{\varepsilon_r\varepsilon_i}(\xi,\zeta)d\xi d\zeta = \frac{1}{\sigma^2\pi} \iint\limits_{\xi^2+\zeta^2<\chi^2} e^{-(\xi^2+\zeta^2)/\sigma^2} d\xi d\zeta.$$

With $\xi = \rho\cos\alpha$ and $\zeta = \rho\cos\alpha$ (the Jacobian of the polar coordinate transformation, $J = |\rho|$), we get

$$P\{\sqrt{\varepsilon_r^2(n) + \varepsilon_i^2(n)} < \chi\} = \frac{1}{\sigma^2\pi} \int_0^{\chi} \int_0^{2\pi} e^{-\frac{\rho^2}{\sigma^2}} \rho d\rho d\alpha = \int_0^{\chi^2/\sigma^2} e^{-\lambda} d\lambda = (1 - e^{-\frac{\chi^2}{\sigma^2}})u(\chi) = F_{|\varepsilon(n)|}(\chi).$$

The probability density function is

$$p_{|\varepsilon(n)|}(\xi) = \frac{dF_{|\varepsilon(n)|}(\xi)}{d\xi} = \frac{2\xi}{\sigma^2} e^{-\xi^2/\sigma^2} u(\xi). \tag{7.84}$$

Example 7.39. 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 σ^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

$$p_y(x) = \frac{2\xi}{\sigma^2} e^{-\frac{\xi^2}{\sigma^2}} u(\xi)$$

The probability that $y(n) \geq A$ is

$$P\{\xi > A\} = 1 - P\{\xi \leq A\} = 1 - \int_0^A \frac{2\xi}{\sigma^2} e^{-\frac{\xi^2}{\sigma^2}} d\xi = 1 - (1 - e^{-A^2/\sigma^2}) = e^{-\frac{A^2}{\sigma^2}}.$$

For $A = 2$ and $\sigma^2 = 1$ we get $P\{\xi > A\} \approx 0.0183$.

The Rayleigh distribution can be related to the χ-squared distribution, which is obtained as the distribution for the sum of squares of N random Gaussian variables, $x_i(n)$, $i = 1,2,\ldots,N$,

$$z(n) = x_1^2(n) + x_2^2(n) + \cdots + x_N^2(n).$$

The distribution of $z(n) = |\varepsilon(n)|^2$, where $|\varepsilon(n)|$ is the Rayleigh distributed variable, is equal to the χ-squared distribution of $z(n)$ with $N = 2$ (see Example 7.14).

7.4.11 Impulsive Noises

This noise is used to model disturbances when strong impulses occur more often than in the case of Gaussian noise. Due to possible stronger pulses, their probability density function decay toward $\pm\infty$ is slower than in the case of Gaussian noise (a definition of the so called heavy-tailed noise is given in Example 7.15).

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 using

$$\varepsilon(n) = \varepsilon_1(n)\varepsilon_2(n) + \varepsilon_3(n)\varepsilon_4(n)$$

where $\varepsilon_i(n)$, $i = 1,2,3,4$ are real-valued Gaussian independent zero-mean noises, Fig. 7.26 (for the variance of this noise see Problem 7.20).

The parameters of the Laplace distributed signal can be estimated from data, as it is done in Section 7.4.5. For the stationary Laplacian distributed random variable $x(n)$, $\mathbf{x} = [x(1), x(2), \ldots, x(N)]^T$, with mean μ, the likelihood maximization problem is equivalent to the log-likelihood minimization problem again stated as

$$(\alpha, \mu)_{MLE} = \arg\Big\{ \min\{ -\ln\Big(p_{\mathbf{x}}(\mathbf{x}|\alpha,\mu)\Big)\}\Big\}, \tag{7.85}$$

with

$$p_{\mathbf{x}}(\mathbf{x}|\alpha,\mu) = \frac{1}{2\alpha}e^{-|x(1)-\mu|/\alpha} \times \cdots \times \frac{1}{2\alpha}e^{-|x(N)-\mu|/\alpha} = \frac{1}{2^N\alpha^N}e^{-||\mathbf{x}-\mu||_1/\alpha} \tag{7.86}$$

Here, we have to minimize the cost function $-\ln\Big(p_{\mathbf{x}}(\mathbf{x}|\alpha,\mu)\Big)$, defined by

$$J(\alpha,\mu) = N\ln(2) + N\ln(\alpha) + \frac{||\mathbf{x}-\mu||_1}{\alpha}, \tag{7.87}$$

where $||\mathbf{x}-\mu||_1$ is the one-norm (L_1-norm) of vector $\mathbf{x}-\mu$. The solution to the L_1-norm minimization problem is presented in Section 7.1.2,

$$\mu = \text{median}\{\mathbf{x}\}.$$

From $\partial J(\alpha,\mu)/\partial\alpha = 0$ follows

$$\alpha = \frac{1}{N}||\mathbf{x}-\mu\}||_1 = \frac{1}{N}||\mathbf{x}-\text{median}\{\mathbf{x}\}||_1.$$

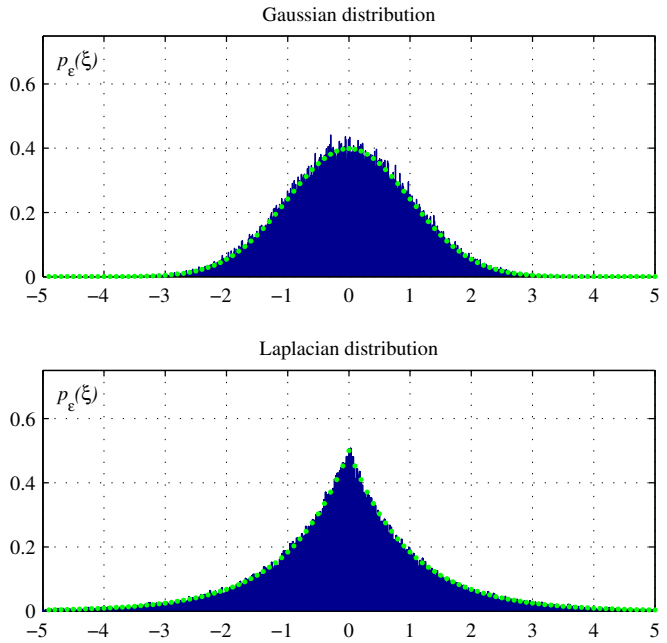


Figure 7.26 The Gaussian and Laplacian noise histograms (with 10000 realizations), with corresponding probability density function (dots).

Example 7.40. A Laplace distributed random signal is simulated as

$$y(n) = x_1(n)x_2(n) + x_3(n)x_4(n) + 1,$$

using $N = 1001$ realizations of the zero-mean Gaussian distributed random variables, $x_1(n)$, $x_2(n)$, $x_3(n)$, and $x_4(n)$, with the same variance $\sigma_x = 1$.

The Laplacian distribution parameters, obtained by minimizing

$$J(\alpha,\mu) = N\ln(2) + N\ln(\alpha) + ||\mathbf{x} - \mu||_1/\alpha,$$

are

$$\mu = \text{median}\{\mathbf{y}\} = 0.98$$

and

$$\alpha = ||\mathbf{y} - \text{median}\{\mathbf{y}\}||_1/N = 1.04,$$

where $\mathbf{y} = [y(1),(y2),\ldots,y(1001)]^T$.

We can also calculate the posterior

$$P(\alpha,\mu|\mathbf{y}) \propto p_{\mathbf{y}}(\mathbf{y}|\alpha,\mu)$$

with

$$p_{\mathbf{y}}(\mathbf{y}|\alpha,\mu) = \frac{1}{(2\alpha)^N} e^{-(|y(1)-\mu|+|y(2)-\mu|+\cdots+|y(N)-\mu|)/\alpha}$$

and present it, for a given N, using discrete sets of α and μ, as in Fig. 7.18.

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 $\varepsilon_2(n)$, that is, as

$$\varepsilon(n) = \frac{\varepsilon_1(n)}{\varepsilon_2(n)}.$$

Another realization of the Cauchy random signal and the definition of the *heavy-tailed* noise are given in Example 7.15.

7.4.12 Poisson Noise

The Poisson noise (or shot noise) is a random signal $\varepsilon(n)$ which can take discrete integer values k with the probability of

$$P(\varepsilon(n) = k) = P(k) = \frac{\lambda^k e^{-\lambda}}{k!} \text{ for } \lambda > 0.$$

The mean value and the variance of $\varepsilon(n)$ are $\mu_\varepsilon = \lambda$ and $\sigma_\varepsilon^2 = \lambda$, respectively (see Problem 7.23). The Poisson random variable is commonly used to model small-probability discrete events. It is typically concerned with the number of events (for example, the number of phone calls in communications or the actual number of particles detected in an image sensor) that occur in a certain (unit) time interval.

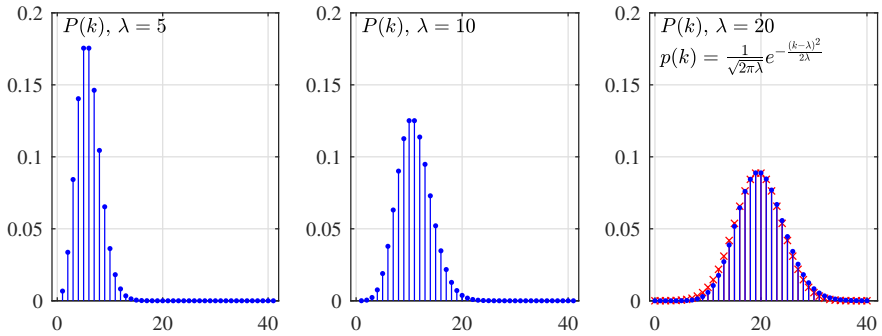


Figure 7.27 Poisson probability for $\lambda = 5$ (left), $\lambda = 10$ (middle), and $\lambda = 20$ (right), along with the Gaussian probability density function (crosses) with the mean value $\mu = \lambda = 20$ and the variance $\sigma^2 = \lambda = 20$.

Example 7.41. Within a long duration, continuous-time signal, an impulsive disturbance appears 15 times per minute, on average. What is the probability that there will be less than 3 impulsive disturbances within a randomly selected continuous-time interval, whose duration is 24 seconds?

★ Since the analyzed interval is 24 seconds, all parameters will be reduced to 24 seconds, as the unit of time. The average number of disturbances within every 24 seconds is $15/60 \times 24 = 6$. This means that the parameter λ in the Poisson distribution is $\lambda = 6$. The probability that there are less than 3 disturbing events in 24 seconds is then equal to the probability that there are either 0 disturbances, $\varepsilon(n) = 0$, or 1 disturbance, $\varepsilon(n) = 1$, or 2 disturbances, $\varepsilon(n) = 2$, within the selected interval, that is

$$P(\varepsilon(n)=0) + P(\varepsilon(n)=1) + P(\varepsilon(n)=2) = \sum_{k=0}^{2} \frac{6^k e^{-6}}{k!} = e^{-6} + 6e^{-6} + \frac{6^2 e^{-6}}{2!} = 0.062.$$

This means that the event of less than 3 disturbances in 24 seconds will occur once in about 16 such intervals.

The probability of, for example, 6 or fewer disturbances in 24 seconds would be 0.6063.

7.4.13 Exponential Random Signal

A random signal with the probability density function

$$p_{x(n)}(\xi) = \frac{1}{\beta} e^{-\xi/\beta} u(\xi)$$

and $\beta > 0$, is called the exponential distributed signal. The expected value of this signal is $\mu_x = \beta$, since

$$\mu_x = \int_0^\infty \frac{\xi}{\beta} e^{-\xi/\beta} d\xi = -\frac{\xi}{\beta} \beta e^{-\xi/\beta} \Big|_0^\infty + \int_0^\infty e^{-\xi/\beta} d\xi = \beta.$$

The variance is $\sigma_x^2 = \beta^2$.

The exponential distributed random variable is often used to model the time elapsed between events that occur randomly over time. The main application area is in studies of the lifetime of systems and components (system reliability and the times between events). The average lifetime for an expectational distributed random variable is β.

The probability distribution of the exponential distributed signal is

$$F_x(\chi) = \int_0^{\chi} \frac{1}{\beta} e^{-\xi/\beta} d\xi = (1 - e^{-\chi/\beta}) u(\chi).$$

The probability that a random variable $x(n)$ will take a value greater than χ is

$$P\{x(n) > \chi\} = 1 - F_x(\chi) = e^{-\chi/\beta} u(\chi).$$

Example 7.42. A random signal $x(n)$ is equal to the length of life of the system denoted by the index n. The average lifetime of this system is 10 years and its life-length is exponentially distributed. What is the probability that the signal value is $x(n) > 20$, meaning that the system n will last more than 20 years?

If the system consists of three components whose life-lengths are statistically independent and exponentially distributed, with average lifetimes $\beta_1 = 5$, $\beta_1 = 10$, and $\beta_3 = 15$ years, respectively, and if the system fails if any of its components fails, what is the average lifetime of the system?

★ The value of the parameter β in the exponential distribution is equal to the expected lifetime, $\beta = 10$. The probability that the system lasts $x(n) > 20$, is

$$P\{x(n) > 20\} = 1 - P\{x(n) \leq 20\} = 1 - F_x(20) = 1 - (1 - e^{-20/10}) = e^{-2} = 0.1353.$$

The probability that the system with three statistically independent components will last longer than χ is equal to the product of the probabilities that each of the components will last longer than χ, that is

$$\begin{aligned} P\{x(n) > \chi\} &= P\{x_1(n) > \chi\} P\{x_2(n) > \chi\} P\{x_3(n) > \chi\} \\ &= (1 - F_{x_1}(\chi))(1 - F_{x_2}(\chi))(1 - F_{x_3}(\chi)) \\ = e^{-\chi/\beta_1} e^{-\chi/\beta_2} e^{-\chi/\beta_3} u(\chi) &= e^{-\chi\left(\frac{1}{\beta_1} + \frac{1}{\beta_2} + \frac{1}{\beta_3}\right)} u(\chi) = e^{-\chi\left(\frac{1}{5} + \frac{1}{10} + \frac{1}{15}\right)} u(\chi) = e^{-\chi/2.7} u(\chi). \end{aligned}$$

The average lifetime is $\beta = 2.7$ and it is shorter than the average life of any of the components.

Example 7.43. Find the Fourier transform, characteristic function, and the moment generating function of the exponentially distributed random variable $x(n)$. What are the moments of this random variable?

★ The Fourier transform of the probability density function of an exponentially distributed random variable is equal to

$$X(\theta) = \int_0^\infty \frac{1}{\beta} e^{-\xi/\beta} e^{-j\theta\xi} d\xi = \frac{1}{1+j\beta\theta}.$$

The characteristic function is equal to

$$\Phi_x(\theta) = X(-\theta) = \frac{1}{1-j\beta\theta},$$

while the moment generating function is related to the Fourier transform as

$$M_x(\theta) = X(-j\theta) = \frac{1}{1-\beta\theta} = 1 + \beta\theta + (\beta\theta)^2 + +(\beta\theta)^3 + \ldots.$$

The moments are $M_1 = \beta$, $M_2 = 2!\beta^2$, $M_3 = 3!\beta^3$, …, $M_N = N!\beta^N$.

The exponential distributed random variable exhibits *memoryless property*, since its probability of exceeding the value $\chi + a$, given that it has exceeded the value a, is

$$P\{x(n) > \chi + a | x(n) > a\} = P\{x(n) > \chi\},$$

regardless of a. The proof is simple using the conditional probability relation

$$P\{x(n) > \chi + a \text{ and } x(n) > a\} = P\{x(n) > \chi + a | x(n) > a\} P\{x(n) > a\}$$

and the fact that $P\{x(n) > \chi + a \text{ and } x(n) > a\} = P\{x(n) > \chi + a\}$, for $\chi \geq 0$, since the event $x(n) > \chi + a$ includes the event $x(n) > a$. These two relations produce

$$P\{x(n) > \chi + a | x(n) > a\} = \frac{P\{x(n) > \chi + a\}}{P\{x(n) > a\}} = \frac{e^{-(\chi+a)/\beta}}{e^{-a/\beta}} = e^{-\chi/\beta} = P\{x(n) > \chi\}.$$

A random variable with the exponential distribution is memoryless because the past does no influence on its future. This means that every instant is like the beginning of a new random period for this random variable. For example, waiting time on the next call in call-center does not depend on the time that has passed since the last call occurred.

7.4.14 Noisy Signals

In real-world scenario, the signals $s(n)$ are commonly corrupted with additive disturbances, denoted by $\varepsilon(n)$. Then, processing methods are applied on the noisy signals,

$$x(n) = s(n) + \varepsilon(n),$$

where $\varepsilon(n)$ is the additive noise. For a deterministic signal $s(n)$, the expected value of the noisy signal $x(n)$ is equal to the sum of the deterministic signal value and the expected value of the noise, that is

$$\mathrm{E}\{x(n)\} = \mathrm{E}\{s(n) + \varepsilon(n)\} = s(n) + \mu_\varepsilon(n).$$

The variance of the noisy signal is not influenced by the deterministic signal,

$$\mathrm{E}\{|x(n) - \mu_\varepsilon(n)|^2\} = \sigma_\varepsilon^2(n).$$

In some application the noise effect is multiplicative and depends on the signal itself. Then, the noisy signal model is

$$x(n) = (1 + \varepsilon(n))s(n).$$

The expected value and the variance of the noisy signal, with multiplicative noise, are given by

$$\mathrm{E}\{x(n)\} = \mathrm{E}\{s(n) + \varepsilon(n)s(n)\} = s(n)(1 + \mu_\varepsilon(n)),$$
$$\mathrm{E}\{|x(n) - \mu_\varepsilon(n)|^2\} = |s(n)|^2\sigma_\varepsilon^2(n).$$

Both the mean and the variance are signal-dependent in the case of multiplicative noise.

Depending on the type of noise, the results obtained so far for various disturbance forms, can be applied to the analysis of noisy signals. This will be the topic of the next sections.

7.5 DISCRETE FOURIER TRANSFORM OF NOISY SIGNALS

In signal processing, the most common signal models are the sinusoidal signals, along with their processing using the Fourier analysis. Influence of noise to this kind of signals and transforms will be studied in this section.

7.5.1 Expected Value and Variance of the DFT

Consider a noisy signal

$$x(n) = s(n) + \varepsilon(n), \tag{7.88}$$

where $s(n)$ is a deterministic useful signal and $\varepsilon(n)$ is an additive noise. The DFT of this signal is

$$X(k) = \sum_{n=0}^{N-1}(s(n) + \varepsilon(n))e^{-j2\pi kn/N} = S(k) + \Xi(k). \tag{7.89}$$

The mean value of $X(k)$ is

$$\mathrm{E}\{X(k)\} = \sum_{n=0}^{N-1} s(n)e^{-j2\pi kn/N} + \sum_{n=0}^{N-1}\mathrm{E}\{\varepsilon(n)\}e^{-j2\pi kn/N} = S(k) + \mathrm{DFT}\{\mu_\varepsilon(n)\}.$$

In the case of a zero-mean noise $\varepsilon(n)$, when $\mu_\varepsilon(n) = 0$, follows

$$\mu_X(k) = \mathrm{E}\{X(k)\} = S(k). \tag{7.90}$$

The variance of $X(k)$, for a zero-mean noise, is

$$\sigma_X^2(k) = \mathrm{E}\{|X(k) - \mu_X(k)|^2\} = \mathrm{E}\{X(k)X^*(k)\} - S(k)S^*(k)$$
$$= \sum_{n_1=0}^{N-1}\sum_{n_2=0}^{N-1}\mathrm{E}\{(s(n_1) + \varepsilon(n_1))(s^*(n_2) + \varepsilon^*(n_2))\}e^{-j2\pi k(n_1-n_2)/N}$$
$$- \sum_{n_1=0}^{N-1}\sum_{n_2=0}^{N-1} s(n_1)s^*(n_2)e^{-j2\pi k(n_1-n_2)/N} = \sum_{n_1=0}^{N-1}\sum_{n_2=0}^{N-1}\mathrm{E}\{\varepsilon(n_1)\varepsilon^*(n_2)\}e^{-j2\pi k(n_1-n_2)/N}. \tag{7.91}$$

For a white noise, with the autocorrelation

$$r_{\varepsilon\varepsilon}(n_1, n_2) = \mathrm{E}\{\varepsilon(n_1)\varepsilon^*(n_2)\} = \sigma_\varepsilon^2\delta(n_1 - n_2),$$

we get

$$\sigma_X^2(k) = \sigma_\varepsilon^2 N. \tag{7.92}$$

If the deterministic signal $s(n)$ is a complex sinusoid, that is

$$s(n) = Ae^{j2\pi k_0 n/N}, \tag{7.93}$$

with the frequency k_0 on the grid, $\omega_0 = 2\pi k_0/N$, then its DFT is

$$S(k) = AN\delta(k - k_0).$$

The *peak signal-to-noise ratio*, being relevant parameter for the DFT based estimation of the signal frequency, is defined by

$$PSNR_{out} = \frac{\max_k |S(k)|^2}{\sigma_X^2} = \frac{A^2N^2}{\sigma_\varepsilon^2 N} = \frac{A^2}{\sigma_\varepsilon^2}N. \tag{7.94}$$

Its logarithmic form, expressed in dB, is $20\log_{10}(AN/\sigma_\varepsilon)$. The value of the peak signal-to-noise ratio increases as N increases. This result is expected, since the signal values are added in phase, increasing the DFT amplitude N times (its power N^2 times), while the noise values are summed up in power.

The noise influence on the DFT of the 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. 7.28.

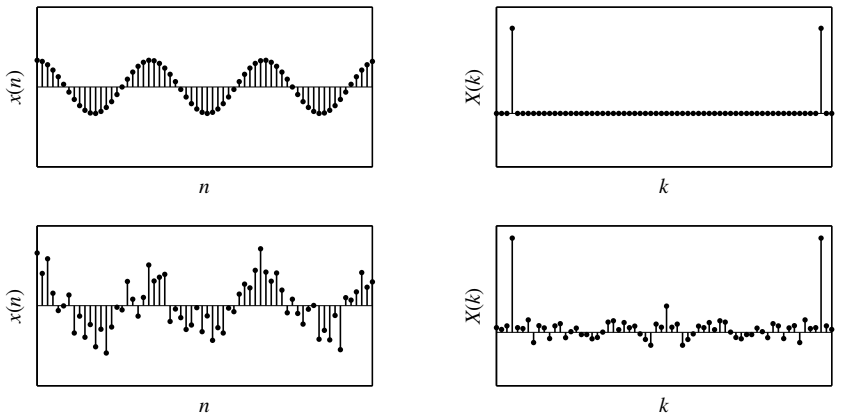


Figure 7.28 Illustration of the noise-free signal, $x(n) = \cos(6\pi n/64)$, and its DFT, Xk (top panels). The same signal is corrupted with an additive zero-mean real-valued Gaussian noise of variance $\sigma_\varepsilon^2 = 1/4$, and shown, along with its DFT (bottom panels).

The *input signal-to-noise ratio* (SNR) for the signal in (7.93) is defined by

$$SNR_{in} = \frac{E_x}{E_\varepsilon} = \frac{\sum_{n=0}^{N-1} |x(n)|^2}{\sum_{n=0}^{N-1} \mathrm{E}\left\{|\varepsilon(n)|^2\right\}} = \frac{NA^2}{N\sigma_\varepsilon^2} = \frac{A^2}{\sigma_\varepsilon^2}. \tag{7.95}$$

If the maximum 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 the 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 (7.89) and (7.92), where $\Xi(k)$ is the noise in the frequency domain, whose variance is equal to $\sigma_\varepsilon^2 N$, we get

$$y(n) = Ae^{j2\pi k_0 n/N} + \frac{\Xi(k_0)}{N}e^{j2\pi k_0 n/N} = x(n) + \varepsilon_X(n).$$

The *output signal-to-noise ratio* is

$$SNR_{out} = \frac{E_x}{E_{\varepsilon_X}} = \frac{\sum_{n=0}^{N-1}|x(n)|^2}{\sum_{n=0}^{N-1}\mathrm{E}\left\{\left|\frac{\Xi(k_0)}{N}e^{j2\pi k_0 n/N}\right|^2\right\}} = \frac{NA^2}{N\frac{N\sigma_\varepsilon^2}{N^2}} = N\frac{A^2}{\sigma_\varepsilon^2} = N \cdot SNR_{in}.$$

Taking $10\log(\circ)$ of both sides we get the output-to-input relation for the signal-to-noise in dB,

$$SNR_{out}[\mathrm{dB}] = 10\log N + SNR_{in}[\mathrm{dB}]. \tag{7.96}$$

7.5.2 Spectral Estimation

In order to improve the representation and estimation performance of the Fourier transform of a noisy signal $s(n) + \varepsilon(n)$, the Fourier transform is commonly calculated using a window function $w(n)$. This topic will be studied again, in detail, in Part V, since the windows play a crucial role in time-frequency analysis. Here, we will present the basic forms and results.

The assumed noise is additive and white, $r_{\varepsilon\varepsilon} = \sigma_\varepsilon^2\delta(n)$, with the zero-mean. The DFT of the signal, multiplied by the window function, is equal to

$$X(k) = \sum_{n=0}^{N-1} w(n)\left[s(n) + \varepsilon(n)\right]e^{-j2\pi kn/N}.$$

The mean value of the windowed DFT is

$$\mu_X(k) = \mathrm{E}\{X(k)\} = \sum_{n=0}^{N-1} w(n)s(n)e^{-j2\pi kn/N} = W(k) *_k S(k),$$

where $W(k) = \mathrm{DFT}\{w(n)\}$ is the DFT of the window and $*_k$ denotes the convolution in frequency.

The variance of $X(k)$ is given by

$$\sigma_X^2(k) = \sum_{n_1=0}^{N-1}\sum_{n_2=0}^{N-1} w(n_1)w^*(n_2)\sigma_\varepsilon^2\delta(n_1 - n_2)e^{-j2\pi k(n_1-n_2)/N} = \sigma_\varepsilon^2\sum_{n=0}^{N-1}|w(n)|^2 = \sigma_\varepsilon^2 E_w, \tag{7.97}$$

where E_w is the energy of the window.

Since we will use mathematical tools that require continuous frequency, consider the Fourier transform of discrete-time noisy signal $x(n) = s(n) + \varepsilon(n)$,

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} w(n)x(n)e^{-j\omega n}, \tag{7.98}$$

where $w(m)$ is a real-valued window, such that $w(0) = 1$. The frequency variable will be kept in continuous form since we will use its derivatives in the explanations that follow. The signal $s(n)$ is deterministic and the noise $\varepsilon(n) = \varepsilon_r(n) + j\varepsilon_i(n)$ is a complex-valued white Gaussian noise with independent and identically distributed real and imaginary parts, $\mathcal{N}(0, \sigma_\varepsilon^2/2)$. The auto-correlation function of this noise is

$$r_{\varepsilon\varepsilon}(m) = \mathrm{E}\{\varepsilon(n)\varepsilon^*(n-m)\} = \sigma_\varepsilon^2\delta(m). \tag{7.99}$$

The expected value of the Fourier transform, for the noisy signal $x(n) = s(n) + \varepsilon(n)$, is

$$\mathrm{E}\{X(e^{j\omega})\} = \mathrm{E}\{\sum_{n=-\infty}^{\infty} w(n)[s(n) + \varepsilon(n)]e^{-j\omega n}\}.$$

Having in mind that $\mathrm{E}\{\varepsilon(n)\} = 0$, follows

$$\mathrm{E}\{X(e^{j\omega})\} = \sum_{n=-\infty}^{\infty} w(n)s(n)e^{-j\omega n}. \tag{7.100}$$

The expected value of the Fourier transform can be written as a convolution of the Fourier transform $W(e^{j\omega})$ of the window $w(n)$,

$$W(e^{j\omega}) = \sum_{n=-\infty}^{\infty} w(n)e^{-j\omega n},$$

and the original Fourier transform, $S(e^{j\omega})$, of the signal $s(n)$, without the window

$$S(e^{j\omega}) = \sum_{n=-\infty}^{\infty} s(n)e^{-j\omega n}.$$

Thus,

$$\mathrm{E}\{X(e^{j\omega})\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(e^{j(\omega-\alpha)})W(e^{j\alpha})d\alpha, \tag{7.101}$$

where the integration is performed over the discrete-time Fourier transform period, $-\pi < \omega \leq \pi$.

7.5.3 Bias in the Fourier Transform of the Windowed Signals

The Fourier transform calculated with a window is biased. The window $w(n)$ causes the bias in the Fourier transform, since its application results in a form that differs from the original Fourier transform without a window. By expanding $S(e^{j(\omega-\alpha)})$ in (7.101) into a Taylor series, around ω,

$$S(e^{j(\omega-\alpha)}) = S(e^{j\omega}) - \frac{\partial S(e^{j\omega})}{\partial\omega}\alpha + \frac{1}{2}\frac{\partial^2 S(e^{j\omega})}{\partial\omega^2}\alpha^2 + \ldots,$$

we get

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} S(e^{j(\omega-\alpha)})W(e^{j\alpha})d\alpha = S(e^{j\omega}) + \frac{1}{2}\frac{\partial^2 S(e^{j\omega})}{\partial\omega^2}m_2 + \ldots, \tag{7.102}$$

where

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} W(e^{j\omega})d\omega = w^2(0) = 1, \quad m_1 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \omega W(e^{j\omega})d\omega = 0, \quad m_2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \omega^2 W(e^{j\omega})d\omega.$$

The first frequency domain moment m_1 (and all other odd moments) of $W(e^{j\omega})$ is equal to zero, since $W(e^{j\omega})$ is an even function (as the Fourier transform of an even, real-valued window function $w(n)$).

From (7.102) follows that the first term is the original Fourier, while the remaining terms introduce the Fourier transform distortion (bias). They can be approximated by

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} S(\omega-\alpha)W(e^{j\alpha})d\alpha - S(e^{j\omega}) = \frac{1}{8}\frac{\partial^2 S(e^{j\omega})}{\partial\omega^2}m_2 + \ldots. \cong \frac{1}{2}\frac{\partial^2 S(e^{j\omega})}{\partial\omega^2}m_2 = \frac{1}{2}b(n,\omega)m_2. \tag{7.103}$$

The bias of the Fourier transform is (approximately)

$$\text{bias}_X(\omega) = \frac{1}{2}b(n,\omega)m_2. \tag{7.104}$$

The Fourier transform bias is highly signal dependent. For the regions where the Fourier transform variations in the frequency direction are small, as described by the second- and higher-order derivatives, this bias is small and vice versa. The signal terms are multiplied by the frequency domain window moments, m_i, in the bias. These moments are small if $W(e^{j\omega})$ is highly concentrated around $\omega = 0$. The bias would be zero if there were no window, that is $W(e^{j\omega}) = 2\pi\delta(\omega)$, $-\pi \le \omega < \pi$. In general, a narrow $W(e^{j\omega})$ requires a wide window $w(n)$ in the discrete-time domain.

7.5.4 Variance in the Fourier transform of Noisy Signals

The Fourier transform variance is defined by

$$\begin{aligned}
\sigma_X^2 &= \mathrm{E}\{X(e^{j\omega})X^*(e^{j\omega})\} - \mathrm{E}\{X(e^{j\omega})\}\mathrm{E}\{X^*(e^{j\omega})\} \\
&= \mathrm{E}\Big\{\Big(\sum_{n=-\infty}^{\infty} w(n)[s(n)+\varepsilon(n)]e^{-j\omega n}\Big)\Big(\sum_{n=-\infty}^{\infty} w(n)[s^*(n)+\varepsilon^*(n)]e^{j\omega n}\Big)\Big\} \\
&-\mathrm{E}\Big\{\sum_{n=-\infty}^{\infty} w(n)[s(n)+\varepsilon(n)]e^{-j\omega n}\Big\}\mathrm{E}\Big\{\sum_{n=-\infty}^{\infty} w(n)[s^*(n)+\varepsilon^*(n)]e^{j\omega n}\Big\} \\
&= \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} w(n)w(m)r_{\varepsilon\varepsilon}(n,m)e^{-j\omega n}e^{j\omega m}.
\end{aligned}$$

A complex-valued Gaussian noise with independent and identically distributed real and imaginary parts, $\mathcal{N}(0,\sigma_\varepsilon^2/2)$) ia assumed. For a white noise, the variance of the Fourier transform estimator reduces to

$$\sigma_X^2 = \sum_{n=-\infty}^{\infty}\sigma_\varepsilon^2 w^2(n) = \sigma_\varepsilon^2 E_w, \tag{7.105}$$

where E_w is the energy of the window. A finite energy window is sufficient to make the variance of $X(e^{j\omega})$ finite for the Gaussian, zero-mean, white noise. We can conclude that the variance increases as the energy of the window, E_w, increases. This means that wide windows will produce big variances, just opposite to the bias which is small for wide windows. Since narrow windows produce large bias and wide windows are characterized by large variances in the Fourier transform estimation, a trade-off is required to balance these two sources of the estimation error.

7.5.5 Bias-to-Variance Trade-Off: Optimum Window Width

The optimum window width can be obtained by minimizing the mean squared error (MSE) defined as a sum of the squared bias and variance

$$e^2 = \text{bias}_X^2(\omega) + \sigma_X^2(\omega). \tag{7.106}$$

Example 7.44. Consider a signal $s(n)$ whose second-order derivative of the Fourier transform is $\partial S(e^{j\omega})/\partial\omega^2$ (higher-order derivatives can be neglected), and the Hann(ing) window $w(n)$ of the width N is used in calculation. Find the optimum window width.

★ For the Hann(ing) window, $E_w = 3N/8$ and $m_2 = 2\pi^2/N^2$, so using (7.103) and (7.105), we get

$$e^2 \cong \frac{\pi^4}{N^4}\left(\frac{\partial^2 S(e^{j\omega})}{\partial\omega^2}\right)^2 + \frac{3N}{8}\sigma_\varepsilon^2. \tag{7.107}$$

It has been assumed that the fourth and other higher-order Fourier transform derivatives can be neglected. From $\partial e^2/\partial N = 0$, the approximation of the optimum window width follows

$$N_{opt}(\omega) \cong \sqrt[5]{\frac{40b^2(\omega)\pi^4}{3\sigma_\varepsilon^2}} \tag{7.108}$$

with $b(\omega) = \partial^2 S(e^{j\omega})/\partial\omega^2$. Roughly speaking, this relation means that small values of the window width (intensive smoothing in frequency direction) should be used at the points where there are no variations in frequency of the Fourier transform, that is, where $b^2(\omega)$ is small. When $b^2(\omega)$ is large, then the window should be wide, meaning less intensive smoothing, that is, keeping the original Fourier transform form, for the points when its variations are high. As far as the noise is concerned, low noise cases (small σ_ε^2) do not require any smoothing of the original Fourier transform in the frequency direction. Thus, wide windows should be used. For a high noise, the Fourier transform smoothing will improve the results.

Of course, in reality, we do not know anything about the signal or its Fourier transform in advance. An algorithm for the estimation of $N_{opt}(\omega)$, without using the value of $b^2(\omega)$, will be presented in the next example.

Example 7.45. The noisy signal

$$x(n) = 2\cos(1.5n) + \sqrt{2}\cos(2.6n) + 150e^{-n^2/2} + 4\varepsilon(n),$$

within $-512 \le n \le 511$, where $\varepsilon(n)$ is the zero-mean, unit-variance Gaussian noise, $\sigma_\varepsilon = 1$, is analyzed using the Fourier transforms, $X_N(k)$, with two Hann(ing) windows, one whose width is $N = 1024$ and the other with $N = 128$. For each frequency index k, we will use better of these two Fourier transforms by checking the confidence intervals intersection.

To simplify the problem, a real-valued and even signal is assumed whose Fourier transform is real-valued. The standard deviation of the real part of the Fourier transform, $X_N(k)$, calculated using the Hann(ing) window of the width N, is $\sigma_{X_N} = \frac{\sigma_\varepsilon}{\sqrt{2}}\sqrt{3N/8}$, while the confidence interval

for the Fourier transform, at a frequency ω with the index k, is

$$[X_N(k) - 2.5\sigma_{X_N}, X_N(k) + 2.5\sigma_{X_N}],$$

where the factor of 2.5 is used for the confidence intervals (probability of almost 0.99), assuming that the noise variance can be estimated from the data. The standard deviation $\sigma_\varepsilon/\sqrt{2}$ was used in σ_{X_N} since the noise was not even and only a half of its power is in the real-valued part of the Fourier transform.

★ For each frequency index k, with the corresponding continuous frequency $\omega = 2\pi k/1024$, the Fourier transform is calculated using $N = 128$, zero-padded up to 1024. This value is denoted by $X_{128}(k)$. Then the Fourier transform with $N = 1024$ is calculated and denoted by $X_{1024}(k)$. The confidence intervals are formed for these two Fourier transform values calculated with two window widths,

$$\left[X_{128}(k) - 7.1\sqrt{\frac{3}{8}128}, \quad X_{128}(k) + 7.1\sqrt{\frac{3}{8}128}\right]$$
$$\left[X_{1024}(k) - 7.1\sqrt{\frac{3}{8}1024}, \quad X_{124}(k) + 7.1\sqrt{\frac{3}{8}1024}\right].$$

If these intervals intersect, then $X(k) = X_{128}(k)$, otherwise $X(k) = X_{1024}(k)$. Namely, if the bias is small, then the Fourier transform $X_N(k)$ calculated using both windows will contain the true value of the Fourier transform (of the noise-free signal). Therefore, for small bias the confidence intervals will intersect, meaning we should use the window with a smaller variance, which is in our experiment $N = 128$. If the bias is large, then it will highly depend on the window width and will move the obtained Fourier transform $X_N(k)$ from its true position. Then, the confidence intervals will be dominated by the bias (different for two windows) and will not contain the true Fourier transform value, meaning that they will not intersect. Since the bias is large, in this case, we should use a small bias window with $N = 1024$. The result is shown in Fig. 7.29. The improvement in the SNR ratio is evident.

This is a simplified version of the intersection of confidence intervals (ICI) method to the window width optimization (Katkovnik-Stankovic method for the window width optimization in time-frequency analysis). For practical applications, the noise variance should also be estimated from the data (see Problem 7.12).

Calculation of higher-order moments and the cross-correlation functions for the Fourier transform of noisy signals could be found in the literature (for the correlation calculation, see the problems).

7.5.6 Periodogram

The power spectral density of signal is commonly estimated using the squared absolute value of the Fourier transform of the signal, called periodogram,

$$P_x(e^{j\omega}) = \frac{1}{2N+1}\left|X(e^{j\omega})\right|^2 = \frac{1}{2N+1}\left|\sum_{n=-N}^{N} x(n)e^{-j\omega n}\right|^2. \tag{7.109}$$

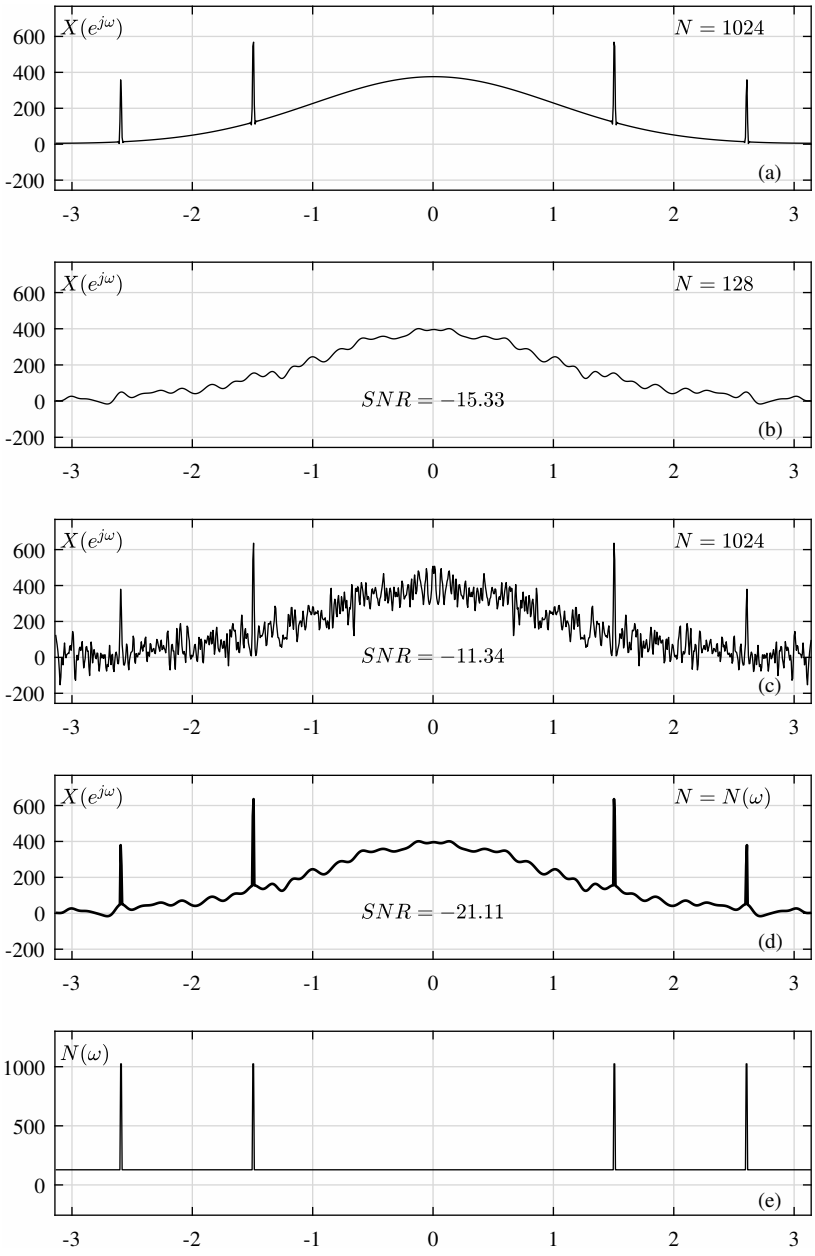


Figure 7.29 Spectral analysis of a signal with two windows in order to approximate optimal window width for Example 7.45.

The periodogram is used as an estimate of the power spectral density,

$$S_{xx}(e^{j\omega}) = \lim_{N\to\infty} \sum_{k=-2N}^{2N} r_{xx}(k)e^{-j\omega k} = \mathrm{FT}\{r_{xx}(n)\} = S_{xx}(e^{j\omega}).$$

As it has been shown in Section 7.3.4, the periodogram is equal to the power spectral density calculated (windowed) by a Bartlett window, that is

$$P_{xx}(e^{j\omega}) = \lim_{N\to\infty} \sum_{k=-2N}^{2N} (1 - \frac{|k|}{2N+1}) r_{xx}(k)e^{-j\omega k} = S_{xx}(e^{j\omega}) *_\omega W(e^{j\omega}), \tag{7.110}$$

where $W(e^{j\omega}) = \mathrm{FT}\{(1 - \frac{|n|}{2N+1})\}$. This means that the periodogram is a biased estimate of the power spectral density, for any signal, except $r_{xx}(k) = C\delta(k)$.

Example 7.46. Find the power spectral density of the random signal

$$x(n) = \varepsilon(n) + 0.5\varepsilon(n-4),$$

where $\varepsilon(n)$ is zero-mean Gaussian noise with unite variance. Find the power spectral density calculated using the periodogram with a window of the width N.

★ The autocorrelation function of this signal is

$$r_{xx}(n) = (1 + 0.5^2)\delta(n) + 0.5\delta(n+4) + 0.5\delta(n-4)$$

Its power spectral density is

$$S_{xx}(\omega) = \mathrm{FT}\{r_{xx}(n)\} = \sum_{n=-\infty}^{\infty} (1.25\delta(n) + 0.5\delta(n+4) + 0.5\delta(n-4))e^{-j\omega n} = 1.25 + \cos(4\omega).$$

If the periodogram is used, with a window whose width is N, then

$$P_{xx}(e^{j\omega}) = \mathrm{E}\{\frac{1}{N}\left|X(e^{j\omega})\right|^2\} = \frac{1}{N}\mathrm{E}\{\left|\sum_{n=-N/2}^{N/2-1} x(n)e^{-j\omega n}\right|^2\} = \sum_{k=-N/2}^{N/2-1} (1 - \frac{|k|}{N}) r_{xx}(k)e^{-j\omega k}$$

$$= \sum_{k=-N/2}^{N/2-1} r_{xx}(k)e^{-j\omega k} - \sum_{k=-N/2}^{N/2-1} \frac{|k|}{N} r_{xx}(k)e^{-j\omega k} = 1.25 + \cos(4\omega) - \frac{4}{N}\cos(4\omega).$$

The bias term is $P_{xx}(e^{j\omega}) - S_{xx}(\omega) = -\frac{4}{N}\cos(4\omega)$.

]

The periodogram of a noisy signal is also a biased estimator of the noise-free periodogram of deterministic signals. Consider the signal $x(n) = s(n) + \varepsilon(n)$, where $s(n)$ is deterministic and $\varepsilon(n)$ is white complex-valued i.i.d. noise with the variance σ_ε^2. Its periodogram is

$$P_x(e^{j\omega}) = \frac{1}{N}\left|\sum_{n=-N/2}^{N/2-1} (s(n) + \varepsilon(n))e^{-j\omega n}\right|^2. \tag{7.111}$$

The expected value of this periodogram is

$$P_{xx}(e^{j\omega}) = \frac{1}{N}\mathrm{E}\{\left|\sum_{n=-N/2}^{N/2-1}(s(n)+\varepsilon(n))e^{-j\omega n}\right|^2\} = \frac{1}{N}\left|\sum_{n=-N/2}^{N/2-1}s(n)e^{-j\omega n}\right|^2 + \sigma_\varepsilon^2. \quad (7.112)$$

The bias is equal to σ_ε^2.

The variance of this estimator can be calculated as well (see Exercise 7.11 and relation (7.161)),

$$\mathrm{Var}\{P_{xx}(e^{j\omega})\} = 2P_s(e^{j\omega})\sigma_\varepsilon^2 + \sigma_\varepsilon^4. \quad (7.113)$$

The variance consists of two parts:

(1) σ_ε^4, which is constant, and

(2) $2P_s(e^{j\omega})\sigma_\varepsilon^2$, being signal-dependent.

Since the variance of periodogram, $\mathrm{Var}\{P_{xx}(e^{j\omega})\}$, is proportional to $P_s(e^{j\omega})$, its highest value is achieved at the maximum of the noise-free signal periodogram, $P_s(e^{j\omega})$, as illustrated in Fig. 7.30 for the noisy chirp signal

$$x(n) = e^{-(n/256)^2}e^{j\frac{\pi}{2}n^2/N} + \frac{1}{4}\varepsilon(n),$$

where $-128 \leq n \leq 127$, and $\varepsilon(n)$ is the zero-mean Gaussian noise with a unit variance.

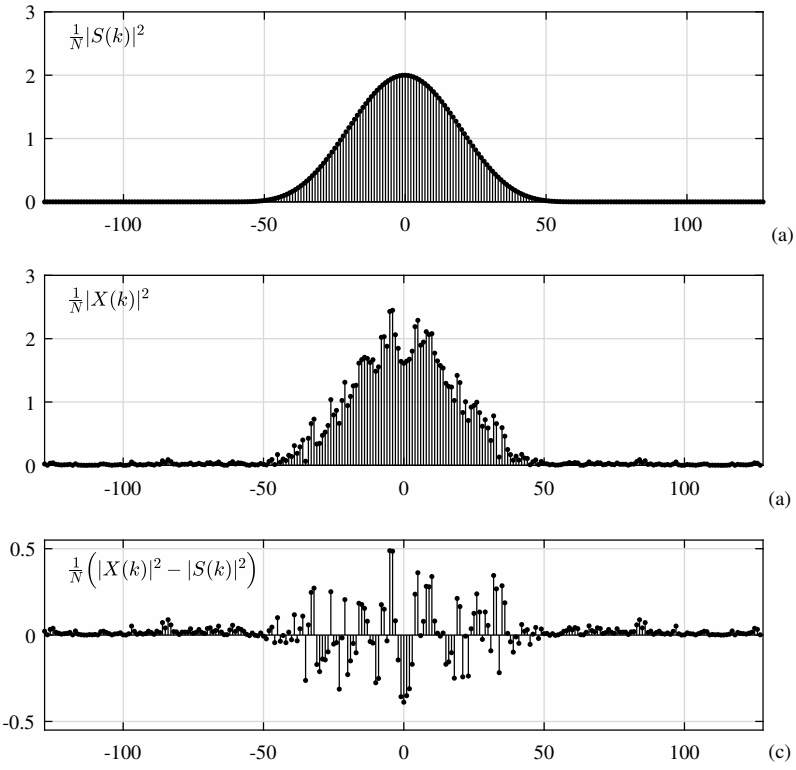


Figure 7.30 Periodogram of a chirp signal (a), chirp-noisy signal (b), and the difference of the previous two periodograms, being highly signal dependent, with variations (and variance) proportional to $|S(k)|^2/N$, (c).

Blackman–Tukey spectral estimator is obtained from (7.110), using a general window

$$P_{xx}(e^{j\omega}) = \sum_{k=-N}^{N} w(k) r_{xx}(k) e^{-j\omega k}. \tag{7.114}$$

The window, $w(k)$, decays smoothly from $w(0) = 1$ toward zero for $k = \pm N$. The frequency domain form of this estimator is equal to the convolution of the true power spectral density and the Fourier transform of the window, $W(e^{j\omega}) = \mathrm{FT}\{w(n)\}$. In the discrete frequency domain, the Blackman–Tukey periodogram can be calculated using

$$P_{xx}(k) = \sum_i W(k-i)\mathrm{DFT}\{r_{xx}(k)\},$$

where $r_{xx}(k)$ is estimated using the standard unbiased autocorrelation estimator

$$\hat{r}_{xx}(\pm k) = \frac{1}{N-k} \sum_{i=0}^{N-k-1} x(k+i)x(i)$$

or the standard biased autocorrelation estimator

$$\hat{r}_{xx}(\pm k) = \frac{1}{N} \sum_{i=0}^{N-k-1} x(k+i)x(i).$$

The biased estimator under-estimates $r_{xx}(k)$ values for large $|k|$, however, they should be small anyway. This estimator avoids possible large outliers in estimating $r_{xx}(k)$ from a small number of samples, for large $|k|$.

Daniell periodogram. In order *to reduce the noise influence*, the smoothed versions of the periodogram are used as the spectral estimators. The simplest smoothed form of the periodogram is

$$P_x^S(k) = \frac{1}{2L+1} \sum_{i=-L}^{L} P_x(k-i) = \frac{1}{2L+1} \sum_{i=-L}^{L} \frac{1}{N} |X(k-i)|^2.$$

Here, the frequency domain window, $W(k)$, takes the simplest possible form of the rectangular window in the Blackman-Tukey method, where $W(i)|X(k-i)|^2/N$ was used. Therefore, *the Daniell spectral estimator is a particular case of the Blackman–Tukey class of spectral estimators*. It can easily be related to the Blackman–Tukey periodogram estimator (7.110) using

$$S_{xx}^A(e^{j\omega}) = \sum_{n=-N/2}^{N/2-1} w(n)\left(1 - \frac{|n|}{N}\right) r_{xx}(n) e^{-j\omega n} = W(k) *_k S_{xx}(k),$$

where the smoothing window in the frequency domain is the Fourier transform of the auto-correlation function window, $w(n)(1 - \frac{|n|}{N})$, and corresponds to

$$P_x^S(k) = \frac{1}{2L+1} \sum_{i=-L}^{L} W(i) P_x(k-i) = \frac{1}{2L+1} \sum_{i=-L}^{L} \frac{1}{N} W(i) |X(k-i)|^2. \tag{7.115}$$

S-method. In the analysis of signals with varying spectral content, the Fourier transform is spread due to the frequency variations of the spectral content within the window (see the stationary phase method in Chapter 1). Then, instead of smoothing the periodogram in the same direction (in-direction smoothing)

in (7.115), *the counter-direction cross-multiplication* can be done, and the spectral estimator

$$SM_x(k) = \frac{1}{2L+1}\sum_{i=-L}^{L}\frac{1}{N}W(i)X(k+i)X^*(k-i)$$

is obtained. This is the so-called S-method based spectral estimator. By increasing the width of L in this method, we could arrive at the Wigner distribution.

Example 7.47. Estimate the power spectral density of the signal

$$x(n) = \cos(200\pi n/N) + \sqrt{2}e^{-(n/64)^2}e^{j\frac{\pi}{4}n^2/N} + \varepsilon(n),$$

where $-128 \leq n \leq 127$, and $\varepsilon(n)$ is a zero-mean Gaussian noise with the unit variance. Use the periodogram with a Hann(ing) window of the width $N = 256$, the Daniell (Blackman-Tukey smoothed) estimator, and the S-method, with the same window. In both, the Blackman-Tukey estimator and the S-method estimator, use $L = 7$ and $W(i) = 1$.

★ Spectral analysis of this random noisy signal using the periodogram, the Daniell (Blackman-Tukey smoothed) estimator, and the S-method based estimator is shown in Fig. 7.31. The periodogram of the noise-free signal is shown in Fig. 7.31(a). Two highly concentrated sinusoidal components and one spread (chirp) component can be noticed. For the noisy signal, the noise almost completely degrades the chirp component in the periodogram with the Hann(ing) window, Fig. 7.31(b). The visibility of this component is significantly improved by smoothing the periodogram as in the Daniell (Blackman-Tukey smoothed) estimator, given in Fig. 7.31(c). In this case, the highly concentrated sinusoidal components are spread as well. Combining the Fourier transform values in the counter-direction, the S-method based spectral estimation is obtained. This estimator preserves a high concentration of the sinusoidal components while improving the concentration of the chirp signal, as shown in Fig. 7.31(d). The S-method based spectral estimator of the noise-free signal is given in Fig. 7.31(e).

]

Bartlett Method and Welch periodogram. The Fourier transform of the signal $x(n)$, whose duration is N, is calculated here over K shorter intervals. The duration of these intervals is M, commonly with the step $R = M$ (Daniell periodogram) or $R = M/2$ (Welch periodogram, in this case a window can also be used). The Fourier transforms of $x(n)$, within these shorter intervals, are

$$X_i(e^{j\omega}) = \frac{1}{M}\sum_{n=0}^{M-1}x(iR+n)e^{-j\omega n}$$

for $i = 0,1,\ldots,K$. The power spectral densities $|X_i(e^{j\omega})|^2$ are averaged to produce

$$P_x^S(\omega) = \frac{1}{K}\sum_{i=0}^{K-1}\left|X_i(e^{j\omega})\right|^2.$$

For a numeric illustration of the Welch periodogram calculation see Example 7.52.

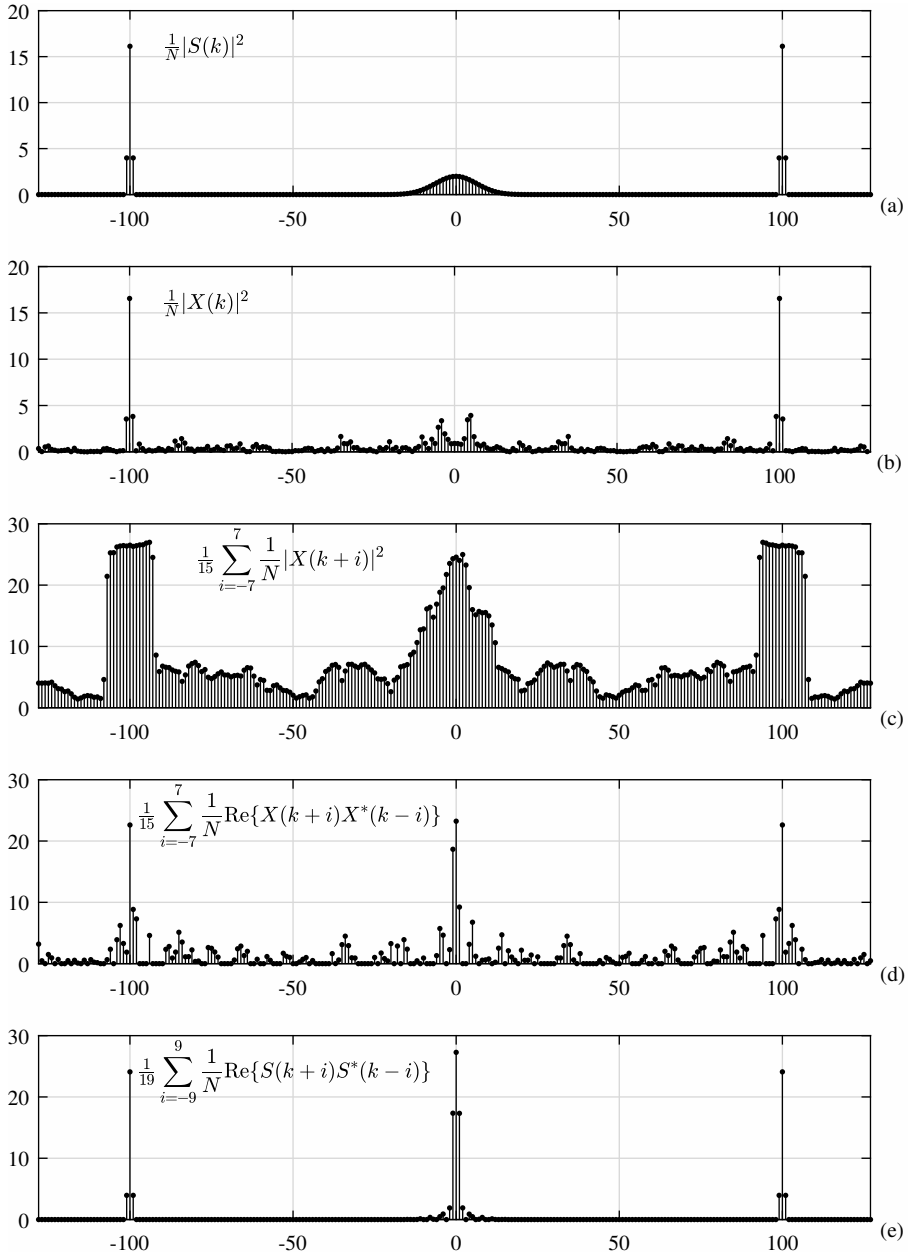


Figure 7.31 Spectral analysis of the random noisy signal, $x(n)$, using the periodogram, the Daniell (Blackman-Tukey smoothed) estimator, and the S-method based spectral estimator. Order and the description of the panels correspond to the task order in Example 7.47.

7.5.7 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 the signal

$$s(n) = Ae^{j2\pi k_0 n/N}.$$

The additive noise $\varepsilon(n)$ is white, complex-valued Gaussian, with zero-mean and independent real and imaginary parts. The variance of noise is σ_ε^2. 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

$$\boldsymbol{\alpha} = \{b,k\}$$

is the set of these parameters. The parameter b is complex-valued. It includes the amplitude and the initial phase of the signal model. For every value of $x(n)$ we may define an error as a difference of the true value $x(n)$ and the assumed model, at the considered instant n,

$$e(n,\boldsymbol{\alpha}) = x(n) - be^{j2\pi kn/N}. \tag{7.116}$$

Since the noise is Gaussian, the probability density function of the error is

$$p(e(n,\boldsymbol{\alpha})) = \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} e^{-|e(n,\boldsymbol{\alpha})|^2/(2\sigma_\varepsilon^2)}.$$

The joint probability density function, for all signal samples from the data set, is equal to the product of the individual probability density functions

$$p_e(e(0,\boldsymbol{\alpha}),e(1,\boldsymbol{\alpha}),\ldots,e(N-1,\boldsymbol{\alpha})) = \frac{1}{(2\pi\sigma_\varepsilon^2)^{N/2}} e^{-\sum_{n=0}^{N-1}|e(n,\boldsymbol{\alpha})|^2/(2\sigma_\varepsilon^2)}.$$

The maximum-likelihood solution for the parameters $\boldsymbol{\alpha} = \{b,k\}$ in obtained by maximizing the probability density function for given values of $x(n)$. Maximization of $p_e(e(0,\boldsymbol{\alpha}),e(1,\boldsymbol{\alpha}),\ldots,e(N-1,\boldsymbol{\alpha}))$ is the same as the minimization of the total squared error,

$$\epsilon(\boldsymbol{\alpha}) = \sum_{n=0}^{N-1} |e(n,\alpha)|^2 = \sum_{n=0}^{N-1} \left| x(n) - be^{j2\pi kn/N} \right|^2. \tag{7.117}$$

The solution to this problem is obtained from $\partial\epsilon(\boldsymbol{\alpha})/\partial b^* = 0$ (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 $\epsilon(\boldsymbol{\alpha})$ and gives the estimate of the signal frequency index k_0, is obtained by replacing the obtained b back into relation (7.117), defining $\epsilon(\boldsymbol{\alpha})$,

$$\epsilon(\boldsymbol{\alpha}) = \sum_{n=0}^{N-1} |x(n) - be^{j2\pi kn/N}|^2 = \left(\sum_{n=0}^{N-1} |x(n)|^2 \right) - N|b|^2.$$

The minimum value of $\epsilon(\boldsymbol{\alpha})$ is achieved when $|b|^2$ (or $|X(k)|^2$) is maximum,

$$\hat{k}_0 = \arg\{\max|X(k)|^2\} = \arg\{\max|X(k)|\}.$$

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 the signal

$$s(n) = Ae^{j\omega_0 n}.$$

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, impulsive with the Laplacian distribution, then the probability density function would be

$$p(e(n,\boldsymbol{\alpha})) = \frac{1}{2\sigma_\varepsilon} e^{-|e(n,\boldsymbol{\alpha})|/\sigma_\varepsilon},$$

and the solution to $\epsilon(\boldsymbol{\alpha}) = \sum_{n=0}^{N-1} |e(n,\boldsymbol{\alpha})|$ minimization would follow from

$$X(k) = N \underset{n=0,1,\ldots,N-1}{\text{median}} \left\{ x(n)e^{-j2\pi kn/N} \right\}.$$

Note that the absolute value of error can be written as

$$|e(n,\boldsymbol{\alpha})| = \left| x(n) - be^{j2\pi kn/N} \right| = \left| x(n)e^{-j2\pi kn/N} - b \right|.$$

Minimization of a sum of this kind of terms is discussed in Part VI.

Example 7.48. The DFT definition, for a given frequency index k, can be understood as

$$\begin{aligned} X(k) &= \sum_{n=0}^{N-1} (s(n) + \varepsilon(n))e^{-j2\pi kn/N} \\ &= N \underset{n=0,1,\ldots,N-1}{\text{mean}} \left\{ (s(n) + \varepsilon(n))e^{-j2\pi kn/N} \right\} \end{aligned} \tag{7.118}$$

Based on the definition of median, discuss when the DFT estimation

$$\begin{aligned} X_R(k) = N \underset{n=0,1,\ldots,N-1}{\text{median}} &\text{Re}\left\{ (s(n) + \varepsilon(n))e^{-j2\pi kn/N} \right\} \\ + jN \underset{n=0,1,\ldots,N-1}{\text{median}} &\text{Im}\left\{ (s(n) + \varepsilon(n))e^{-j2\pi kn/N} \right\} \end{aligned} \tag{7.119}$$

can produce better results than (7.118). Calculate the value of $X(0)$ using (7.118) and estimate its value by (7.119) for the signal

$$s(n) = \exp(j4\pi n/N)$$

with $N = 8$, and the additive noise

$$\varepsilon(n) = 2001\delta(n) - 204\delta(n-3).$$

Which of these two estimates is closer to the noise-free DFT value?

★If a strong impulsive noise is expected in the signal, then the mean value will be highly sensitive to this kind of noise. As it is stated, the median based calculation is less sensitive to strong impulsive disturbances. For the signal

$$s(n) = \exp(j\pi n/2) = [1, j, -1, -j, 1, j, -1, -j]$$

and the given noise $\varepsilon(n)$, the value of $X(0)$ is equal to

$$X(0) = \sum_{n=0}^{7} (s(n) + \varepsilon(n)) = 0 + 2001 - 204 = 805.$$

The median-based estimation is

$$\begin{aligned} X_R(0) &= 8\text{median}\{2002,\ 0, -1, -204,\ 1,\ 0, -1,\ 0\} \\ &\quad + j8\text{median}\{0,\ 1,\ 0, -1,\ 0,\ 1,\ 0, -1\} = 0 + j0. \end{aligned} \tag{7.120}$$

Obviously the median-based estimate is not influenced by this impulsive noise. In this case it produced better estimate (the exact value) of the considered noise-free DFT element $X(0)$.

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 the signal frequency is assumed on the frequency grid, this case can be understood as the signal frequency position detection. Available observations of the signal are

$$x(n) = s(n) + \varepsilon(n), \text{ for } 0 \le n \le N-1,$$

where $\varepsilon(n)$ is a complex zero-mean i.i.d. Gaussian white noise, with variance σ_ε^2. Its DFT is

$$X(k) = \sum_{n=0}^{N-1} (s(n) + \varepsilon(n)) e^{-j2\pi kn/N} = NA\delta(k - k_0) + \Xi(k),$$

with $\sigma_X^2(k) = \sigma_\varepsilon^2 N$ and $\mathrm{E}\{\Xi(k)\} = 0$. The real and imaginary parts of the DFT $X(k_0)$, at the signal position $k = k_0$, are the Gaussian random variables, with the total variance $\sigma_\varepsilon^2 N$, or

$$\mathcal{N}(NA, \sigma_\varepsilon^2 N/2), \qquad \mathcal{N}(0, \sigma_\varepsilon^2 N/2), \tag{7.121}$$

respectively, where a real-valued A is assumed without any loss of generality.

The real and imaginary part of the DFT of noise only, $X(k) = \Xi(k)$, for $k \neq k_0$ are zero-mean random variables with the same variance

$$\mathcal{N}(0, \sigma_\varepsilon^2 N/2).$$

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, $k \neq k_0$, is Rayleigh-distributed (7.84)

$$q(\xi) = \frac{2\xi}{\sigma_\varepsilon^2 N} e^{-\xi^2/(\sigma_\varepsilon^2 N)}, \ \ \xi \ge 0.$$

The DFT at a noise only position takes a value greater than χ, with probability

$$Q(\chi) = \int_{\chi}^{\infty} \frac{2\xi}{\sigma_\varepsilon^2 N} e^{-\xi^2/(\sigma_\varepsilon^2 N)} d\xi = \exp(-\frac{\chi^2}{\sigma_\varepsilon^2 N}). \tag{7.122}$$

The probability that a DFT of noise only is lower than χ is $[1 - Q(\chi)]$. 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 χ is $[1-Q(\chi)]^M$. Probability that at least one of M DFT noise only values is greater than χ, is

$$G(\chi) = 1 - [1 - Q(\chi)]^M. \tag{7.123}$$

The probability density function for the absolute DFT values at the position of the signal (whose real and imaginary parts are described by (7.121)) is Rice-distributed

$$p(\xi) = \frac{2\xi}{\sigma_\varepsilon^2 N} e^{-(\xi^2+N^2A^2)/(\sigma_\varepsilon^2 N)} I_0(2NA\xi/(\sigma_\varepsilon^2 N)),\ \xi \geq 0, \tag{7.124}$$

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 the estimation occurs. To calculate this probability, consider the absolute DFT value of a signal at and around ξ. The DFT value at the signal position is within ξ and $\xi + d\xi$ with the probability $p(\xi)d\xi$, where $p(\xi)$ is defined by (7.124). The probability that at least one of M DFT noise only values is above ξ in amplitude, is equal to

$$G(\xi) = 1 - [1 - Q(\xi)]^M.$$

Thus, the probability that the absolute value of the DFT of the signal component is within ξ and $\xi + d\xi$ and that at least one of the absolute DFT noise only values exceeds the DFT signal value is equal to $G(\xi)p(\xi)d\xi$. Considering all possible values of ξ, from (7.122) and (7.123), the probability of the wrong signal frequency detection follows as

$$\begin{aligned} P_E &= \int_0^\infty G(\xi)p(\xi)d\xi = \int_0^\infty \left(1 - \left[1 - \exp(-\frac{\xi^2}{\sigma_\varepsilon^2 N})\right]^M\right) \\ &\quad \times \frac{2\xi}{\sigma_\varepsilon^2 N} e^{-(\xi^2+N^2A^2)/(\sigma_\varepsilon^2 N)} I_0(2NA\xi/(\sigma_\varepsilon^2 N))d\xi. \end{aligned} \tag{7.125}$$

An approximation of this expression can be obtained 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),

$$P_E \cong 1 - \left[1 - \exp(-\frac{NA^2}{\sigma_\varepsilon^2 N})\right]^M. \tag{7.126}$$

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.

7.6 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 expected value of the output signal $y(n)$ is given by

$$\mu_y(n) = \mathrm{E}\{y(n)\} = \sum_{k=-\infty}^{\infty} h(k)\mathrm{E}\{x(n-k)\} = \sum_{k=-\infty}^{\infty} h(k)\mu_x(n-k) = h(n) *_n \mu_x(n). \tag{7.127}$$

For a stationary input signal the expected value is in the form

$$\mu_y = \mu_x \sum_{k=-\infty}^{\infty} h(k) = \mu_x H(e^{j0}). \tag{7.128}$$

The cross-correlation of the output and input signal is

$$r_{yx}(n,m) = \mathrm{E}\{y(n)x^*(m)\} = \sum_{k=-\infty}^{\infty} \mathrm{E}\{x(k)x^*(m)\}h(n-k) = \sum_{k=-\infty}^{\infty} r_{xx}(k,m)h(n-k). \tag{7.129}$$

For a stationary input signal, with changes of indices $n-m=l$ and $k-m=p$, we get

$$r_{yx}(l) = \sum_{p=-\infty}^{\infty} r_{xx}(p)h(l-p) = r_{xx}(l) *_l h(l).$$

The cross-correlation of the input signal, $x(n)$, and the output signal, $y(n)$, is obtained as the convolution of the input signal autocorrelation and the impulse response, $h(n)$. The z-transform of both sides of this equation gives

$$R_{yx}(z) = R_{xx}(z)H(z).$$

The cross-correlation of the input signal and the output signal is

$$r_{xy}(n,m) = \mathrm{E}\{x(n)y^*(m)\} = \sum_{k=-\infty}^{\infty} \mathrm{E}\{x(n)x^*(k)\}h^*(m-k) = \sum_{k=-\infty}^{\infty} r_{xx}(n,k)h^*(m-k).$$

For a stationary signal, with $n-m=l$ and $n-k=p$, we get that the cross-correlation of the input signal and the output signal is equal to the convolution of the input signal autocorrelation and the reversed and conjugated impulse response, that is

$$r_{xy}(l) = \sum_{p=-\infty}^{\infty} r_{xx}(p)h^*(p-l) = r_{xx}(l) *_l h^*(-l).$$

The z-transform of both sides maps this equation into the z-transform domain

$$\sum_{l=-\infty}^{\infty} r_{xy}(l)z^{-l} = \sum_{l=-\infty}^{\infty}\sum_{p=-\infty}^{\infty} r_{xx}(p)h^*(p-l)z^{-l} = \sum_{k=-\infty}^{\infty}\sum_{p=-\infty}^{\infty} r_{xx}(p)h^*(k)z^{-p}\left(z^{-1}\right)^{-k}$$

$$R_{xy}(z) = R_{xx}(z)H^*(\frac{1}{z^*}).$$

The Fourier transform form the last equation is of the form

$$S_{xy}(e^{j\omega}) = S_{xx}(e^{j\omega})H^*(e^{j\omega}). \tag{7.130}$$

Similarly, the autocorrelation of the output signal $y(n)$ is defined by

$$r_{yy}(n,m) = \mathrm{E}\{y(n)y^*(m)\} = \sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} \mathrm{E}\{x(l)x^*(k)\}h(n-l)h^*(m-k). \tag{7.131}$$

After some straightforward transformations, we get the z-transform domain relation between the autocorrelations of the stationary input and the output signal

$$R_{yy}(z) = R_{xx}(z)H(z)H^*(\frac{1}{z^*}).$$

The Fourier transform of output signal autocorrelation function in terms of the Fourier transform of the input signal autocorrelation function and the system frequency response is given by

$$S_{yy}(e^{j\omega}) = S_{xx}(e^{j\omega})\left|H(e^{j\omega})\right|^2, \tag{7.132}$$

proving that $S_{xx}(e^{j\omega})$ is indeed the power density function. By taking a narrow-pass filter with the unit amplitude $\left|H(e^{j\omega})\right|^2 = 1$ for $\omega_0 \leq \omega < \omega_0 + d\omega$, we will get the spectral density of the signal $x(n)$ within that small frequency range.

Example 7.49. The linear time-invariant system is defined by

$$y(n) = x(n) + ax(n-1) + a^2x(n-2).$$

The input signal is a zero-mean white noise $\varepsilon(n)$ with the variance σ_ε^2. Find the cross-correlation of the input signal and the output signal and the autocorrelation of the output signal. For $a = -1$ find the power spectral density of the output signal.

★The system transfer function is

$$H(z) = 1 + az^{-1} + a^2z^{-2}.$$

Since the input signal is a white noise of variance σ_ε^2 its autocorrelation is, by definition,

$$r_{xx}(n) = r_{\varepsilon\varepsilon}(n) = \sigma_\varepsilon^2\delta(n).$$

The power spectral density of the input signal is obtained as the Fourier transform of the autocorrelation function, that is

$$S_{xx}(\omega) = \sum_{n=-\infty}^{\infty} r_{xx}(n)e^{-j\omega n} = \sigma_\varepsilon^2.$$

The z-transform of the autocorrelation function of the input signal is obtained as

$$R_{xx}(z) = \sum_{n=-\infty}^{\infty} r_{xx}(n)z^{-n} = \sigma_\varepsilon^2 \sum_{n=-\infty}^{\infty} \delta(n)z^{-n} = \sigma_\varepsilon^2.$$

The z-transform of the autocorrelation function of the output signal, for the linear time-invariant system, is equal to

$$R_{yy}(z) = R_{xx}(z)H(z)H^*(1/z^*)$$
$$= \sigma_\varepsilon^2\left[1 + a^2 + a^4 + a(1+a^2)(z+z^{-1}) + a^2(z^2+z^{-2})\right].$$

The autocorrelation function of the output signal is equal to the inverse z-transform of $R_{yy}(z)$,

$$r_{yy}(n) = \sigma_\varepsilon^2(1+a^2+a^4)\delta(n) + \sigma_\varepsilon^2 a(1+a^2)(\delta(n+1)+\delta(n-1)) + \sigma_\varepsilon^2 a^2(\delta(n+2)+\delta(n-2)).$$

The power spectral density of the output signal is

$$S_{yy}(\omega) = R_{yy}(e^{j\omega}) = \sigma_\varepsilon^2(1+a^2+a^4+2a(1+a^2)\cos\omega + 2a^2\cos(2\omega)),$$

while the z-transform of the cross-correlation of the input and output signal is

$$R_{yx}(z) = H(z)R_{xx}(z) = (1+az^{-1}+a^2z^{-2})\sigma_\varepsilon^2.$$

Its inverse z-transform of $R_{yx}(z)$ is equal to the cross-correlation, $r_{yx}(n)$,

$$r_{yx}(n) = \sigma_\varepsilon^2(\delta(n)+a\delta(n-1)+a^2\delta(n-2)).$$

For $a=-1$ the power spectral density function of the output signal is

$$S_{yy}(\omega) = \sigma_\varepsilon^2(3-4\cos\omega+2\cos(2\omega)) = \sigma_\varepsilon^2(1-4\cos\omega+4\cos^2\omega) = \sigma_\varepsilon^2(1-2\cos\omega)^2.$$

Example 7.50. For the discrete-time system defined by

$$y(n) - 1.3y(n-1) + 0.36y(n-2) = x(n)$$

with the random input signal $x(n)=\varepsilon(n)$, $\mu_\varepsilon = 0$ and $r_{\varepsilon\varepsilon}(n)=\delta(n)$, find:
(a) The expected value $\mu_y(n)$ and the autocorrelation $r_{yy}(n)$ of the output signal,
(b) The power spectral density functions $S_{yy}(\omega)$ and $S_{yx}(\omega)$.

★(a)The expected value of the output signal is obtained from

$$\mu_y = \mu_x H(e^{j0}) = \mu_\varepsilon H(e^{j0}) = 0.$$

The z-transform of the autocorrelation of the output signal, $y(n)$, is

$$R_{yy}(z) = R_{xx}(z)H(z)H(1/z),$$

since $H(z)$ is the z-transform of a real-valued signal, when

$$H^*(1/z^*) = \sum_{n=-\infty}^{\infty}(h(n)(1/z^*)^{-n})^* = \sum_{n=-\infty}^{\infty} h(n)(1/z)^{-n} = H(1/z).$$

The autocorrelation of the input signal, $x(n)$, is

$$R_{xx}(z) = 1.$$

The transfer function of the considered system has the following form

$$H(z) = \frac{1}{1-1.3z^{-1}+0.36z^{-2}} = \frac{1}{(1-0.9z^{-1})(1-0.4z^{-1})}.$$

Therefore, the autocorrelation of the output signal is

$$R_{yy}(z) = \frac{1}{(1-0.9z^{-1})(1-0.4z^{-1})(1-0.9z)(1-0.4z)}$$

or

$$R_{yy}(z) = \frac{25}{8}\left[\frac{z}{(z-0.4)(z-1/0.4)} - \frac{z}{(z-0.9)(z-1/0.9)}\right].$$

The inverse z-transform of $R_{yy}(z)$ is

$$r_{yy}(n) = \frac{25}{8}\left[(0.9)^{|n|}\frac{0.9}{0.19} - (0.4)^{|n|}\frac{0.4}{0.84}\right].$$

(b) The power spectral density of the output signal is obtained in the form

$$S_{yy}(\omega) = R_{yy}(z)_{|z=e^{j\omega}} = \frac{1}{(1.16-0.8\cos\omega)(1.81-1.8\cos\omega)},$$

while the cross-power spectral density function $S_{yx}(\omega)$ is equal to $R_{yx}(z)$ at $z = e^{j\omega}$,

$$S_{yx}(\omega) = R_{yx}(z)_{|z=e^{j\omega}} = H(z)R_{xx}(z)|_{z=e^{j\omega}}$$
$$= \frac{1}{1-1.3\cos\omega+0.36\cos 2\omega + j(1.3\sin\omega - 0.36\sin 2\omega)}.$$

Example 7.51. The white noise $\varepsilon(n)$ with variance σ_ε^2 and zero mean is an input to a linear time-invariant system. If the impulse response of the system is $h(n)$ show that

$$\mathrm{E}\{x(n)y(n)\} = h(0)\sigma_\varepsilon^2$$

and

$$\sigma_y^2 = \sigma_\varepsilon^2 \sum_{n=-\infty}^{\infty} |h(n)|^2 = \sigma_\varepsilon^2 E_h,$$

where $y(n)$ is the output of this system.

★The expected value of the product of the input signal and the output signal is

$$\mathrm{E}\{x(n)y(n)\} = E\left\{\sum_{k=-\infty}^{\infty} h(k)x(n)x(n-k)\right\}.$$

Since the impulse response is a deterministic signal

$$\mathrm{E}\{x(n)y(n)\} = \sum_{k=-\infty}^{\infty} h(k)\mathrm{E}\{x(n)x(n-k)\} = \sum_{k=-\infty}^{\infty} h(k)r_{xx}(k).$$

the autocorrelation function of the input signal is

$$r_{xx}(n) = \sigma_\varepsilon^2\delta(n),$$

producing

$$\mathrm{E}\{x(n)y(n)\} = \sum_{k=-\infty}^{\infty} h(k)\sigma_\varepsilon^2\delta(k) = h(0)\sigma_\varepsilon^2.$$

The variance of output signal is defined by

$$\sigma_y^2 = \mathrm{E}\{y(n)y^*(n)\} - \mathrm{E}\{y(n)\}\,E\{y^*(n)\}$$

or

$$\sigma_y^2 = \mathrm{E}\left\{\sum_{k=-\infty}^{\infty} h(k)x(n-k)\sum_{k=-\infty}^{\infty} h^*(k)x^*(n-k)\right\}$$
$$-\mathrm{E}\left\{\sum_{k=-\infty}^{\infty} h(k)x(n-k)\right\}\mathrm{E}\left\{\sum_{k=-\infty}^{\infty} h^*(k)x^*(n-k)\right\}.$$

The output signal is a zero-mean signal, since

$$\mathrm{E}\{y(n)\} = \mathrm{E}\{y^*(n)\} = \sum_{k=-\infty}^{\infty} h(k)\mathrm{E}\{x(n-k)\} = 0.$$

The variance of the output signal can be written in the form

$$\sigma_y^2 = \sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} h(k)h^*(l)E\{x(n-k)x^*(n-l)\} = \sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} h(k)h^*(l)r_{xx}(l-k).$$

Since $r_{xx}(n) = \sigma_\varepsilon^2\delta(n)$, that is, $r_{xx}(l-k) = \sigma_\varepsilon^2\delta(l-k)$, only the terms with $l = k$ remain in the double summation expression for the variance σ_y^2, producing

$$\sigma_y^2 = \sigma_\varepsilon^2 \sum_{k=-\infty}^{\infty} |h(k)|^2 = \sigma_\varepsilon^2 E_h.$$

7.6.1 Spectral Estimation of Narrowband Signals

A narrowband random signal with N_p components around the frequencies ω_1, ω_2, and ω_{N_p} can be considered, from a spectral point of view, as an output of the system whose transfer function is

$$H(z) = \frac{G}{(1-r_1e^{j\omega_1}z^{-1})(1-r_2e^{j\omega_2}z^{-1})\ldots(1-r_{N_p}e^{j\omega_{N_p}}z^{-1})}$$
$$= \frac{G}{1+a_1z^{-1}+a_2z^{-2}+\cdots+a_{N_p}z^{-N_p}}.$$

when the input is a white noise. The amplitudes of the poles r_i, $i = 1,2,\ldots,N_p$, are inside (and close to) the unit circle. The discrete-time domain description of this system is

$$y(n) + a_1y(n-1) + a_2y(n-2) + \cdots + a_{N_p}y(n-N_p) = Gx(n), \tag{7.133}$$

where $x(n)$ is a white noise with variance $\sigma_x^2 = 1$, the autocorrelation $r_{xx}(k) = \delta(k)$, and the spectral energy density $S_{xx}(\omega) = 1$. For a given narrowband random signal $y(n)$, the task is to find coefficients a_i and G.

The autocorrelation of the real-valued output signal is obtained after the multiplication of the difference equation by $y(n-k)$,

$$y(n)y(n-k)+a_1y(n-1)y(n-k)+\cdots+a_{N_p}y(n-N_p)y(n-k)=Gx(n)y(n-k),$$

and the expected value calculation,

$$\mathrm{E}\{y(n)y(n-k)+a_1y(n-1)y(n-k)+\cdots+a_{N_p}y(n-N_p)y(n-k)\}=\mathrm{E}\{Gx(n)y(n-k)\}.$$

For $k=0$, follows

$$r_{yy}(0)+a_1r_{yy}(0-1)+a_2r_{yy}(0-2)+\cdots+a_{N_p}r_{yy}(0-N_p)=G^2.$$

For $k>0$ and a causal system, we may find that $r_{xy}(k)=h(-k)=0$. It is also clear from (7.133) that $x(n)$ is related to $y(n)$ and that any $y(n-k)$, for $k>0$, does not include $x(n)$, meaning that $\mathrm{E}\{x(n)y(n-k)\}=0$, and

$$r_{yy}(k)+a_1r_{yy}(k-1)+a_2r_{yy}(k-2)+\cdots+a_{N_p}r_{yy}(k-N_p)=0.$$

The previous equations are known as the Yule-Walk equations. The matrix form of this system of equations is

$$\begin{bmatrix} r_{yy}(0) & r_{yy}(1) & \dots & r_{yy}(N_p) \\ r_{yy}(1) & r_{yy}(0) & \dots & r_{yy}(N_p-1) \\ \dots & \dots & \dots & \dots \\ r_{yy}(N_p) & r_{yy}(N_p-1) & \dots & r_{yy}(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ a_2 \\ \dots \\ a_{N_p} \end{bmatrix} = \begin{bmatrix} G^2 \\ 0 \\ 0 \\ \dots \\ 0 \end{bmatrix}. \tag{7.134}$$

The system is solved for the unknown system coefficients $[a_0, a_1, a_2,\dots,a_{N_p}]$ with $G=1$. Then, the coefficients are normalized as $[a_0, a_1, a_2,\dots,a_{N_p}]/a_0$ with $G=1/a_0$. The spectral energy density of $y(n)$ follows with $S_{xx}(\omega)=1$ as

$$S_{yy}(\omega)=\left|\frac{G}{1+a_1e^{-j\omega}+a_2e^{-j2\omega}+\cdots+a_{N_p}e^{-jN_p\omega}}\right|^2. \tag{7.135}$$

This is the autoregressive (AR) spectral estimation.

Note that the autocorrelation functions for real-valued $y(n)$, defined within $0\le n\le N-1$, can be estimated, using

$$r_{yy}(k)=\frac{1}{N-k}\sum_{n=0}^{N-1-k}y(n+k)y(n) \qquad \text{for } 0\le k\le N-1, \tag{7.136}$$

and $r_{yy}(k)=r_{yy}(-k)$ for $-N+1\le k<0$. These values are then used in (7.134) for the autoregressive spectral estimation.

Next, we will comment the estimated autocorrelation within the basic definition of the power spectral density framework, Section 7.3.4. Relation (7.136) corresponds to the unbiased estimation of the autocorrelation function. The power spectral density, according to (7.32), is calculated as $S_{yy}(\omega)=\mathrm{FT}\{r_{yy}(k)\}$.

Since the autocorrelation estimates for a large k use only a small number of signal samples in averaging, they are not reliable. It is common to apply a triangular (Bartlett) window function

($w(k) = (N - |k|)/N$) to reduce the weight of these estimates in the Fourier transform calculation

$$w(k)r_{yy}(k) = w(k)\frac{1}{N-k}\sum_{n=0}^{N-1-k} y(n+k)y(n) = \frac{1}{N}\sum_{n=0}^{N-1-k} y(n+k)y(n) \tag{7.137}$$

for $0 \leq k \leq N-1$. Since the window is used, this autocorrelation function estimate is biased. The Fourier transform of the biased autocorrelation function $w(k)r_{yy}(k) = (1-|k|/N)r_{yy}(k)$ is the power spectral density $P_{yy}(\omega) = \mathrm{FT}\{(1-|k|/N)r_{yy}(k)\}$ defined by (7.34).

Example 7.52. Consider the random signal

$$y(n) = 2\cos(1.4n + \varphi_1) + \sqrt{2}\sin(1.6n + \varphi_2) + 10e^{-(n-N/2)^2/16} + \varepsilon(n),$$

within $0 \leq n \leq N-1 = 127$, where φ_1 and φ_2 are random variables uniformly distributed from $-1/2$ rad to $1/2$ rad, while $\varepsilon(n)$ is the zero-mean, unit-variance Gaussian noise. Plot the power spectral density calculated using:

(a) The Fourier transform of $r_{yy}(k)$

$$S_{yy}(\omega) = \mathrm{FT}\{r_{yy}(k)\} = \sum_{k=-N+1}^{N-1} r_{yy}(k)e^{-j\omega k}.$$

where $r_{yy}(k)$ is calculated using

$$r_{yy}(\pm k) = \frac{1}{N-k}\sum_{i=0}^{N-k-1} x(k+i)x(i).$$

(b) The Fourier transform of the signal $y(n)$

$$P_{yy}(\omega) = \frac{1}{N}\left|\sum_{n=0}^{N-1} y(n)e^{-j\omega n}\right|^2.$$

(c) The power spectrum in (b) corresponds to $\mathrm{FT}\{w_B(k)r_{yy}(k)\}$, where $w_B(k)$ is the Bartlett window whose width is equal to the width of the autocorrelation function r_{yy}.

(d) The Fourier transform of the signal $y(n)$ over $K = 7$ shorter intervals. The duration of these intervals is $M = 32$, with the step $R = M/2$. The Fourier transforms of $y(n)$, within these shorter intervals, are

$$Y_i(e^{j\omega}) = \frac{1}{M}\sum_{n=0}^{M-1} y(iR+n)e^{-j\omega n}$$

for $i = 0,1,\ldots,6$. The power spectral densities $|Y_i(e^{j\omega})|^2$ are averaged to produce (Welch periodogram)

$$S_{yy}^A(\omega) = \frac{1}{K}\sum_{i=0}^{K-1}\left|Y_i(e^{j\omega})\right|^2.$$

(e) Relation (7.135) with appropriately estimated coefficients a_i and G, along with the relations (7.134) and (7.136).

★The results are shown in Fig. 7.32, in order from (a) to (e).

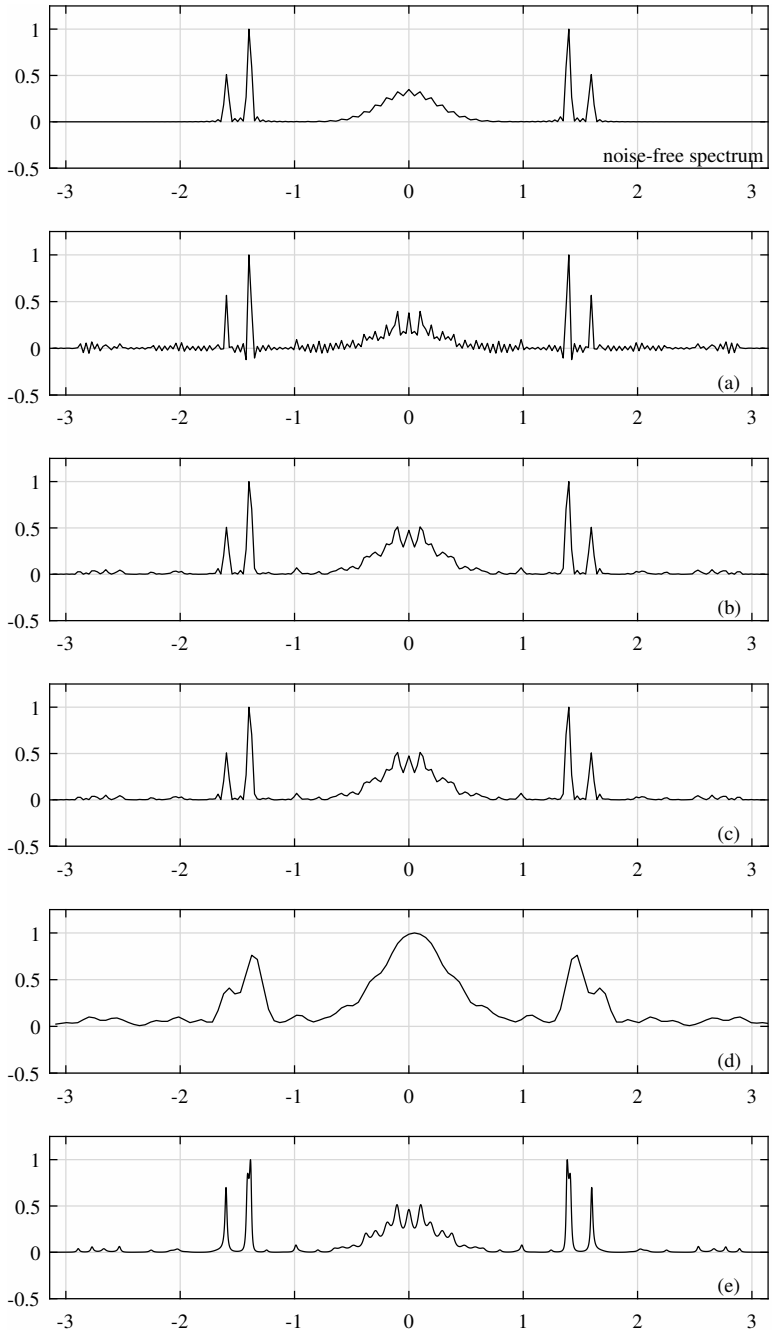


Figure 7.32 Spectral analysis of a signal with random phases (normalized values). Order and the description of the panels correspond to the task order in Example 7.52.

7.6.2 Detection and Matched Filter

Detection of an unknown deterministic signal in a high noise environment is of crucial interest in many real-world applications. In this case the problem is in testing the hypothesis

$$\begin{aligned} H_0 &: \text{ Signal is not present in the observed noisy signal} \\ H_1 &: \text{ Signal is present in the observed noisy signal} \end{aligned} \tag{7.138}$$

Hypothesis testing is described in Section 7.4.9.

The decision-making process based on the hypothesis testing is improved if we provide the best possible observation parameter for the decision. Here we will present the detection of a known signal in a white noise using the matched filter, which is designed to maximize the output signal value.

7.6.2.1 Matched Filter

Consider a general signal form

$$x(n) = s(n) + \varepsilon(n),$$

where $s(n)$ is a known signal with the Fourier transform $S(e^{j\omega})$ and $\varepsilon(n)$ is a white noise with the power spectral density σ_ε^2. The problem is to find a system with the maximum output if the input $x(n)$ contains the signal $s(n)$. The output signal is used to test the null hypothesis H_0: *the signal* $s(n)$ *is not present in* $x(n)$.

The output of the system with the impulse response $h(n)$ and the frequency response $H(e^{j\omega})$, to the input signal $x(n)$, is of the form

$$y(n) = y_s(n) + y_\varepsilon(n),$$

where $y_s(n)$ and $y_\varepsilon(n)$ are the system outputs to the input signals $s(n)$ and $\varepsilon(n)$, respectively. For the output signal $y_s(n)$ holds

$$Y_s(e^{j\omega}) = H(e^{j\omega})S(e^{j\omega}).$$

The power spectral density of $y_s(n)$ is equal to

$$\left|Y_s(e^{j\omega})\right|^2 = \left|H(e^{j\omega})\right|^2 \left|S(e^{j\omega})\right|^2.$$

The power of the output noise is given by

$$E\{|y_\varepsilon(n)|^2\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|H(e^{j\omega})\right|^2 \sigma_\varepsilon^2 d\omega.$$

The output signal $y(n)$, at an instant n_0, is

$$y_s(n_0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} H(e^{j\omega})S(e^{j\omega})e^{j\omega n_0}d\omega, \quad \text{with} \quad |y_s(n_0)|^2 = \left|\frac{1}{2\pi}\int_{-\pi}^{\pi} H(e^{j\omega})S(e^{j\omega})e^{j\omega n_0}d\omega\right|^2.$$

The aim is to maximize the output signal at an instant n_0 if the input signal contains $s(n)$. According to Schwartz's inequality (for its discrete form see Part VI)

$$\left|\frac{1}{2\pi}\int_{-\pi}^{\pi} H(e^{j\omega})S(e^{j\omega})e^{j\omega n_0}d\omega\right|^2 \le \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|S(e^{j\omega})\right|^2 d\omega \;\; \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|H(e^{j\omega})\right|^2 d\omega,$$

the peak output signal-to-noise ratio is

$$PSNR = \frac{|y_s(n_0)|^2}{\mathrm{E}\{|y_\varepsilon(n)|^2\}} \leq \frac{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|S(e^{j\omega})\right|^2 d\omega \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|H(e^{j\omega})\right|^2 d\omega}{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|H(e^{j\omega})\right|^2 \sigma_\varepsilon^2 d\omega}.$$

This ratio reaches its maximum when the equality sign holds

$$PSNR_{\max} = \frac{1}{2\pi\sigma_\varepsilon^2}\int_{-\pi}^{\pi}\left|S(e^{j\omega})\right|^2 d\omega = \frac{E_s}{\sigma_\varepsilon^2}.$$

The maximum ratio in Schwartz's inequality is achieved for

$$H(e^{j\omega}) = kS^*(e^{j\omega})e^{-j\omega n_0}.$$

In the time domain, the impulse response is

$$h(n) = ks^*(n_0 - n).$$

This system is called *matched filter*. Its impulse response is matched to the signal form. The matched filter maximizes the ratio of the output signal and the noise and is used in the detection, to make a decision if the known signal $s(n)$ exists in the noisy signal $x(n)$.

If the additive noise is Gaussian distributed then the null hypothesis (there is no deterministic signal in the input signal) rejection region, for significance level $\alpha = 0.001$ is

$$\text{Probability}\{|y(n_0)| > \lambda\} = 1 - \text{erf}(\frac{\lambda}{\sqrt{2}\sigma_y}) < 0.001,$$
$$|y(n_0)| > 3.2905\sigma_y$$

For the application, where an order of 1000 nonzero samples is expected in the output signal, the significance level must be small. Otherwise, we will have many false positive results.

Example 7.53. The matched filter is illustrated on the detection of the *chirp signal*

$$s(n) = e^{-2(n/128)^2}\cos(8\pi(n/128)^2 + \pi n/8)$$

in a Gaussian white noise of the variance $\sigma_\varepsilon^2 = 1$. The output of the matched filter is calculated for $n_0 = 0$ using the known signal,

$$y(n) = x(n) *_n s(-n) = \sum_{m=-\infty}^{\infty}(s(m) + \varepsilon(m))s(m-n).$$

The maximum value of the output signal is reached at $n = 0$, when

$$y(0) = \sum_{m=-\infty}^{\infty}(s(m) + \varepsilon(m))s(m) \quad \text{and} \quad \mathrm{E}\{y(0)\} = \sum_{m=-\infty}^{\infty} s^2(m) = E_s = 56.86.$$

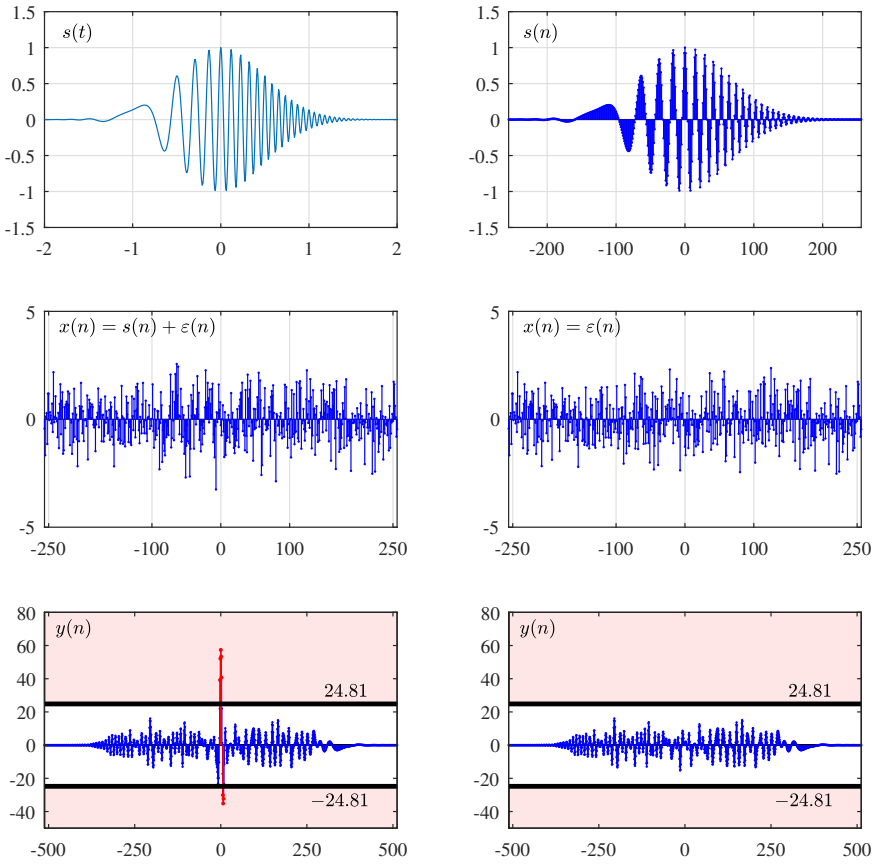


Figure 7.33 Illustration of the matched filter: Signal $s(t)$, $s(n)$. Input noisy signal $x(n) = s(n) + \varepsilon(n)$, contains the signal $s(n)$. Input signal $x(n) = \varepsilon(n)$ does not contain the signal $s(n)$. Corresponding outputs from the matched filter $y(n) = x(n) * s(-n)$ are presented bellow the input signal panels. The null hypothesis rejection region is shaded.

The variance of the output noise is derived in Example 7.51 as

$$\sigma_y^2 = \sigma_\varepsilon^2 \sum_{n=-\infty}^{\infty} |h(n)|^2 = \sigma_\varepsilon^2 \sum_{n=-\infty}^{\infty} |s(-n)|^2 = \sigma_\varepsilon^2 E_s,$$

The null hypothesis rejection region is defined by

$$|y(n_0)| > 3.2905\sigma_y = 3.2905\sigma_\varepsilon\sqrt{E_s} = 24.81.$$

Two cases are shown in Fig. 7.33: (1) When the input signal contains $s(n)$ and (2) when the input signal does not contain $s(n)$. We can see that the output of the matched filter has an easily detectable peak at $n = 0$ for the case then the input signal contains $s(n)$. There is no such a peak in $y(n)$ when the input signal $x(n)$ is noise only. Therefore, the null hypothesis can be rejected

in the case presented in the left panels, while it can not be rejected for the case shown in the right panels.

7.6.2.2 Two Hypothesis Decision

Consider a signal $s(n)$ that can take one of two constant values $s(n) = A_1$ or $s(n) = A_2$, corrupted by an additive random noise $\varepsilon(n)$:

(1) $x(n) = A_1 + \varepsilon(n)$ or

(2) $x(n) = A_2 + \varepsilon(n)$.

Assume that the probabilities of these two signal states are $P(A_1)$ and $P(A_2)$, such that $P(A_1) + P(A_2) = 1$. In this experiment, a value of the signal $x(n) = y$ is observed and the question is which of these two hypothesis is true:

$$\begin{aligned} H_1:&\ \text{Signal is of the form } x(n) = A_1 + \varepsilon(n) \\ H_2:&\ \text{Signal is of the form } x(n) = A_2 + \varepsilon(n). \end{aligned} \tag{7.139}$$

Bayes' formula for the state $x(n) = A_1 + \varepsilon(n)$ is

$$P(A_1|y)\ p(y)dy = p(y|A_1)dy\ P(A_1)$$

where $p(y)dy$ is the probability that y takes a specific value within $[y, y + dy)$. This relation can be written as

$$P(A_1|y) = \frac{p(y|A_1)P(A_1)}{p(y)}.$$

Similarly, Bayes' formula for the state $x(n) = A_2 + \varepsilon(n)$ produces

$$P(A_2|y) = \frac{p(y|A_2)P(A_2)}{p(y)}.$$

Since $P(A_1|y)$ is the probability of A_1 if y occurred and $P(A_2|y)$ is the probability of A_2 if the same y occurred, then a criterion to state that the hypothesis H_1 is true can be defined by

$$P(A_1|y) > P(A_2|y).$$

This relation produces the decision relations

$$\begin{aligned} H_1:&\quad \text{is true if}\quad p(y|A_1)P(A_1) > p(y|A_2)P(A_2) \\ H_2:&\quad \text{is true if}\quad p(y|A_1)P(A_1) < p(y|A_2)P(A_2) \end{aligned} \tag{7.140}$$

which are commonly written in the form

$$p(y|A_1)P(A_1) \underset{H_2}{\overset{H_1}{\gtrless}} p(y|A_2)P(A_2) \quad \text{or} \quad \frac{p(y|A_1)}{p(y|A_2)} \underset{H_2}{\overset{H_1}{\gtrless}} \frac{P(A_2)}{P(A_1)},$$

where $p(y|A_1)/p(y|A_2)$ is the likelihood ratio.

In the logarithmic form, the log-likelihood ratio criterion for the decision is given by

$$\ln\Big(\frac{p(y|A_1)}{p(y|A_2)}\Big) \underset{H_2}{\overset{H_1}{\gtrless}} \ln\Big(\frac{P(A_2)}{P(A_1)}\Big). \tag{7.141}$$

The decision threshold, d, is obtained from

$$\ln\Big(\frac{p(d|A_1)}{p(d|A_2)}\Big)=\ln\Big(\frac{P(A_2)}{P(A_1)}\Big).$$

For the Gaussian probability density function of the disturbance $\varepsilon(n)$, the signal $x(n)$ in the case $x(n)=A_1+\varepsilon(n)$ is distributed as

$$p_{\xi}(\xi|A_1)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(\xi-A_1)^2/(2\sigma_x^2)}.$$

For $x(n)=A_1+\varepsilon(n)$ the probability density is

$$p_{\xi}(\xi|A_2)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(\xi-A_2)^2/(2\sigma_x^2)}.$$

The decision threshold is then obtained from

$$e^{-((d-A_1)^2-(d-A_2)^2)/(2\sigma_x^2)}=P(A_2)/P(A_1)$$

or from the logarithmic form

$$2d(A_1-A_2)-A_1^2+A_2^2=2\sigma_x^2\ln(\frac{P(A_2)}{P(A_1)}).$$

The threshold value, for the Gaussian distribution, is

$$d=\frac{\sigma_x^2\ln(\frac{P(A_2)}{P(A_1)})}{A_1-A_2}+\frac{A_1+A_2}{2}.$$

Obviously, the special case when $P(A_2)=P(A_1)$, produces $d=(A_1+A_2)/2$.

When a signal sample $x(n)=y$ is obtained as a result of the experiment, the hypothesis testing rule is now simple,

$$y\underset{H_2}{\overset{H_1}{\gtrless}}d. \tag{7.142}$$

Example 7.54. Consider the random signal with two possible forms: (1) $x(n)=A_1+\varepsilon(n)=1+\varepsilon(n)$ or (2) $x(n)=A_2+\varepsilon(n)=-1+\varepsilon(n)$, where $\varepsilon(n)$ is the zero-mean Gaussian distributed random variable with variance $\sigma_\varepsilon^2=0.5$. Assume that the probabilities of these two states are $P(A_1)=1/3$ and $P(A_2)=2/3$. Find the decision threshold d.

★The posterior probability functions, $P(A_1)p_\varepsilon(\xi-A_1)$ and $P(A_2)p_\varepsilon(\xi-A_2)$, are shown in Fig. 7.34. The decision threshold is obtained from

$$d=\frac{\sigma_x^2\ln(\frac{P(A_2)}{P(A_1)})}{A_1-A_2}+\frac{A_1+A_2}{2}=\frac{0.5^2\ln(\frac{2/3}{1/3})}{1-(-1)}+\frac{1+(-1)}{2}=0.08664.$$

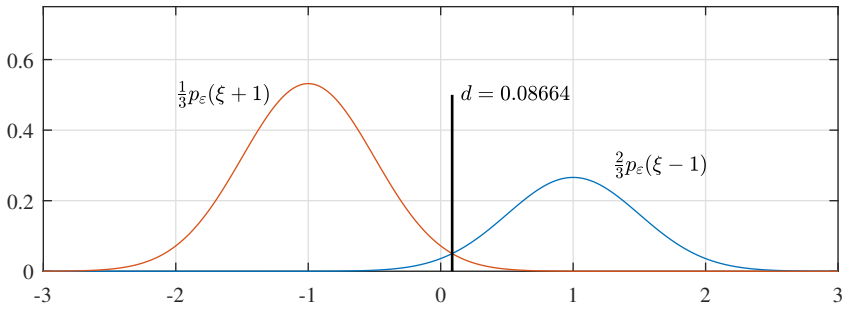


Figure 7.34 Illustration of the detection thershold calculation.

7.6.3 Optimal Wiener Filter

Assume that the input signal is $x(n)$ and that it contains an information about the desired signal $d(n)$. The input signal is processed by a system whose impulse response is $h(n)$. The output signal is $y(n) = h(n) *_n x(n)$. The task here is to find the impulse response $h(n)$ of the system such that the difference of the desired signal and the output signal, denoted as the error

$$e(n) = d(n) - y(n),$$

is minimum in the mean squared sense, that is,

$$h(n) = \min_{h(n)}\{\mathrm{E}\{|e(n)|^2\}\}.$$

The mean squared error is

$$\mathrm{E}\{|e(n)|^2\} = \mathrm{E}\{\left|d(n) - \sum_{m=-\infty}^{\infty} h(m)x(n-m)\right|^2\}.$$

The minimum value is obtained from

$$\frac{\partial \mathrm{E}\{|e(n)|^2\}}{\partial h^*(k)} = E\left\{2\left(d(n) - \sum_{m=-\infty}^{\infty} h(m)x(n-m)\right)x^*(n-k)\right\} = 0. \tag{7.143}$$

This relation states that expected value of the product of the error signal $e(n) = d(n) - y(n)$ and the input signal $x^*(n-k)$ is zero

$$\mathrm{E}\{2e(n)x^*(n-k)\} = 0$$

for any k. For signals satisfying this relation we say that they are normal to each other.

Relation (7.143) can be written as

$$\mathrm{E}\left\{\sum_{m=-\infty}^{\infty} h(m)x(n-m)x^*(n-k)\right\} = \mathrm{E}\{d(n)x^*(n-k)\}$$

or

$$\sum_{m=-\infty}^{\infty} h(m)r_{xx}(k-m) = r_{dx}(k).$$

Taking the z-transform of both sides of the last equation we get

$$H(z)R_{xx}(z) = R_{dx}(z).$$

The transfer function of the optimal filter is

$$H(z) = \frac{R_{dx}(z)}{R_{xx}(z)}.$$

For a special case, when the input signal is the desired signal $d(n)$ with an additive noise

$$x(n) = d(n) + \varepsilon(n),$$

where $\varepsilon(n)$ is uncorrelated with the desired signal, the optimal (Wiener) filtering relation follows

$$H(z) = \frac{R_{dd}(z)}{R_{dd}(z) + R_{\varepsilon\varepsilon}(z)}$$

since

$$r_{dx}(k) = E\{d(n)x^*(n-k)\} = E\{d(n)[d^*(n-k) + \varepsilon^*(n-k)]\} = r_{dd}(k).$$

Here we used $E\{d(n)\varepsilon^*(n-k)\} = 0$, since $d(n)$ and $\varepsilon(n)$ are uncorrelated. Also

$$r_{xx}(k) = E\{[d(n) + \varepsilon(n)][d^*(n-k) + \varepsilon^*(n-k)]\} = r_{dd}(k) + r_{\varepsilon\varepsilon}(k).$$

The frequency response of the optimal filter is given by

$$H(e^{j\omega}) = \frac{S_{dd}(\omega)}{S_{dd}(\omega) + S_{\varepsilon\varepsilon}(\omega)}.$$

Example 7.55. A signal $x(n) = d(n) + \varepsilon(n)$ is processed by an optimal filter. Power spectral density of $d(n)$ is $S_{dd}(\omega)$. If the signal $d(n)$ and the additive noise $\varepsilon(n)$, whose power spectral density is $S_{\varepsilon\varepsilon}(\omega)$, are independent find the output signal-to-noise ratio.

★For this signal and noise, according to (7.132), we have

$$S_{yy}(e^{j\omega}) = \left|H(e^{j\omega})\right|^2 S_{xx}(e^{j\omega}) = \left|\frac{S_{dd}(\omega)}{S_{dd}(\omega) + S_{\varepsilon\varepsilon}(\omega)}\right|^2 S_{xx}(e^{j\omega}) = \frac{S_{dd}^2(\omega)}{S_{dd}(\omega) + S_{\varepsilon\varepsilon}(\omega)}$$

since $S_{xx}(e^{j\omega}) = S_{dd}(\omega) + S_{\varepsilon\varepsilon}(\omega)$. The output signal-to-noise ratio is

$$SNR = \frac{\frac{1}{2\pi}\int_{-\pi}^{\pi} S_{dd}(\omega)\left|H(e^{j\omega})\right|^2 d\omega}{\frac{1}{2\pi}\int_{-\pi}^{\pi} S_{\varepsilon\varepsilon}(\omega)\left|H(e^{j\omega})\right|^2 d\omega}.$$

Note that the input signal-to-noise ratio is

$$SNR_i = \frac{\frac{1}{2\pi}\int_{-\pi}^{\pi} S_{dd}(\omega)d\omega}{\frac{1}{2\pi}\int_{-\pi}^{\pi} S_{\varepsilon\varepsilon}(\omega)d\omega}.$$

The optimal prediction system follows with the input signal $x(n) = d(n-1) + \varepsilon(n-1)$ and the desired signal $d(n)$. Transfer function of the optimal predictor is obtained from

$$r_{dx}(k) = E\{d(n)x^*(n-k)\} = E\{d(n)[d^*(n-1-k) + \varepsilon^*(n-1-k)]\} = r_{dd}(k+1)$$

and

$$r_{xx}(k) = E\{[d(n-1) + \varepsilon(n-1)][d^*(n-1-k) + \varepsilon^*(n-1-k)]\} = r_{dd}(k) + r_{\varepsilon\varepsilon}(k).$$

as

$$H(z) = \frac{zS_{dd}(z)}{S_{dd}(z) + S_{\varepsilon\varepsilon}(z)}$$

since

$$\sum_{k=-\infty}^{\infty} r_{dd}(k+1)z^{-k} = \sum_{k=-\infty}^{\infty} r_{dd}(k)z^{-k+1} = zS_{dd}(z).$$

The optimal smoothing is the case when the desired signal is $d(n)$ and we can use its future value(s). This processing follows with $x(n) = d(n+1) + \varepsilon(n+1)$ as

$$H(z) = \frac{z^{-1}S_{dd}(z)}{S_{dd}(z) + S_{\varepsilon\varepsilon}(z)}.$$

Example 7.56. The input signal is $x(n) = s(n) + \varepsilon(n)$, where $d(n) = s(n)$ is the desired signal and $\varepsilon(n)$ is a noise. If the autocorrelation functions of the signal and noise are $r_{ss}(n) = 4^{-|n|}$ and $r_{\varepsilon\varepsilon}(n) = 2\delta(n)$, respectively, and the cross-correlation of the signal and noise is $r_{s\varepsilon}(n) = \delta(n)$, design the optimal filter.

★The optimal filter transfer function is

$$H(z) = R_{dx}(z)/R_{xx}(z)$$

where are

$$R_{dx}(z) = R_{ss}(z) + R_{s\varepsilon}(z) \quad \text{and} \quad R_{xx}(z) = R_{ss}(z) + 2R_{s\varepsilon}(z) + R_{\varepsilon\varepsilon}(z).$$

Based on the correlation functions, we can calculate the z-transforms

$$R_{ss}(z) = \sum_{n=-\infty}^{\infty} r_{ss}(n)z^{-n} = \sum_{n=-\infty}^{\infty} 4^{-|n|}z^{-n} = \sum_{n=-\infty}^{-1} 4^n z^{-n} + \sum_{n=0}^{\infty} 4^{-n}z^{-n} =$$
$$= \frac{z/4}{1 - z/4} + \frac{1}{1 - 1/(4z)} = \frac{-3.75z}{(z-0.25)(z-4)},$$

while $R_{s\varepsilon}(z) = 1$ and $R_{\varepsilon\varepsilon}(z) = 2$.

The transfer function of the optimal filter is

$$H(z) = \frac{R_{ss}(z) + R_{s\varepsilon}(z)}{R_{ss}(z) + 2R_{s\varepsilon}(z) + R_{\varepsilon\varepsilon}(z)} = \frac{0.25z^2 - 2z + 0.25}{z^2 - 5.1875z + 1}.$$

The optimal systems realization using the FIR filters will be presented within the introductory part of the chapter dealing with adaptive discrete systems in Part III.

7.7 QUANTIZATION EFFECTS

In order to process analog signals using computers they have to be converted into numbers stored into registers of a finite precision. Continuous-time signals are transformed into digital signals using analog-to-digital (A/D) converters. This operation is done in two steps. First, the continuous-time signal is converted into a discrete-time signal by taking samples of the continuous-time signal at discrete-time instants (sampling)

$$x(n) = x(n\Delta t)\Delta t.$$

Next, the discrete-time signal, with continuous amplitudes of samples, is converted into a digital signal

$$x_Q(n) = Q[x(n)]$$

with discrete-valued amplitudes (quantization). This process is illustrated in Fig. 7.35. The error caused by the quantization of the discrete-time signal amplitudes is called the *quantization noise*.

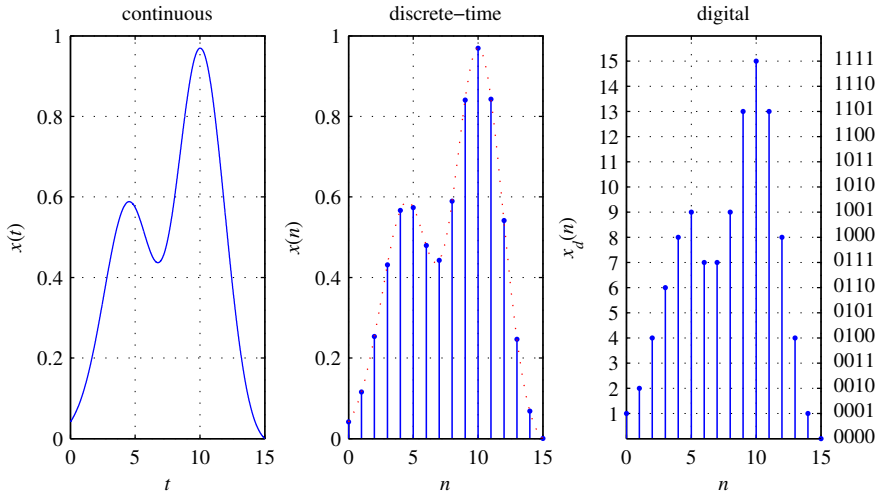


Figure 7.35 Illustration of a continuous signal and its discrete-time and digital version.

The quantization noise influences results of signal processing in several ways:

-Input signal quantization error, described by an additive quantization noise. This influence (in the form of an additive input noise that depends on quantization step Δ) can be modeled as the uniform noise with values between $-\Delta/2$ and $\Delta/2$.

-Quantization of the results of arithmetic operations. It depends on the way how the calculations are performed.

-Quantization of the coefficients in algorithms. Usually this kind of error is neglected in analysis since it is deterministic (comments on the errors in the coefficients are given in the chapter dealing with the realizations of discrete-time systems).

In order to make appropriate analysis of the quantization effects, common assumptions are:

1) random variables corresponding to the quantization errors are uncorrelated, that is, the quantization error is a white noise process with the uniform distribution,

2) the error sources are uncorrelated with one another, and

3) all the errors are uncorrelated with the input signal and, consequently, with all signals in the system.

7.7.1 Input signal quantization

For registers with b bits the digital signal values $x_Q(n)$ are coded into binary format. Assume that the registers with b bits are used and that all input signals are normalized to the range

$$0 \leq x(n) < 1.$$

The binary numbers are written within the register as

$$\boxed{a_{-1}}\boxed{a_{-2}}\boxed{a_{-3}}\boxed{\ldots}\boxed{a_{-b}}.$$

The value of $x_Q(n)$ is

$$x_Q(n) = a_{-1}2^{-1} + a_{-2}2^{-2} + \cdots + a_{-b}2^{-b}.$$

The maximum number that can be written within this format is $0.111\ldots11$ representing $1 - 2^{-(b+1)}$. Common number of bits b ranges from 8 to 24.

For reducing the signal number of digits to b bits, *rounding or truncation* is used. An example of the quantization with $b = 4$ bits is presented in Fig. 7.35, where the maximum value of $x_d(n) = x_Q(n)$ is denoted by 1111, meaning $2^{-1} + 2^{-2} + 2^{-3} + 2^{-4} = 15/16$.

For the case with positive and negative numbers, one extra bit is used for the sign. The registers are now with $b + 1$ bits. The first bit is the sign bit and the remaining b bits represent the signal absolute value

$$\boxed{s}\boxed{a_{-1}}\boxed{a_{-2}}\boxed{a_{-3}}\boxed{\ldots}\boxed{a_{-b}}\ .$$

In computers, negative numbers are commonly represented in a complement of 2 form.

In order to distinguish these two cases we will use register of length b, meaning no sign bit exists and register of length $b + 1$, where the sign bit is used.

Example 7.57. In a register with $b = 8$ bits, the binary number $x_Q(n)$

$$\boxed{1}\boxed{0}\boxed{1}\boxed{1}\boxed{0}\boxed{0}\boxed{1}\boxed{0}$$

has the decade system value

$$x_Q(n) = 1 \cdot 2^{-1} + 1 \cdot 2^{-3} + 1 \cdot 2^{-4} + 1 \cdot 2^{-7} = \frac{89}{128} = 0.6953.$$

The decimal point is at the position just before the first digit. The values of $x_Q(n)$, in this register, are

$$0 \leq x_Q(n) \leq \frac{255}{256}$$

with the quantization step $1/256$.

The quantization error is a difference in the amplitude of the original signal and the quantized signal

$$e(n) = x(n) - x_Q(n).$$

When the rounding approach is used, the maximum absolute error can be a half of the last digit weight,

$$-\frac{1}{2}2^{-b} \leq x(n) - x_Q(n) < \frac{1}{2}2^{-b}$$
$$-\frac{1}{2}\Delta \leq x(n) - x_Q(n) < \frac{1}{2}\Delta$$

where

$$\Delta = 2^{-b}.$$

We can also write $|e(n)| \leq 2^{-(b+1)} = \frac{1}{2}\Delta$.

In the example from Fig. 7.35, the quantization step is $2^{-4} = 1/16$ and the error is within $|e(n)| \leq \frac{1}{2}\frac{1}{16}$.

The error values are equally probable within the defined interval, with the probability density function

$$p_e(\xi) = \begin{cases} \frac{1}{\Delta} & \text{for } -\frac{1}{2}\Delta \leq \xi < \frac{1}{2}\Delta \\ 0 & \text{elsewhere.} \end{cases},$$

The quantization error of the signal $x(n)$ may be described as an additive uniform white noise.

The expected value of the quantization error, with the rounding approach, is

$$\mu_e = \mathrm{E}\{e(n)\} = \int_{-\Delta/2}^{\Delta/2} \xi p_e(\xi) d\xi = 0.$$

The variance of rounding quantization error is

$$\sigma_e^2 = \int_{-\Delta/2}^{\Delta/2} \frac{1}{\Delta}(\xi - \mu_e)^2 d\xi = \frac{1}{12}\Delta^2.$$

When the truncation is used, the quantization error is within

$$0 \leq x(n) - x_Q(n) < \Delta$$

or $0 \leq e(n) < \Delta$, with the expected value

$$\mu_e = \mathrm{E}\{e(n)\} = \frac{\Delta}{2}$$

and the variance

$$\sigma_e^2 = \int_0^{\Delta} \frac{1}{\Delta}(\xi - \frac{\Delta}{2})^2 d\xi = \frac{1}{12}\Delta^2.$$

Example 7.58. The DFT of a signal $x(n)$ is calculated using the quantized version of this signal

$$x_Q(n) = Q[x(n)] = x(n) + e(n).$$

Quantization is done by an A/D converter with $b + 1 = 8$ bits, using rounding. The DFT is calculated on a high precision computer with $N = 1024$ signal samples. Find the expected value and variance of the calculated DFT.

★The DFT of the quantized signal is

$$X_Q(k) = \sum_{n=0}^{N-1} [x(n) + e(n)]\, e^{-j2\pi kn/N}.$$

Its expected value is

$$\mu_{X_Q}(k) = \mathrm{E}\{X_Q(k)\} = \sum_{n=0}^{N-1} x(n) e^{-j2\pi kn/N} = X(k).$$

The variance is

$$\sigma^2_{X_Q}(k) = \sum_{n_1=0}^{N-1} \sum_{n_2=0}^{N-1} \sigma_e^2 \delta(n_1 - n_2) e^{-j2\pi k(n_1-n_2)/N} = \sigma_e^2 N = \frac{1}{12} 2^{-14} 2^{10} = \frac{1}{192}.$$

The noise in the DFT is a sum of many independent noises from the input signal and coefficients. Thus, it is Gaussian distributed with standard deviation $\sigma_{X_Q} = 0.072$. It may significantly influence the signal DFT values, especially if they are not well concentrated or if there are signal components with small amplitudes.

Example 7.59. How the input signal quantization error influences the results of:

(a) the weighted sum

$$X_s = \sum_{n=0}^{N-1} a_n x(n) \quad \text{and}$$

(b) the product

$$X_P = \prod_{n=0}^{N-1} x(n).$$

★If the quantized values $x_Q(n) = Q[x(n)] = x(n) + e(n)$ of the signal $x(n)$ are used in calculation instead of the true signal values then:

(a) The estimator of a weighted sum is

$$\hat{X}_s = \sum_{n=0}^{N-1} a_n x_Q(n) = \sum_{n=0}^{N-1} a_n x(n) + \sum_{n=0}^{N-1} a_n e(n).$$

Obviously, the total error is

$$e_{X_s} = \sum_{n=0}^{N-1} a_n e(n).$$

It is Gaussian distributed since there are many small errors $e(n)$, for large N. It has been assumed that the weighting coefficients are such that they allow many signal values to influence the result with similar weights.

The expected value of the total error is

$$\mu_{X_s} = \mathrm{E}\{e_{X_s}\} = \sum_{n=0}^{N-1} a_n \mathrm{E}\{e(n)\} = 0,$$

for rounding. The variance of the total error is

$$\sigma^2_{X_s} = \sum_{n=0}^{N-1} a_n^2 \mathrm{var}\{e(n)\} = \frac{1}{12}\Delta^2 \sum_{n=0}^{N-1} a_n^2.$$

(b) The estimator of the product is

$$\hat{X}_P = \prod_{n=0}^{N-1} x_Q(n) = \prod_{n=0}^{N-1} (x(n) + e(n)).$$

Assuming that the individual errors are small so that all the higher-order error terms, containing the error products, for example, $e(n)e(m)$ or $e(n)e(m)e(l)$, could be neglected we get

$$\hat{X}_P \cong \prod_{n=0}^{N-1} x(n) + \sum_{m=0}^{N-1} \prod_{\substack{n=0 \\ n\neq m}}^{N-1} x(n)e(m).$$

The quantization effect caused error is

$$e_{X_P} = \sum_{m=0}^{N-1} \prod_{\substack{n=0 \\ n\neq m}}^{N-1} x(n)e(m).$$

It is interesting to note that the relative error is additive since

$$r_{X_P} = \frac{e_{X_P}}{X_P} = \frac{\sum_{m=0}^{N-1} \prod_{\substack{n=0 \\ n\neq m}}^{N-1} x(n)e(m)}{\prod_{n=0}^{N-1} x(n)} = \sum_{m=0}^{N-1} \frac{e(m)}{x(m)} = \sum_{m=0}^{N-1} r_x(m).$$

The expected value is zero if the rounding is used. The variance is signal-dependent,

$$\sigma^2_{X_p} = \sum_{m=0}^{N-1} \prod_{\substack{n=0 \\ n\neq m}}^{N-1} x^2(n)\mathrm{var}\{e(n)\} = \frac{1}{12}\Delta^2 \sum_{m=0}^{N-1} \prod_{\substack{n=0 \\ n\neq m}}^{N-1} x^2(n).$$

7.7.2 Quantization of the results

In the quantization of results, after basic arithmetic operations are performed, we can distinguish two cases. The first case is when the fixed-point arithmetic is used. The register here assumes that the decimal point is positioned at the fixed place. All data are written with respect to this assumed decimal point position. In the floating-point arithmetic, the numbers are written in the sign-mantissa-exponent format. The quantization error is then produced on mantissa only.

7.7.2.1 Fixed point arithmetic

Fixed point arithmetic assumes that the decimal point position is at a fixed place. The common assumption is that all input values and the intermediate results, in this case, are normalized so that $0 \leq x(n) < 1$ or $-1 < x(n) < 1$, if the sign bit is used.

In multiplications, the result of a multiplication

$$x_Q(n)x_Q(m)$$

will, in general, produce a result of $2b$ digits. It should be quantized in the same way as the input signal

$$Q[x_Q(n)x_Q(m)] = x_Q(n)x_Q(m) + e(n,m)$$

where $e(n,m)$ is the quantization error satisfying all the previous properties with

$$-\frac{1}{2}\Delta \leq e(m,n) \leq \frac{1}{2}\Delta.$$

Example 7.60. Find the expected value of the quantization error for

$$r(n) = \sum_{m=0}^{N-1} x(n+m)x(n-m)$$

where $x(n)$ is quantized and the product of signals is quantized to b bits as well. Assume that the signal values are such that their additions will not cause overflow.

★For this calculation, the model is

$$r_Q(n) = \sum_{m=0}^{N-1} \left[x_Q(n+m)x_Q(n-m) + e(n+m,n-m)\right]$$

$$= \sum_{m=0}^{N-1} \{[x(n+m)+e(n+m)]\,[(x(n-m)+e(n-m)] + e(n+m,n-m)\}.$$

The expected value is

$$\mathrm{E}\{r_Q(n)\} = \sum_{m=0}^{N-1} x(n+m)x(n-m) + \mathrm{E}\{\sum_{m=0}^{N-1} e(n+m)e(n-m)\}$$

$$= r(n) + \mathrm{E}\{e^2(n)\} = r(n) + \frac{1}{12}\Delta^2,$$

since it is assumed that errors for two different signal samples are not correlated $\mathrm{E}\{e(n+m)e(n-m)\} = 0$ for $m \neq 0$ and the signal and errors are not correlated, $\mathrm{E}\{x(n+m)e(n-m)\} = 0$, for any m and n.

In general, the additions cause quantization errors as well. Namely, by adding two values $0 \leq x(n) < 1$, the result could be greater than 1. In order to avoid the overflow, the input values are shifted in the register to the right (appropriately divided), causing the quantization error.

In the case when complex-valued numbers are used in calculation, the quantization of the real part and the imaginary part is done separately,

$$x_Q(n) = Q[x(n)] = Q[\mathrm{Re}\{x(n)\} + j\,\mathrm{Im}\{Q[x(n)]\}] = x(n) + e_r(n) + je_i(n).$$

Since the real and imaginary part are independent, with the same variance, the variance of the quantization error for a complex-valued signal is given by

$$\sigma_e^2 = 2\frac{1}{12}\Delta^2 = \frac{1}{6}\Delta^2.$$

For the additions the variance is doubled as well.

In case of multiplications one complex-valued multiplication requires four real-valued multiplications, introducing four errors. The quantization variance of a complex-valued multiplication is

$$\sigma_e^2 = 4\frac{1}{12}\Delta^2 = \frac{1}{3}\Delta^2.$$

If the values of a signal $x(n)$ are not small we have to ensure that no overflow occurs during the calculations using the fixed-point arithmetic. Consider a real-valued random white signal whose samples are within $-1 < x(n) < 1$, with the variance σ_x^2. The registers of $b+1$ bits are assumed, with one bit being used for the sign. As an example consider the expected value calculation

$$X_N = \frac{1}{N}\sum_{n=0}^{N-1} x(n).$$

We have to be sure that an overflow will not occur during the expected value calculation. All sums should stay within the interval $(-1,1)$.

One approach to calculate X_N is in dividing the input signal values by N and summing them, that is

$$X_N = \frac{x(0)}{N} + \frac{x(1)}{N} + \cdots + \frac{x(N-1)}{N}.$$

Then we are sure that no result will be outside the interval $(-1,1)$. Division of the signal samples by N introduces an additive quantization noise,

$$\hat{X}_N = \frac{x(0)}{N} + e(0) + \frac{x(1)}{N} + e(1) + \cdots + \frac{x(N-1)}{N} + e(N-1).$$

Variance of the equivalent noise $e(0) + e(1) + \cdots + e(N-1)$ is

$$\sigma_e^2 = \frac{1}{12}\Delta^2 N = \frac{1}{12}2^{-2b}N.$$

Since the variance of $x(n)/N$ is σ_x^2/N^2, the variance of $\hat{X}_N$ is

$$\sigma_{\hat{X}_N}^2 = N\frac{\sigma_x^2}{N^2} + \frac{1}{12}\Delta^2 N.$$

Ratio of the variances corresponding to the signal and the noise in the result is

$$SNR = \frac{N\frac{\sigma_x^2}{N^2}}{\frac{1}{12}\Delta^2 N} = \frac{1}{N^2}\frac{\sigma_x^2}{\frac{1}{12}\Delta^2} = \frac{1}{N^2}\frac{\sigma_x^2}{\frac{1}{12}2^{-2b}}$$

or in [dB]

$$SNR = 10\log\left(\frac{1}{N^2}\frac{\sigma_x^2}{\frac{1}{12}2^{-2b}}\right) = 20\log\sigma_x - 20\log N - 20\log 2^{-b} + 10\log(12)$$

$$= 20\log\sigma_x - 20\frac{\log_2 N}{\log_2 10} - 20\frac{\log_2 2^{-b}}{\log_2 10} + 10.8 = 20\log\sigma_x - 6.02(m-b) + 10.8,$$

where $N = 2^m$. Obviously, by increasing the number of samples N to $2N$ will keep the same SNR if b is increased for one bit, since $(m+1-(b+1)) = m-b$.

Another way to calculate the mean is in performing the summation step by step, according to the presented scheme, for $N = 8$,

$$X_N = \frac{\frac{\frac{x(0)}{2}+\frac{x(1)}{2}}{2}+\frac{\frac{x(2)}{2}+\frac{x(3)}{2}}{2}}{2}+\frac{\frac{\frac{x(4)}{2}+\frac{x(5)}{2}}{2}+\frac{\frac{x(6)}{2}+\frac{x(7)}{2}}{2}}{2}.$$

Here, two adjunct signal values $x(n)$ are divided by $1/2$ first. They are added then, avoiding possible overflows. The error in one step is

$$\frac{x(n)}{2}+e(n)+\frac{x(n+1)}{2}+e(n+1)=\frac{x(n)+x(n+1)}{2}+e_n^{(2)}.$$

The error

$$e_n^{(2)} = e(n)+e(n+1)$$

has the variance

$$\text{var}\left\{e_n^{(2)}\right\} = \frac{1}{12}\Delta^2+\frac{1}{12}\Delta^2 = \frac{1}{6}\Delta^2.$$

After every division by 2, the result is shifted in the register to the right and a quantization error is created. Thus, the error model, due to the addition quantization, is

$$\begin{aligned}
\hat{X}_N &= \frac{\frac{\frac{x(0)}{2}+\frac{x(1)}{2}+e_0^{(2)}}{2}+\frac{\frac{x(2)}{2}+\frac{x(3)}{2}+e_2^{(2)}}{2}+e_0^{(4)}}{2}+\frac{\frac{\frac{x(4)}{2}+\frac{x(5)}{2}+e_4^{(2)}}{2}+\frac{\frac{x(6)}{2}+\frac{x(7)}{2}+e_6^{(2)}}{2}+e_4^{(4)}}{2}+e_0^{(8)} \\
&= \frac{x(0)}{N}+\frac{x(1)}{N}+\cdots+\frac{x(N-1)}{N} \\
&+ \frac{e_0^{(2)}}{N/2}+\frac{e_2^{(2)}}{N/2}+\cdots+\frac{e_{N-2}^{(2)}}{N/2} \\
&+ \frac{e_0^{(4)}}{N/4}+\cdots+\frac{e_{N-4}^{(4)}}{N/4} \\
&\vdots \\
&+ \frac{e_0^{(N)}}{N/N}.
\end{aligned} \tag{7.144}$$

The variance of all quantization noises is the same $\sigma_e^2 = \frac{1}{6}\Delta^2 = \frac{1}{6}2^{-2b}$. Notice that the noises in the first stage are divided by $N/2$, due to divisions by 2 in the next stages of summation. Their variance is reduced for $N^2/4$. The value of the variance of errors in these stages is

$$\begin{aligned}
\text{var}\{\frac{e_0^{(2)}}{N/2}+\frac{e_2^{(2)}}{N/2}+\cdots+\frac{e_{N-2}^{(2)}}{N/2}\} &= \frac{1}{6}\Delta^2\frac{1}{N^2/4}\frac{N}{2} = \frac{1}{6}\Delta^2\frac{2}{N} \\
\text{var}\{\frac{e_0^{(4)}}{N/4}+\cdots+\frac{e_{N-4}^{(4)}}{N/4}\} &= \frac{1}{6}\Delta^2\frac{1}{N^2/16}\frac{N}{4} = \frac{1}{6}\Delta^2\frac{4}{N} \\
&\vdots \\
\text{var}\{\frac{e_0^{(N)}}{N/N}\} &= \frac{1}{6}\Delta^2\frac{1}{N^2/N^2}\frac{N}{N} = \frac{1}{6}\Delta^2\frac{N}{N} = \frac{1}{6}\Delta^2\frac{2^m}{N}.
\end{aligned}$$

where $2^m = N$. The total variance of $\hat{X}_N$ is

$$\begin{aligned}\sigma^2_{\hat{X}_N} &= N\frac{\sigma_x^2}{N^2} + \frac{1}{6}\Delta^2\frac{2}{N} + \frac{1}{6}\Delta^2\frac{4}{N} + \cdots + \frac{1}{6}\Delta^2\frac{2^m}{N} \\ &= \frac{\sigma_x^2}{N} + \frac{1}{6}\Delta^2\frac{2}{N}(1+2+\cdots+2^{m-1}) = \frac{\sigma_x^2}{N} + \frac{1}{6}\Delta^2\frac{2}{N}\frac{1-2^m}{1-2} \\ &= \frac{\sigma_x^2}{N} + \frac{1}{6}\Delta^2\frac{2}{N}(N-1) = \frac{\sigma_x^2}{N} + \frac{1}{3}\Delta^2(1-\frac{1}{N}).\end{aligned} \tag{7.145}$$

Ratio of the variances $\frac{\sigma_x^2}{N}$ and $\frac{1}{3}\Delta^2(1-\frac{1}{N})$, corresponding to the output signal-to-noise ratio, is

$$SNR = \frac{\frac{\sigma_x^2}{N}}{\frac{1}{3}\Delta^2(1-\frac{1}{N}).} = \frac{\sigma_x^2}{\frac{1}{3}\Delta^2(N-1)} \cong \frac{1}{N}\frac{\sigma_x^2}{\frac{1}{3}2^{-2b}} = 3\sigma_x^2 2^{2(b-m/2)}.$$

Significant improvement (for an order of N) is obtained using this scheme for the summation, instead of the direct one. In dB the ratio is

$$SNR \cong 10\log\left(3\sigma_x^2 2^{2(b-m/2)}\right) = 20\log\sigma_x - 6.02(\frac{m}{2} - b) + 4.8.$$

If the signal values were complex then $2^{-2b}/12$ would be changed to $2^{-2b}/6$.

7.7.2.2 Discrete rounding error

The previous results are common in literature. They are derived assuming that the variances of the errors are the same and obtained assuming uniform nature of the quantization errors. However these results differ from the ones obtained by statistical analysis. The reason is in the quantization error distribution and variance. Namely, after the high precision signal $x(n)$ is divided by 2 and stored into $(b+1)$-bit registers, the errors in $x(n)/2 + e(n)$ are uniform, with $-\Delta/2 \leq e(n) < \Delta/2$. When these values are stored into registers, then in every next stage, when we calculate $[x(n)/2 + e(n)] + [x(n+1)/2 + e(n+1)]/2$ the input values $x(n)/2 + e(n)$ and $x(n+1)/2 + e(n+1)$ are already stored in the $(b+1)$-bit registers. Division by 2 is just a one bit shift to the right. This shift cases one bit error. Therefore this one bit error is discrete in amplitude

$$e_d \in \{-\Delta/2, 0, \Delta/2\},$$

with probabilities

$$P_d(\pm\Delta/2) = 1/4 \quad \text{and} \quad P_d(0) = 1/2.$$

The expected value of this kind of error is zero, provided that the rounding is done in such a way that it takes values $\pm\Delta/2$ with equal probability (various *tie-breaking algorithms* for rounding exist). The variance of $e_d^{(i)}$ is

$$\text{var}\left\{e_n^{(i)}\right\} = 2\text{var}\{e_d\} = 2[\frac{1}{4}(-\frac{\Delta}{2})^2 + \frac{1}{4}(\frac{\Delta}{2})^2] = \frac{1}{4}\Delta^2, \text{ for } i > 2.$$

The total variance of $\hat{X}_N$ is then of the form

$$\sigma^2_{\hat{X}_N} = N\frac{\sigma_x^2}{N^2} + \frac{1}{4}\Delta^2\frac{2}{N} + \frac{1}{4}\Delta^2\frac{4}{N} + \cdots + \frac{1}{4}\Delta^2\frac{2^m}{N} = \frac{\sigma_x^2}{N} + \frac{1}{2}\Delta^2(1-\frac{4}{3N}),$$

instead of (7.145). Signal-to-noise ratio is given by

$$SNR = \frac{\frac{\sigma_x^2}{N}}{\frac{1}{2}\Delta^2(1-\frac{4}{3N})} \cong 2\sigma_x^2 2^{2(b-m/2)}.$$

The previous analysis corresponds to the calculation of the DFT coefficient $X(0)$ when the input signal is a random uniform signal, whose values are within $-1 < x(n) < 1$, with variance σ_x^2. A model for the element $X(k)$, with all quantization errors included, is

$$\hat{X}(k) = \frac{1}{N}\sum_{n=0}^{N-1}\left\{[x(n)+e_i(n)]\,W_N^{nk} + e_m(n)\right\} = \sum_{n=0}^{N-1} y(n),$$

where $e_i(n)$ is the input signal quantization error and $e_m(n)$ is the multiplication quantization error. The variances for complex-valued signals are

$$\text{var}\{e_i(n)\} = 2\frac{1}{12}\Delta^2 = \frac{1}{6}\Delta^2, \quad \text{var}\{e_m(n)\} = 4\frac{1}{12}\Delta^2 = \frac{1}{3}\Delta^2.$$

Moreover, we have to provide that the additions do not produce an overflow. If we use the calculation scheme, presented for $N = 8$, as

$$\hat{X}(k) = \frac{\frac{\frac{y(0)}{2}+\frac{y(1)}{2}+e_0^{(2)}}{2} + \frac{\frac{y(2)}{2}+\frac{y(3)}{2}+e_2^{(2)}}{2} + e_0^{(4)}}{2} + \frac{\frac{\frac{y(4)}{2}+\frac{y(5)}{2}+e_4^{(2)}}{2} + \frac{\frac{y(6)}{2}+\frac{y(7)}{2}+e_6^{(2)}}{2} + e_4^{(4)}}{2} + e_0^{(8)},$$

then in every addition, the terms should be divided by 2. This division introduces the quantization error. In the first step,

$$\frac{y(n)}{2} + e(n) + \frac{y(n+1)}{2} + e(n+1) = \frac{1}{2}\{[x(n)+e_i(n)]\,W_N^{nk} + e_m(n) + [x(n+1)+e_i(n+1)]W_N^{(n+1)k} + e_m(n+1)\} + e(n) + e(n+1).$$

The total error in this step is

$$e_n^{(2)} = \frac{e_i(n)W_N^{nk} + e_m(n) + e_i(n+1)W_N^{(n+1)k} + e_m(n+1)}{2} + e(n) + e(n+1)$$

with the variance

$$\text{var}\{e_n^{(2)}\} = \frac{1}{4}\left(\frac{1}{6}\Delta^2 + \frac{1}{3}\Delta^2 + \frac{1}{6}\Delta^2 + \frac{1}{3}\Delta^2\right) + 2\frac{1}{6}\Delta^2 = \frac{7}{12}\Delta^2.$$

In all other steps, within the errors $e_0^{(4)}$ to $e_0^{(N)}$, just the addition errors appear. Their variance, for complex-valued terms, is

$$\text{var}\{e_n^{(i)}\} = 2\frac{1}{6}\Delta^2.$$

Therefore, the variance of

$$\hat{X}_N = \frac{x(0)}{N} + \frac{x(1)}{N} + \cdots + \frac{x(N-1)}{N} + \frac{e_0^{(2)}}{N/2} + \frac{e_2^{(2)}}{N/2} + \cdots + \frac{e_{N-2}^{(2)}}{N/2} + \frac{e_0^{(4)}}{N/4} + \cdots + \frac{e_{N-4}^{(4)}}{N/4} + \ldots + \frac{e_0^{(N)}}{N/N} \tag{7.146}$$

is obtained using

$$\begin{aligned}
\text{var}\{\frac{e_0^{(2)}}{N/2}+\frac{e_2^{(2)}}{N/2}+\cdots+\frac{e_{N-2}^{(2)}}{N/2}\} &= \frac{7}{12}\Delta^2\frac{1}{N^2/4}\frac{N}{2}=\frac{7}{12}\Delta^2\frac{2}{N}\\
\text{var}\{\frac{e_0^{(4)}}{N/4}+\cdots+\frac{e_{N-4}^{(4)}}{N/4}\} &= \frac{1}{3}\Delta^2\frac{1}{N^2/16}\frac{N}{4}=\frac{1}{3}\Delta^2\frac{4}{N}\\
&\cdots\\
\text{var}\{\frac{e_0^{(N)}}{N/N}\} &= \frac{1}{3}\Delta^2\frac{1}{N^2/N^2}\frac{N}{N}=\frac{1}{3}\Delta^2\frac{2^m}{N}.
\end{aligned}$$

The total variance of $\hat{X}_N$ is

$$\begin{aligned}
\sigma^2_{\hat{X}_N} &= \frac{\sigma_x^2}{N}+\frac{1}{3}\Delta^2\frac{2}{N}(\frac{3}{4}+1+2+\cdots+2^{m-1})\\
&= \frac{\sigma_x^2}{N}+\frac{2}{3}\Delta^2\frac{N-\frac{1}{4}}{N}\cong\frac{\sigma_x^2}{N}+\frac{2}{3}\Delta^2
\end{aligned}$$

with

$$SNR = 10\log\frac{3\sigma_x^2}{2N\Delta^2}=20\log\sigma_x-6.02(\frac{m}{2}-b)+1.76.$$

If the described discrete nature of the quantization error amplitude, after the first quantization step, is taken into account (provided that the rounding is done in such a way that the error takes values $\pm\Delta/2$ with equal probability), then with

$$\text{var}\left\{e_n^{(i)}\right\}=4\text{var}\{e_d\}=\frac{1}{2}\Delta^2,$$

for $i>2$, the variance of $\hat{X}_N$ follows

$$\sigma^2_{\hat{X}_N}=\frac{\sigma_x^2}{N}+\frac{\Delta^2}{N}(\frac{7}{6}+2+4+\cdots+2^{m-1})=\frac{\sigma_x^2}{N}+\Delta^2\frac{N-\frac{5}{6}}{N}\cong\frac{\sigma_x^2}{N}+\Delta^2.$$

If the FFT is calculated using the fixed-point arithmetic and the signal is uniform, distributed within $-1<x(n)<1$, with the variance σ_x^2, then in order to avoid an overflow the signal could be divided at the input with N and the standard FFT could be used, as in Fig. 7.36.

An improvement in the SNR can be achieved if the scaling is done not to the input signal $x(n)$ by N, but by $1/2$ in every butterfly, as shown in Fig. 7.37. The improvement achieved here is due to the fact that the quantization errors appearing in the early butterfly stages are divided by $1/2$ and therefore reduced at the output, as in (7.144). An improvement of the order of N is achieved in the output signal-to-noise ratio.

7.7.2.3 Floating point arithmetic

Fixed point arithmetic is simple, but could be inefficient if the signal values within a wide range of amplitudes are expected. For example, if we can expect the signal values

$$\begin{aligned}
x_Q(n_1) &= 1011111110101.010\\
x_Q(n_2) &= 0.0000000000110101,
\end{aligned}$$

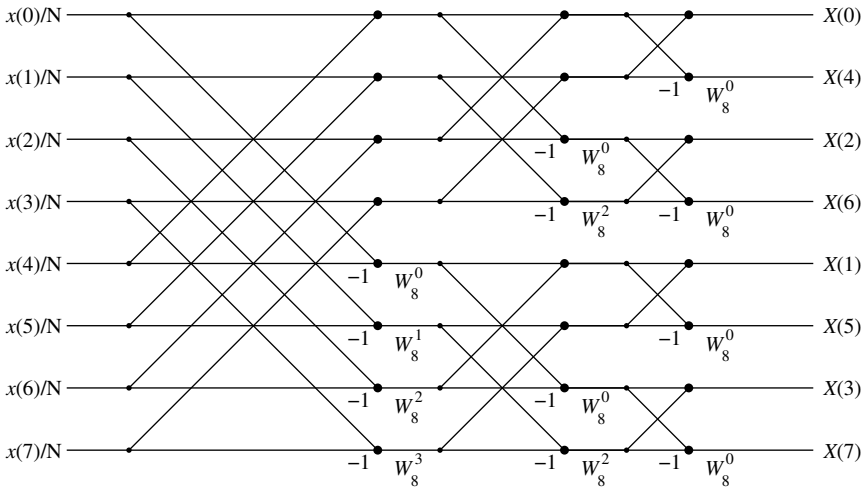


Figure 7.36 The FFT calculation scheme obtained using the decimation-in-frequency for $N = 8$ with the signal being divided by N in order to avoid an overflow when the fixed point arithmetic is used.

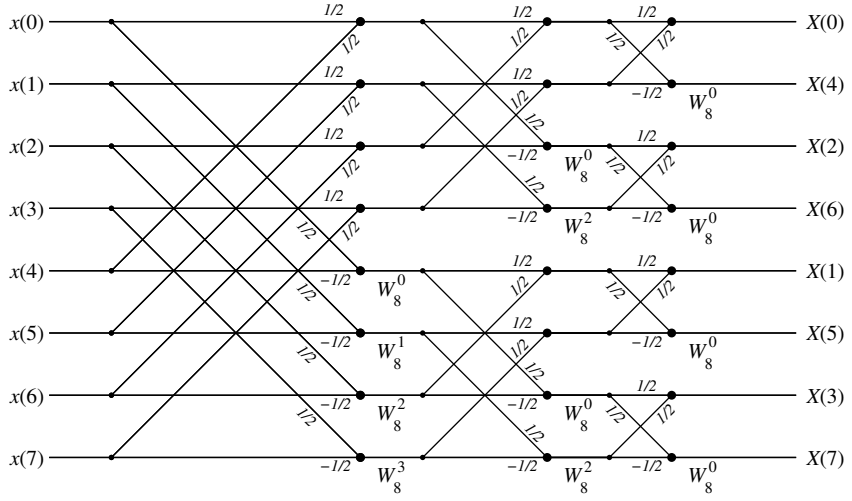


Figure 7.37 The FFT calculation scheme obtained using the decimation-in-frequency for $N = 8$ with the signal being divided by $1/2$ in every butterfly in order to avoid an overflow when the fixed-point arithmetic is used.

then obviously fixed-point arithmetic would require large registers so that both values can be stored without loosing their significant digits. However, we can represent these signal values into the exponential form as

$$x_Q(n_1) = 1.011111110101010 \times 2^{12}$$
$$x_Q(n_2) = 1.10101 \times 2^{-11}.$$

The exponential format of numbers is then written within the register in the following format

s_n	s_e	e_1	e_2	e_3	e_4	e_5	e_6	e_7	m_{-1}	m_{-2}	m_{-3}	$\dots$	m_{-b}

where:

s_n is the sign of number (1 for a positive number and 0 for a negative number)
s_e is the sign of exponent (1 for a positive exponent and 0 for a negative exponent)
$e_1 e_2 \dots e_7$ is the binary format of the exponent, and
$m_{-1} m_{-2} \cdots_{-b}$ is the mantissa, assuming that the integer value is always 1, it is omitted.

Within this format, the previous signal value $x_Q(n_1)$, with a register of 19 bits in total, is

1	1	**0**	**0**	**0**	**1**	**1**	**0**	**0**	0	1	1	1	1	1	1	0	1

while $x_Q(n_2)$ is

1	0	**0**	**0**	**0**	**1**	**0**	**1**	**1**	1	0	1	0	1	0	0	0	0

If the exponent cannot be written within the defined number of bits (here 7), the computer has to stop the calculation and indicate "*overflow*", that is, the number cannot fit into the register. For mantissa, the values are just rounded to the available number of bits. In the implementations based on the floating-point arithmetic, the quantization affects the mantissa only. The relative error in mantissa is again

$$|e(n)| \leq 2^{-(b+1)} = \frac{1}{2}\Delta.$$

The error in signal is multiplied by the exponent. Since we can say that the exponent value is of the signal order, we can write

$$Q[x(n)] = x_Q(n) = x(n) + e(n)x(n) = x(n)(1 + e(n)).$$

The quantization error behaves here as a multiplicative uniform noise. Thus, for the floating-point representation, multiplicative errors appear.

The floating-point additions also produce the quantization errors, which are represented by a multiplicative noise. During additions, the number of bits may increase. This increase in the number of bits requires mantissa shift, what causes multiplicative error.

In addition to the *IEEE standard* when the total number of bits is 32 (23 for mantissa and 7 for exponent) we will mention two standard formats for the telephone signal coding. The μ-law pulse-coded modulation (PCM) is used in the North America and the A-law PCM is used in European telephone networks. They use 8-bit representations with a sign bit, 3 exponent bits, and 4 mantissa bits

s	e_1	e_2	e_3	m_1	m_2	m_3	m_4

The μ-law encoding takes a 14-bit signed signal value (its two's complement representation) as input, adds 33 (binary 100001) and converts it to an 8-bit value. The encoding formula in the μ-law is

$$(-1)^s \left[2^{e+1}(m + 16.5) - 33 \right].$$

This is a 14-bit signed integer from -8031 to $+8031$.

The sign bit s is set to 1 if the input sample is negative. It is set to 0 if the input sample is positive. Number 0 is written as

$$\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{0}.$$

Example 7.61. As an example consider the positive numbers from $+1$ to $+30$. They are written as $+2^1(m+16.5)-33$ with 15 quantization steps equal to 2 (starting from $m=1$ to $m=15$). Then the numbers from $+31$ to $+94$ are written as $+2^2(m+16.5)-33$ with 16 quantization steps equal to 4 (with m from 0 to 15). The last interval for positive numbers is from $+4063$ to $+8158$ written as $+2^8(m+16.5)-33$ with 16 quantization intervals (with m from 0 to 15) of the width 256. The range of the input values is from -8159 to $+8159$ ($\pm 2^{13}$) with the minimum step size 2 for the smallest amplitudes.

The *compression function* corresponding to the μ-law encoding format of signal $0 \leq |x| \leq 1$ is

$$F(x) = \text{sign}(x)\frac{\ln(1+\mu|x|)}{\ln(1+\mu)}.$$

with $\mu = 255$.

Example 7.62. Write the number $a = 456$ in the binary μ-law format.

★The number to be represented by $2^{e+1}(m+16.5)$ is $456+33=489$. The mantissa range is $0 \leq m \leq 15$. This means that the exponent $(e+1)$ should be such that

$$0+16.5 \leq \frac{489}{2^{e+1}} \leq 15+16.5$$

for the range $16.5 \leq m+16.5 \leq 31.5$. It is easy to conclude that $489/16 = 30.5625$, meaning $e+1=4$ with $m+16.5=30.5625$. The nearest integer value of m is $m=14$. Therefore $\hat{a} = 2^{3+1} \times (14+16.5) - 33 = 455$ is the nearest μ-law format number to a. The binary form is

$$\boxed{0}\boxed{0}\boxed{1}\boxed{1}\boxed{1}\boxed{1}\boxed{1}\boxed{0}.$$

The quantization step for this range of numbers is $2^4 = 16$. It means that the closest possible smaller number is 439, while the next possible larger number would be 471. It is the last number with the quantization step $2^{e+1} = 16$.

Example 7.63. Write a model for the calculation of

$$r(n,m) = x(n+m)x(n-m)$$

if the quantization error is caused by the floating-point registers with b bits for the mantissa. What is the expected value? Write the model for

$$y(n) = x(n) + x(n+1).$$

The considered signals are real-valued.

★The quantization model for this calculation is given by

$$\hat{r}(n,m) = x(n+m)(1+e(n+m))x(n-m)(1+e(n-m))(1+e(n+m,n-m)).$$

The expected value is

$$\begin{aligned}\mathrm{E}\{\hat{r}(n)\} &= x(n+m)x(n-m) + \mathrm{E}\{e(n+m)e(n-m)\} \\ &= r(n) + \mathrm{E}\{e^2(n)\}\delta(m) = r(n) + \frac{1}{12}\Delta^2\delta(m).\end{aligned}$$

For $y(n)$, the model is

$$\hat{y}(n) = [x(n)(1+e(n)) + x(n+1)(1+e(n+1))](1+e(n,n+1)),$$

where $e(n,n+1)$ is the is the multiplicative noise that models the addition error.

7.8 PROBLEMS

Problem 7.1. Signal $x_{20i}(n)$, for $i = 01,02,..,15$, is the monthly average of the maximum daily temperatures in a city, measured from the year 2001 to 2015. The values of this signal are given in Table 7.2. If we can assume that the signal for every individual month is Gaussian, find the probability that the average of maximum daily temperatures: (a) in January is lower than 2, (b) in January is higher than 12.

Problem 7.2. Available are M realizations of the random variable $x_i(n)$, $i = 1,2,\ldots,M$, at an instant n. The variance of $x(n)$ is estimated in two possible scenarios:

(a) The mean value is know in advance and it is equal to zero, $\mu_x(n) = 0$.

(b) The mean value is not known and it is estimated from data as

$$\mu_x(n) = \frac{1}{M}\left(x_1(n) + x_2(n) + \cdots + x_M(n)\right)$$

How the estimate of the variance in (a) is related to the variance estimate in (b)?

Problem 7.3. A random variable $x(n)$ is recorded in $N = 10$ trials

$$\mathbf{x} = [0.26,\ 0.31, 0.64,\ 0.99, 1.00,\ 0.92,\ 0.85,\ 0.73, 0.58, 0.15]^T,$$

with different independent random variable, t_n, values given in the vector form

$$\mathbf{t} = [-0.8, -0.83, -0.60, -0.10, -0.01,\ 0.28,\ 0.39, 0.52,\ 0.65,\ 0.92]^T.$$

The random variable $x(n)$ is modeled using the fifth-order polynomial

$$x(n) = a_0 + a_1 t_n + a_2 t_n^2 + a_3 t_n^3 + a_4 t_n^4 + a_5 t_n^5$$

or in the matrix form

$$\mathbf{x} = \mathbf{Ta},$$

where $\mathbf{T}$ is the matrix with columns $\mathbf{t}^m$, that is $\mathbf{T} = [\mathbf{t}^0,\ \mathbf{t}^1, \ldots, \mathbf{t}^5]$, and $\mathbf{t}^m$ is the notation for the column vector with the elements t_n^m, $n = 1,2,\ldots,N$ (see the form matrix definition in (7.14)).

(a) Estimate the model parameters using the polynomial fitting (the least-squares solution to the minimization of $J(\mathbf{a}) = ||\mathbf{x} - \mathbf{Ta}||_2^2$),

$$\hat{\mathbf{a}} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{x}. \tag{7.147}$$

.

(b) Estimate the model parameters with the ridge regression model (the solution to the minimization of $J(\mathbf{a}) = ||\mathbf{x} - \mathbf{Ta}||_2^2 + \lambda||\mathbf{a}||_2^2$) in the form

$$\hat{\mathbf{a}} = (\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\mathbf{x}, \tag{7.148}$$

with $\lambda = 0.1$.

(c) Repeat the calculations in (a) and (b) with an increased additive noise in the data

$$\mathbf{x} = [0.35,\ 0.33,\ 0.57,\ 0.92,\ 0.94,\ 0.89,\ 0.87,\ 0.86,\ 0.44,\ 0.29]^T.$$

(d) Predict the value $x(1.12)$ in all considered cases. Use the result in (a) as the reference.

(e) Find the bias and the covariance matrix of the regression ridge estimator as a function of λ, when the noise in the true data $\mathbf{s}$ is white, with the variance σ_ε^2 and the considered signal is $\mathbf{x} = \mathbf{s} + \boldsymbol{\varepsilon}$ (*advanced topic*).

Problem 7.4. The cumulative probability distribution function $F(\chi)$ is given as

$$F(\chi) = \begin{cases} 0 & \chi \leq 0 \\ \chi/2 & 0 < \chi \leq 1 \\ 1/2 & 1 < \chi \leq 2 \\ (\chi-1)/2 & 2 < \chi \leq 3 \\ 1 & \chi > 3. \end{cases}$$

Find the probability density function $p(\xi)$ and the probability that $x(n) < 2.5$.

Problem 7.5. The probability density function of a random variable $x(n)$ is

$$p_x(\xi) = ae^{-b|\xi|}, -\infty < \xi < \infty,$$

where a and b are the constants. Find the relation between a and b. What is the cumulative probability distribution function for $a = 1$?

Problem 7.6. A random signal $x(n)$ is characterized by the probability density function

$$p_x(\xi) = \frac{\lambda}{2} e^{-\lambda|\xi|}, \ \lambda > 0.$$

Find the expected value and variance of $x(n)$.

Problem 7.7. The joint probability density function of signals $x(n)$ and $y(n)$ is

$$p_{xy}(\xi,\zeta) = \begin{cases} k\xi e^{-\xi(\zeta+1)}, & 0 \leq \xi < \infty \ \ 0 \leq \zeta < \infty \\ 0 & \text{elsewhere.} \end{cases}$$

Find the value of constant k.

Problem 7.8. Consider two independent random signals $x(n)$ and $y(n)$ with probability density functions $p_{x(n)}(\xi)$ and $p_{y(n)}(\xi)$. A new random signal is defined is such a way that it takes the greater value of the signals $x(n)$ and $y(n)$ at each instant n,

$$z(n) = \max\{x(n), y(n)\}.$$

Find the probability distribution and the probability density function of the random signal $z(n)$.

Problem 7.9. A set of $N = 10$ balls is considered, with an equal number of balls being marked with 1 (or white) and 0 (or black). A random signal $x(n)$ corresponds to drawing four balls in a row. It has four samples $x(0)$, $x(1)$, $x(2)$, and $x(3)$ corresponding to these draws. The signal values are equal to the number (color) associated with the randomly drawn ball. If k is the number of values 0 that appear in the signal (number of black balls), write the probability for $k = 0$. Generalize the result for N balls and M signal samples.

Problem 7.10. The random signal $x(n)$ is zero-mean Gaussian distributed with the the probability density function

$$p_x(\xi) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\xi^2/(2\sigma_x^2)}.$$

Show that the variance of this random variable is equal to σ_x^2.

Problem 7.11. The random signal $x(n)$ is zero-mean Gaussian distributed random variable with the variance σ_2^2. Find the median of $x(n)$ and the median of $|x(n)|$.

Problem 7.12. (a) Consider a zero-mean Gaussian distributed random noise $\varepsilon(n)$ with variance σ_ε^2. Find the variance of $y(n) = \varepsilon(n) - \varepsilon(n-1)$ and relate it to the sample median of $|y(n)| = |\varepsilon(n) - \varepsilon(n-1)|$.

(b) Sow that, if a signal $x(n)$ consists of a slow-varying deterministic signal $s(n)$ such that $|s(n) + \varepsilon(n) - s(n-1) - \varepsilon(n-1)| \approx |\varepsilon(n) - \varepsilon(n-1)|$, the noise standard deviation can estimated using

$$\hat{\sigma}_\varepsilon = \frac{1}{\sqrt{2}} \frac{1}{0.6745} \operatorname*{median}_{n=2,3,\ldots,N} \{ \Big| x(n) - x(n-1) \Big| \}.$$

(c) Check this result on the signal and noise from Example 7.45.

Problem 7.13. The random signal $x(n)$ is such that $x(n) = 0$ with probability 0.8. In all other cases $x(n)$ is Gaussian random variable with the expected value 3 and the variance equal to 2. Find the expected value and the variance of $x(n)$.

Problem 7.14. The signal $\varepsilon(n)$ is a Gaussian noise with the expected value $\mu_\varepsilon = 0$ and the variance σ_ε^2. Find the probability that $|\varepsilon(n)| > A$. If the signal length is $N = 2000$, find the expected number of samples with amplitudes higher than $A = 10$, assuming that $\sigma_\varepsilon^2 = 2$. What is the result for $A = 4$ and $\sigma_\varepsilon^2 = 2$.

Problem 7.15. The random signal $x(n)$ is a Gaussian noise with the expected value 0 and the variance σ_x^2. The signal has a large number N of samples. A random sequence $y(n)$ is formed using M samples from the signal $x(n)$ with the lowest amplitudes. Find μ_y and σ_y.

Problem 7.16. Consider the signal $s(n) = A\delta(n - n_0)$ and a zero-mean Gaussian noise $\varepsilon(n)$ with variance σ_ε^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 the event A that the maximum value of $x(n) = s(n) + \varepsilon(n)$ is obtained at $n = n_0$.

Problem 7.17. The random signal $x(n)$ is a Gaussian noise with the expected value 0 and the variance σ_x^2. A random sequence $y(n)$ is formed by omitting the samples from the signal $x(n)$ whose amplitudes are higher than A. Find the probability density function of the sequence $y(n)$. Find μ_y and σ_y.

Problem 7.18. The signal samples $x(n)$ are such that

$$x(n) = \begin{cases} A + \varepsilon(n), & \text{for } n \in \mathbb{N}_x \\ \varepsilon(n), & \text{otherwise} \end{cases}$$

where $\varepsilon(n)$ is a Gaussian noise with the expected value $\mu_\varepsilon = 0$ and the variance σ_ε^2, $A > 0$ is a constant and $\mathbb{N}_x$ is a nonempty set of discrete-time instants. The threshold-based criterion is used to detect if an arbitrary time instant n belongs to the set $\mathbb{N}_x$

$$n \in \mathbb{N}_x \quad \text{if } x(n) > T,$$

where T is the threshold. Find the value of threshold T if the probability of false detection is 0.01.

Problem 7.19. The signal $x(n)$ is a random Gaussian sequence with the expected value $\mu_x = 5$ and the variance $\sigma_x^2 = 1$. The signal $y(n)$ is a random Gaussian sequence, independent from $x(n)$, with the expected value $\mu_y = 1$ and the variance $\sigma_y^2 = 1$. If we consider $N = 1000$ samples of these signals, find the expected number of time instants where $x(n) > y(n)$ holds.

Problem 7.20. Let $x(n)$ and $y(n)$ be independent real-valued white Gaussian random variables with expected values $\mu_x = \mu_y = 0$ and the variances σ_x^2 and σ_y^2. Show that the random variable

$$z = \frac{1}{M} \sum_{n=1}^{M} x(n)y(n)$$

has the variance

$$\sigma_z^2 = \frac{1}{M}\sigma_x^2\sigma_y^2.$$

Problem 7.21. Find the moments, M_i, and the cumulants, K_i, (up to the fourth-oder) of the Gaussian distributed random variable

$$p_x(\theta) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(\xi-\mu)^2/(2\sigma^2)}.$$

The moments can be calculated form

$$\Phi_x(-j\theta) = \int_{-\infty}^{\infty} p_{x(n)}e^{\theta\xi}d\xi = 1 + \theta M_1 + \frac{1}{2!}\theta^2 M_2 + \frac{1}{3!}\theta^3 M_3 + \dots , \tag{7.149}$$

while the cumulants follow from

$$\ln(\Phi_x(-j\theta)) = 1 + \theta K_1 + \frac{1}{2!}\theta^2 K_2 + \frac{1}{3!}\theta^3 K_3 + \dots , \tag{7.150}$$

The function $M_x(\theta) = \Phi_x(-j\theta) = \mathrm{E}\{e^{\theta\xi}\}$ is called the moment generating function.

Problem 7.22. A random signal $\varepsilon(n)$ is stationary and Cauchy distributed with the probability density function

$$p_{\varepsilon(n)}(\xi) = \frac{a}{1+\xi^2}.$$

Find the coefficient a, expected value, and the variance of this signal.

Problem 7.23. Find the expected value and the variance of the Poisson distributed random variable

$$P(x(n) = k) = P(k) = \frac{\lambda^k e^{-\lambda}}{k!} \text{ for } \lambda > 0.$$

Problem 7.24. The causal system is defined by

$$y(n) = x(n) + 0.5y(n-1).$$

The input signal is $x(n) = a\delta(n)$ with the random amplitude a. The random variable a is uniformly distributed within the interval from 4 to 5. Find the expected value and autocorrelation of the output signal. Is the output signal WSS?

Problem 7.25. Consider the Hilbert transformer with the impulse response

$$h(n) = \begin{cases} \frac{2}{\pi}\frac{\sin^2(n\pi/2)}{n}, & n \neq 0 \\ 0, & n = 0 \end{cases}.$$

The input signal to this transformer is a white noise with the variance equal to 1.

(a) Find the autocorrelation function of the output signal.

(b) Find the cross-correlation of the input and the output signal. Show that the cross-correlation is an antisymmetric function.

(c) Find the autocorrelation and the power spectral density function of the analytic signal $\varepsilon_a(n) = \varepsilon(n) + j\varepsilon_h(n)$, where $\varepsilon_h(n) = \varepsilon(n) *_n h(n)$.

Problem 7.26. Consider the causal system

$$y(n) - ay(n-1) = x(n).$$

If the input signal is a white noise $x(n) = \varepsilon(n)$, with the autocorrelation function $r_{\varepsilon\varepsilon}(n) = \sigma_\varepsilon^2\delta(n)$, find the autocorrelation and the power spectral density of the output signal.

Problem 7.27. Consider the linear time-invariant system whose input is

$$x(n) = \varepsilon(n)u(n)$$

and the impulse response is

$$h(n) = a^n u(n),$$

where $\varepsilon(n)$ is a stationary real-valued noise with the expected value μ_ε and the autocorrelation $r_{\varepsilon\varepsilon}(n,m) = \sigma_\varepsilon^2\delta(n-m) + \mu_\varepsilon^2$. Find the expected value and the variance of the output signal.

Problem 7.28. Find the expected value, the autocorrelation, and the power spectral density of the random signal

$$x(n) = \varepsilon(n) + \sum_{k=1}^{N} a_k e^{j(\omega_k n + \theta_k)},$$

where $\varepsilon(n)$ is a stationary real-valued noise with the expected value μ_ε and the autocorrelation $r_{\varepsilon\varepsilon}(n,m) = \sigma_\varepsilon^2\delta(n-m) + \mu_\varepsilon^2$ and θ_k are random variables, uniformly distributed over the interval $-\pi < \theta_k \leq \pi$. All random variables are statistically independent.

Problem 7.29. Find a stable optimal filter if the correlation functions for the desired signal and additive noise are $r_{ss}(n) = 0.25^{|n|}$, $r_{s\varepsilon}(n) = 0$ and $r_{\varepsilon\varepsilon}(n) = \delta(n)$. Discuss the filter causality.

Problem 7.30. Calculate the DFT value $X(2)$ of the signal $s(n) = \exp(j4\pi n/N)$, with $N = 8$, corrupted by the additive noise $\varepsilon(n) = 2001\delta(n) - 204\delta(n-3)$, using

$$X(k) = \sum_{n=0}^{N-1} (s(n) + \varepsilon(n))e^{-j2\pi kn/N}$$

and estimate the DFT using

$$X_R(k) = N \operatorname*{median}_{n=0,1,..,N-1} \operatorname{Re}\left\{(s(n) + \varepsilon(n))e^{-j2\pi kn/N}\right\}$$
$$+ jN \operatorname*{median}_{n=0,1,..,N-1} \operatorname{Im}\left\{(s(n) + \varepsilon(n))e^{-j2\pi kn/N}\right\}.$$

Problem 7.31. The spectrogram is one of the most commonly used tools in time-frequency analysis. Its form is

$$S_x(n,k) = \left|\sum_{i=0}^{N-1} x(n+i)w(i)e^{-j\frac{2\pi}{N}ik}\right|^2$$

where the signal is $x(n) = s(n) + \varepsilon(n)$, with $s(n)$ being the desired deterministic signal and $\varepsilon(n)$ being a complex-valued, zero-mean white Gaussian noise, with the variance σ_ε^2 and independent and identically distributed (i.i.d.) real and imaginary parts. The window function is $w(i)$. Using the rectangular window of the width N find:

a) the expected value of $S_x(n,k)$,

b) the variance of $S_x(n,k)$.

Note: For a Gaussian random signal $\varepsilon(n)$, holds

$$E\{\varepsilon(l)\varepsilon^*(m)\varepsilon^*(n)\varepsilon(p)\} = E\{\varepsilon(l)\varepsilon^*(m)\}E\{\varepsilon^*(n)\varepsilon(p)\}$$
$$+E\{\varepsilon(l)\varepsilon^*(n)\}E\{\varepsilon^*(m)\varepsilon(p)\} + E\{\varepsilon(l)\varepsilon(p)\}E\{\varepsilon^*(m)\varepsilon^*(n)\}. \quad (7.151)$$

Problem 7.32. The basic time-frequency distribution is the Fourier transform, whose discrete-time form reads

$$W_x(n,\omega) = \sum_{k=-L}^{L} x(n+k)x^*(n-k)e^{-j2\omega k},$$

where the signal is given by $x(n) = s(n) + \varepsilon(n)$, with $s(n)$ being the desired deterministic signal and $\varepsilon(n)$ being the complex-valued, zero-mean white Gaussian noise whose variance is σ_ε^2. The real and imaginary parts of the noise are independent and identically distributed (i.i.d.). Find:

a) the expected value of $W_x(n,\omega)$,

b) the variance of $W_x(n,\omega)$.

Use the previous problem note. Find the variance for an FM signal, when $|s(n)| = A$.

Problem 7.33. A random signal $s(n)$ carries an information. Its autocorrelation function is $r_{ss}(n) = 4(0.5)^{|n|}$. A noise with variance of autocorrelation $r_{\varepsilon\varepsilon}(n) = 2\delta(n)$ is added to the signal. Find the optimal filter for:

(a) $d(n) = s(n)$ - optimal filtering,

(b) $d(n) = s(n-1)$ - optimal smoothing,

(c) $d(n) = s(n+1)$ - optimal prediction.

Problem 7.34. Design an optimal filter if the autocorrelation function of the signal is $r_{ss}(n) = 3(0.9)^{|n|}$. The autocorrelation of noise is $r_{\varepsilon\varepsilon}(n) = 4\delta(n)$, while the cross–correlation of the signal and noise is $r_{s\varepsilon}(n) = 2\delta(n)$.

Problem 7.35. The power spectral densities of the signal $S_{dd}(e^{j\omega})$ and of the input noise is $S_{\varepsilon\varepsilon}(e^{j\omega})$ are given in Fig. 7.38. Show that the frequency response of the optimal filter $H(e^{j\omega})$ is presented in Fig. 7.38(bottom). Find the SNR at the input and the output of the optimal filter.

Problem 7.36. Find the expected value of the quantization error of the Fourier transform (its pseudo form over-sampled in frequency)

$$W_x(n,k) = \sum_{m=0}^{N-1} x(n+m)x(n-m)e^{-j2\pi mk/N},$$

where $x(n)$ is a real-valued quantized signal. The product of signals is quantized to b bits as well. Neglect the quantization of the coefficients $e^{-j2\pi mk/N}$ and the quantization of their products with the signal.

7.9 EXERCISE

Exercise 7.1. Signal $x_{20i}(n)$ is equal to the monthly average of the maximum daily temperatures in a city measured from year 2001 to 2015. If we can assume that the signal for an individual month is Gaussian find the probability that the average of the maximum temperatures: (a) in July is lower than 25, (b) in August is higher than 39.

Exercise 7.2. The random signal $x(n)$ is such that $x(n) = x_1(n)$ with probability p. In all other cases $x(n)$ is $x_2(n)$. If the expected value and the variance of $x_1(n)$ and $x_2(n)$ are μ_{x_1}, $\sigma_{x_1}^2$ and μ_{x_2}, $\sigma_{x_2}^2$, respectively, find the expected value and the variance of $x(n)$.

Result: $\mu_x = p\mu_{x_1} + (1-p)\mu_{x_2}$ and

$$\begin{aligned}\sigma_x^2 &= p\left[\mathrm{E}\{x_1^2(n)\} - \mu_x^2\right] + (1-p)\left[\mathrm{E}\{x_2^2(n)\} - \mu_x^2\right] \\ &= p[\sigma_{x_1}^2 + \mu_{x_1}^2 - \mu_x^2] + (1-p)[\sigma_{x_2}^2 + \mu_{x_2}^2 - \mu_x^2] \\ &= p\sigma_{x_1}^2 + (1-p)\sigma_{x_2}^2 + p(1-p)(\mu_{x_1} - \mu_{x_2})^2.\end{aligned}$$

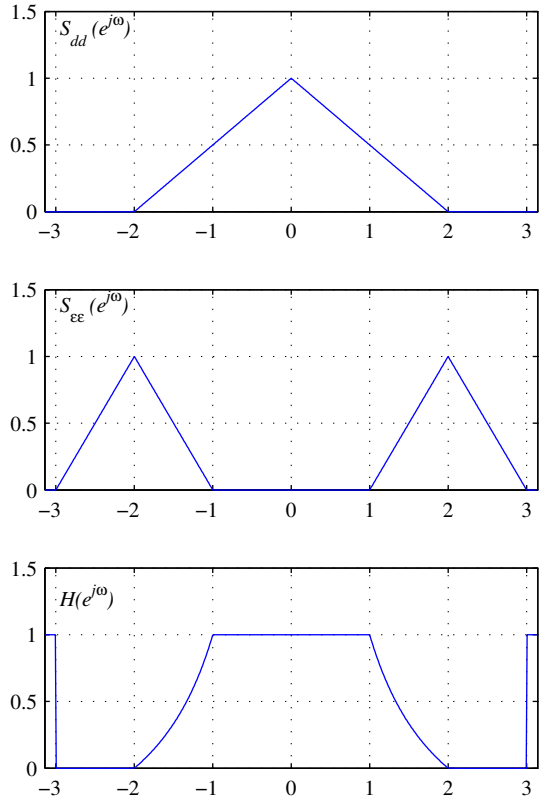


Figure 7.38 Power spectral densities of the signal $\left|S(e^{j\omega})\right|^2$ and input noise $S_{\varepsilon\varepsilon}(e^{j\omega})$ along with the frequency response of an optimal filter $H(e^{j\omega})$.

Exercise 7.3. Find the expected value and the variance of a white uniform noise whose values are within the interval $-a \leq x(n) \leq a$. If this signal is an input to the FIR system with the impulse response $h(n) = 1$ for $1 \leq n \leq N$ and $h(n) = 0$ elsewhere, find the expected value and the variance of the output signal.

Exercise 7.4. Consider the signal $x(n)$ equal to the Gaussian zero-mean noise with the variance σ_ε^2. A new noise $y(n)$ is formed using the values of $x(n)$ lower than median value. Find the expected value and the variance of this new noise $y(n)$. Result: $\sigma_y^2 = 0.1426\sigma_\varepsilon^2$.

Exercise 7.5. The causal system is defined by

$$y(n) - \frac{1}{2}y(n-1) = x(n).$$

The input signal is the causal part of a white noise $\varepsilon(n)$,

$$x(n) = \varepsilon(n)u(n)$$

where $\mu_\varepsilon = 0$ and $r_{\varepsilon\varepsilon}(n) = \sigma_\varepsilon^2\,\delta(n)$. Find the expected value value and the autocorrelation $r_{yy}(n,m)$ of the output signal. What is the cross-correlation between the input signal and the output signal $r_{yx}(n,m)$. Show that for $n \to \infty$ the output signal tends to a WSS signal.

Exercise 7.6. (a) Calculate the DFT value $X(4)$ for $x(n) = \exp(j4\pi n/N)$ with $N = 16$.

(b) Calculate the DFT of a noisy signal $x(n) + \varepsilon(n)$, where the noise realization is $\varepsilon(n) = 1001\delta(n) - 899\delta(n-3) + 561\delta(n-11) - 32\delta(n-14)$.

(c) Estimate the DFT using the noisy signal $x(n) + \varepsilon(n)$ and

$$X_R(k) = N \underset{n=0,1,..,N-1}{\text{median}} \text{Re}\left\{(x(n)+\varepsilon(n))e^{-j2\pi kn/N}\right\} + jN \underset{n=0,1,..,N-1}{\text{median}} \text{Im}\left\{(x(n)+\varepsilon(n))e^{-j2\pi kn/N}\right\}.$$

Discuss the results.

Exercise 7.7. The power spectral densities of the desired signal $S_{dd}(e^{j\omega})$ and the input noise $S_{\varepsilon\varepsilon}(e^{j\omega})$ are given in Fig. 7.39 for two cases. One on the left panels and the other on the right panels. Show that the frequency response of the optimal filter $H(e^{j\omega})$ is given in Fig. 7.39(bottom panel for both cases of the signal and noise). Find the SNR at the input and the output of the optimal filter in both cases.

Exercise 7.8. Find the transfer function of the optimal filter for the signal $x(n) = s(n) + \varepsilon(n)$, where $\varepsilon(n)$ is a white noise with the autocorrelation $r_{\varepsilon\varepsilon}(n) = N\delta(n)$, and $s(n)$ is the random signal obtained as the output of the first-order linear system to the white noise with the autocorrelation $r_{ss}(n) = a^{|n|}$, $0 < a < 1$. The signal and noise are not correlated.

Exercise 7.9. A random signal $s(n)$ carries an information. Its autocorrelation function is $r_{ss}(n) = \frac{1}{4^{|n|}}$. A noise with the autocorrelation $r_{\varepsilon\varepsilon}(n) = 0.5\delta(n)$ is added to the signal. Find the optimal filter for:

(a) $d(n) = s(n)$ - optimal filtering,

(b) $d(n) = s(n-1)$ - optimal smoothing,

(c) $d(n) = s(n+1)$ - optimal prediction.

Exercise 7.10. Find the power spectral densities of the signals whose autocorrelation functions are:

(a) $r_{xx}(n) = \delta(n) + 2\cos(0.\pi n)$,

(b) $r_{xx}(n) = -4\delta(n+1) + 7\delta(n) - 4\delta(n-1)$, and

(c) $r_{xx}(n) = 2a\cos(\omega_0 n) + \sum_{k=0}^{\infty} \sigma^2(1/2)^k\delta(n-k)$.

Exercise 7.11. Find the expected value and variance of the periodogram, $P_{xx}(e^{j\omega})$, of a deterministic signal $s(n)$ corrupted by the white noise with variance σ_ε^2,

$$P_{xx}(e^{j\omega}) = \frac{1}{N}\text{E}\{\left|\sum_{n=-N/2}^{N/2-1}(s(n)+\varepsilon(n))e^{-j\omega n}\right|^2\}. \tag{7.152}$$

Hint: For the variance calculation use

$$\text{Var}\{P_{xx}(e^{j\omega})\} = \frac{1}{N^2}\text{E}\{\sum_{n_1=-N/2}^{N/2-1}\sum_{n_2=-N/2}^{N/2-1}\sum_{n_3=-N/2}^{N/2-1}\sum_{n_4=-N/2}^{N/2-1}(s(n_1)+\varepsilon(n_1))(s(n_2)+\varepsilon(n_2)) \times (s(n_3)+\varepsilon(n_3))^*(s(n_4)+\varepsilon(n_4))^* e^{-j\omega(n_1+n_2-n_3-n_4)}\} - |\text{E}\{P_{xx}(e^{j\omega})\}|^2$$

and relation (7.151).

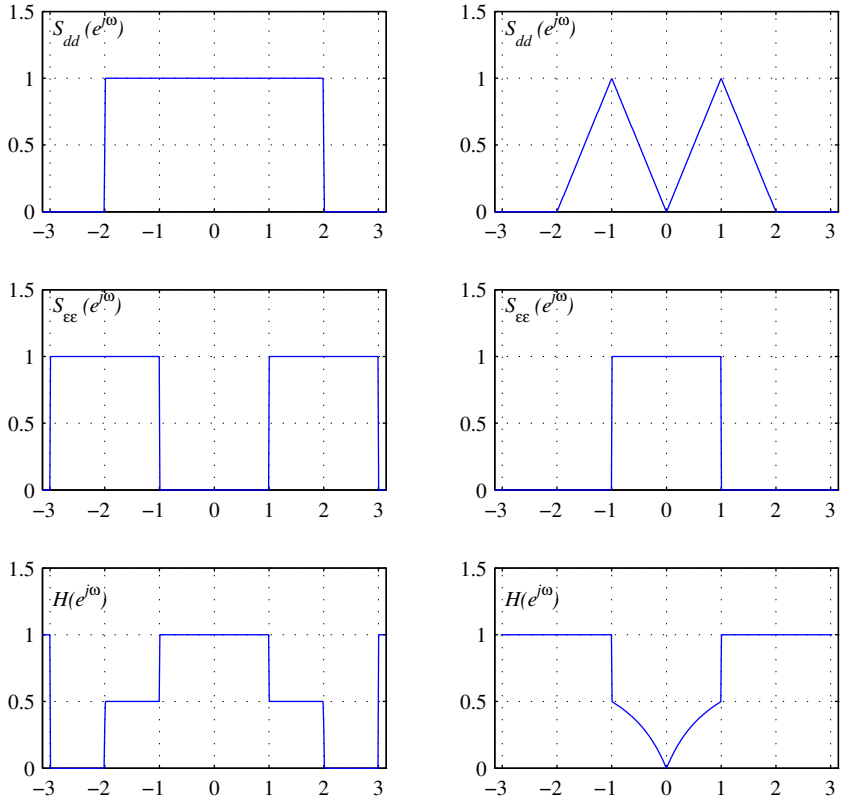


Figure 7.39 Power spectral densities of the signal $\left|S(e^{j\omega})\right|^2$ and the input noise $S_{\varepsilon\varepsilon}(e^{j\omega})$ along with the frequency responses of the optimal filters $H(e^{j\omega})$. Two cases are shown, one on the left panels and the other on the right panels.

7.10 SOLUTIONS

Solution 7.1. (a) The expected value of the temperature for January, Table 7.2, is

$$\mu_x(1) = 7.2667.$$

The standard deviation for January, calculated over 15 years, is $\sigma_x(1) = 2.6196$. The probability that the average maximum temperature in January is lower than 2 is

$$P(x(1) < 2) = \int_{-\infty}^{2} \frac{1}{\sigma_x(1)\sqrt{2\pi}} e^{-\frac{(\xi-\mu_x(1))^2}{2\sigma_x^2(1)}} d\xi = 0.5\left[1 - \text{erf}(\frac{7.2667-2}{2.7115\sqrt{2}})\right] = 0.0260.$$

This means that this event will occur once in about 40 years.

(b) The average maximum temperature is higher than 12 with the probability

$$P(x(1) > 12) = \int_{12}^{\infty} \frac{1}{\sigma_x(1)\sqrt{2\pi}} e^{-\frac{(\xi-\mu_x(1))^2}{2\sigma_x^2(1)}} d\xi = 0.5\left[1 - \text{erf}(\frac{12-7.2667}{2.7115\sqrt{2}})\right] = 0.0404.$$

This means that this event will happen once in about 25 years.

Solution 7.2. (a) In the scenario when the expected value is a priory known, $\mu_x(n) = 0$, the variance estimation is

$$\sigma_x^2(n) = \frac{1}{M}\Big(x_1^2(n) + x_2^2(n) + \cdots + x_M^2(n)\Big) - \mu_x^2(n).$$

(b) When the mean is also estimated from the data, the variance will be denoted by $s_x^2(n)$, and it is equal to

$$s_x^2(n) = \frac{1}{M}\left(\Big(x_1(n) - \frac{1}{M}\sum_{i=1}^{M} x_i(n)\Big)^2 + \cdots + \Big(x_M(n) - \frac{1}{M}\sum_{i=1}^{M} x_i(n)\Big)^2\right)$$

$$= \frac{1}{M}\sum_{j=1}^{M}\Big(x_j(n) - \frac{1}{M}\sum_{i=1}^{M} x_i(n)\Big)^2 = \frac{1}{M}\sum_{j=1}^{M} x_j^2(n) - \Big(\frac{1}{M}\sum_{i=1}^{M} x_i(n)\Big)^2$$

$$= \frac{1}{M}\sum_{j=1}^{M} x_j^2(n) - \frac{1}{M^2}\sum_{j=1}^{M}\sum_{i=1}^{M} x_i(n)x_j(n) = \frac{1}{M}\sum_{j=1}^{M} x_j^2(n) - \frac{1}{M^2}\sum_{j=1}^{M} x_j^2(n) - \frac{1}{M^2}\sum_{j=1}^{M}\sum_{\substack{i=1\\ i\neq j}}^{M} x_i(n)x_j(n)$$

$$= \frac{M-1}{M^2}\sum_{j=1}^{M} x_j^2(n) - \frac{M^2-M}{M^2}\frac{1}{M^2-M}\sum_{j=1}^{M}\sum_{\substack{i=1\\ i\neq j}}^{M} x_i(n)x_j(n).$$

In the second summation, the denominator $(M^2 - M)$ is used since there are exactly $(M^2 - M)$ terms in it, and the mean value estimate is

$$\hat{\mu}_x^2(n) \approx \frac{1}{M^2-M}\sum_{j=1}^{M}\sum_{\substack{i=1\\ i\neq j}}^{M} x_i(n)x_j(n).$$

Therefore, the the variance, $s_x^2(n)$, can be written in the form

$$s_x^2(n) = \frac{M-1}{M}\Big(\frac{1}{M}\sum_{j=1}^{M} x_j^2(n) - \frac{1}{M^2-M}\sum_{j=1}^{M}\sum_{\substack{i=1\\ i\neq j}}^{M} x_i(n)x_j(n)\Big)$$

$$= \frac{M-1}{M}\Big(\frac{1}{M}\sum_{j=1}^{M} x_j^2(n) - \hat{\mu}_x^2(n)\Big) \approx \frac{M-1}{M}\sigma_x^2(n)$$

This means that the variance with the true mean value, $\sigma_x^2(n)$, is (approximately) related to the variance with the estimated mean value, $s_x^2(n)$, as

$$\sigma_x^2(n) \approx \frac{M}{M-1}s_x^2(n) = \frac{1}{M-1}\left(\Big(x_1(n) - \frac{1}{M}\sum_{i=1}^{M} x_i(n)\Big)^2 + \cdots + \Big(x_M(n) - \frac{1}{M}\sum_{i=1}^{M} x_i(n)\Big)^2\right).$$

Solution 7.3. (a) The model parameters, obtained as the solution to the least squares minimization problem of $J(\mathbf{a}) = ||\mathbf{x} - \mathbf{T}\mathbf{a}||_2^2$, for the data

$$\mathbf{t} = [-0.8, -0.83, -0.60, -0.10, -0.01,\ 0.28,\ 0.39, 0.52,\ 0.65,\ 0.92]^T$$

and

$$\mathbf{x} = [0.26,\ 0.31, 0.64,\ 0.99, 1.00,\ 0.92,\ 0.85,\ 0.73, 0.58, 0.15]^T.$$

are given by (see the matrix $\mathbf{T}$ definition in (7.14))

$$\hat{\mathbf{a}} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{x} = [0.9997, -0.0043, -0.9926,\ 0.0275, -0.0107, -0.0293]^T. \tag{7.153}$$

The estimated model is shown in Fig. 7.40(a) by the dotted line. Since the noise is small (cased by rounding the data to two decimals), the model fits the data accurately.

(b) When the regularization constant is added, the solution to the ridge regression minimization of $J(\mathbf{a}) = ||\mathbf{x} - \mathbf{T}\mathbf{a}||_2^2 + \lambda||\mathbf{a}||_2^2)$, in the form

$$\hat{\mathbf{a}} = (\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\mathbf{x} = [0.9225,\ 0.0133, -0.5775,\ 0.0005, -0.3871, -0.0028]^T \tag{7.154}$$

is obtained. In this case, a small bias in fitting the data can be observed from Fig. 7.40(b).

(c) When a stronger additive noise is present in the data

$$\mathbf{x} = [0.35,\ 0.33,\ 0.57,\ 0.92,\ 0.94,\ 0.89,\ 0.87,\ 0.86,\ 0.44,\ 0.29]^T,$$

the least squares and the ridge regression solutions are, respectively,

$$\hat{\mathbf{a}} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{x} = [0.9515,\ 0.4443, -1.1449, -1.6335,\ 0.3664, 1.3640]^T \tag{7.155}$$

and

$$\hat{\mathbf{a}} = (\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\mathbf{x} = [0.8740,\ 0.0795, -0.5277, -0.0622, -0.2434, -0.0068]. \tag{7.156}$$

The model results are shown in Fig. 7.40(c) and (d), respectively. The higher-order model coefficients in $\hat{\mathbf{a}}$ are larger in the solution when the regularization is not used.

(d) The predicted values of $x(1.12)$ in all considered cases are indicated by a circle. We can see that the moderate noise causes significant deviation (over-fitting) of the results, Fig. 7.40(c), if the regularization is not used. In the case with a very small noise, the regularizations slightly worsen the results, by introducing the bias, as shown in Fig. 7.40(b).

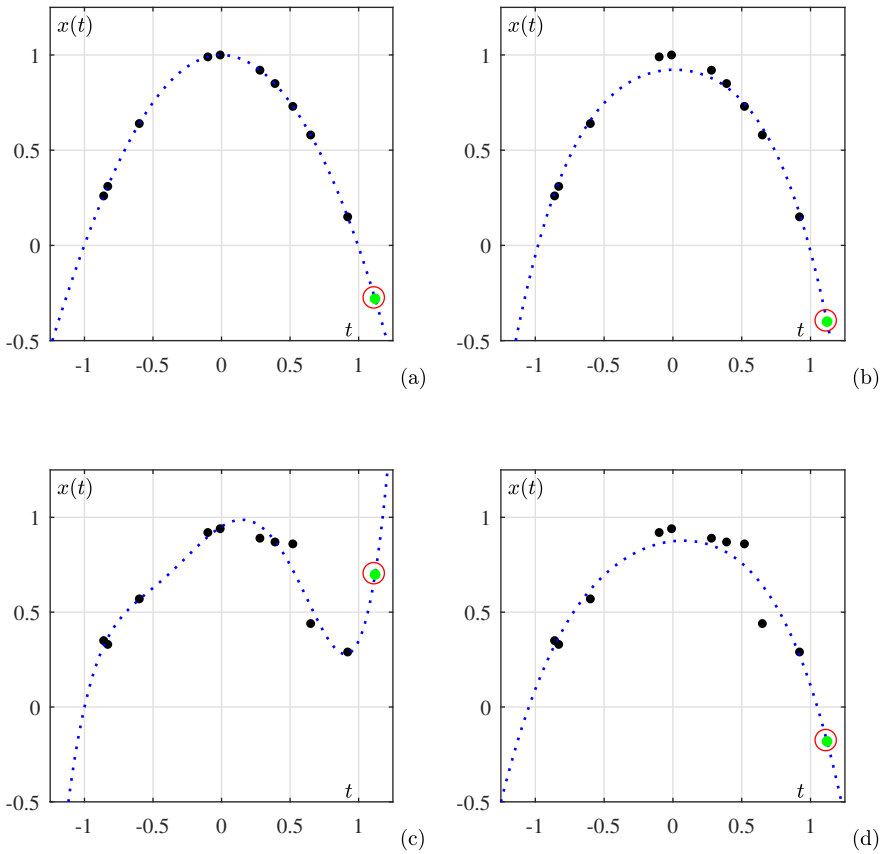


Figure 7.40 Polynomial fitting example. (a) Data with a small noise (the black dots) and the polynomial fitting with the least-squares solution (the dotted line). (b) Data with a small noise and the polynomial fitting with the regression model using regularization constant $\lambda = 0.1$, producing a small deviation from the data. (c) Data with moderate noise and the polynomial fitting with the least-squares solution. The noise causes significant deviations and an over-fitted model. (d) Data with moderate noise and the polynomial fitting with the regression model using the regularization constant $\lambda = 0.1$, keeping the energy of all model coefficients low. The predicted value at $x(1.12)$ is marked by the circle.

(e) *Advanced topic:* Assume that the observed signal consists of the true data, $\mathbf{s}$ and the noise $\boldsymbol{\varepsilon}$, that is $\mathbf{x} = \mathbf{s} + \boldsymbol{\varepsilon}$. The parameters of the true model, $\mathbf{Ta} = \mathbf{s}$, are the solution to the least-squares minimization of $J(\mathbf{a}) = ||\mathbf{s} - \mathbf{Ta}||_2^2$, that is

$$\mathbf{a} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{s}.$$

The bias of the ridge regression model estimate is obtained from

$$\hat{\mathbf{a}} = (\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\mathbf{x} = (\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T(\mathbf{s} + \boldsymbol{\varepsilon}),$$

$$\mathrm{E}\{\hat{\mathbf{a}}\} = (\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\mathbf{s}, \text{ as}$$

$$\mathrm{bias}(\hat{\mathbf{a}}) = \mathrm{E}\{\hat{\mathbf{a}}\} - \mathbf{a} = \left((\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1} - (\mathbf{T}^T\mathbf{T})^{-1}\right)\mathbf{T}^T\mathbf{s}.$$

For $\lambda = 0$ the estimator is unbiased, $\text{bias}(\hat{\mathbf{a}}) = 0$, while for large λ the bias increases toward $|\text{bias}(\hat{\mathbf{a}})| = |(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{s}|$.

The covariance matrix of the estimator is, by definition,

$$\text{Cov}(\hat{\mathbf{a}}) = \text{E}\{(\hat{\mathbf{a}} - \text{E}\{\hat{\mathbf{a}}\})(\hat{\mathbf{a}} - \text{E}\{\hat{\mathbf{a}}\})^T\} = \text{E}\{\Big((\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\boldsymbol{\varepsilon}\Big)\Big((\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\boldsymbol{\varepsilon}\Big)^T\}$$
$$= \sigma_\varepsilon^2\Big((\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\Big)\Big((\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\Big)^T = \sigma_\varepsilon^2(\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\mathbf{T}(\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}.$$

since $\text{E}\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\} = \sigma_\varepsilon^2\mathbf{I}$.

The covariance matrix for $\lambda = 0$ is $\text{Cov}(\hat{\mathbf{a}}) = \sigma_\varepsilon^2(\mathbf{T}^T\mathbf{T})^{-1}$. Its elements decrease toward 0 as λ increases.

Obviously, there is an optimal value of parameter λ, when the bias-to-variance trade off, $||\text{bias}(\hat{\mathbf{a}})||_2^2 + \text{Trace}(\text{Cov}(\hat{\mathbf{a}}))$, is minimum. $\text{Trace}(\text{Cov}(\hat{\mathbf{a}}))$ represents the sum of diagonal elements (variances) in the covariance matrix (7.24).

Cross-validation method. In practice, finding the best parameter λ value is a difficult problem. One way to find the best λ is the so-called leave-one-out *cross-validation method*, where:

1. The estimation of the model parameters is done with a preselected set of λ values, leaving out one (or more) data point $(t_n, x(n))$.
2. The estimation is repeated for other data points being leaved out, one by one.
3. The total squared error of the model predicted values with respect to the corresponding leaved out values is calculated for each of the considered λ.
4. The best λ is the one which produces the smallest total prediction error.

Solution 7.4. For the cumulative probability distribution function

$$F(\chi) = \begin{cases} 0 & \chi \le 0 \\ \chi/2 & 0 < \chi \le 1 \\ 1/2 & 1 < \chi \le 2 \\ (\chi - 1)/2 & 2 < \chi \le 3 \\ 1 & \chi > 3 \end{cases}$$

the probability density function is obtained as the derivative of $F(\chi)$,

$$p_x(\xi) = \frac{dF(\xi)}{d\xi} = \begin{cases} 0 & \xi \le 0 \\ 1/2 & 0 < \xi \le 1 \\ 0 & 1 < \xi \le 2 \\ 1/2 & 2 < \xi \le 3 \\ 0 & \xi > 3. \end{cases}$$

The probability of $x(n) < 2.5$ is $P(x(n) < 2.5) = F(2.5) = 0.75$.

Solution 7.5. Integral of the probability density function over $(-\infty, \infty)$ is

$$\int_{-\infty}^{\infty} p_x(\xi)d\xi = 1.$$

Therefore,

$$\int_{-\infty}^{\infty} ae^{-b|\xi|}d\xi = a\left[\int_{-\infty}^{0} e^{b\xi}d\xi + \int_{0}^{\infty} e^{-b\xi}d\xi\right] = \frac{2a}{b} = 1,$$

resulting in $b = 2a$.

For $a = 1$, the probability density function is $p_x(\xi) = e^{-2|\xi|}$ for $-\infty < \xi < \infty$. The probability distribution function is

$$F_x(\chi) = \int_{-\infty}^{\chi} p_x(\xi)d\xi = \int_{-\infty}^{\chi} e^{2\xi}d\xi = \frac{e^{2\chi}}{2}$$

for $-\infty < \chi < 0$, and

$$F_x(\chi) = \frac{1}{2} + \int_0^{\chi} p_\epsilon(\xi)d\xi = \frac{1}{2} + \int_0^{\chi} e^{-2\xi}d\xi = 1 - \frac{e^{-2\chi}}{2}$$

for $0 < \chi < \infty$.

Solution 7.6. The expected value is

$$\mathrm{E}_x = \int_{-\infty}^{\infty} \xi p_x(\xi)d\xi = \frac{1}{2}\int_{-\infty}^{0} \xi e^{\lambda\xi}d\xi + \frac{1}{2}\int_0^{\infty} \xi e^{-\lambda\xi}d\xi = -\frac{1}{2}\int_0^{\infty} \xi e^{-\lambda\xi}d\xi + \frac{1}{2}\int_0^{\infty} \xi e^{-\lambda\xi}d\xi = 0.$$

The variance of $x(n)$ is obtained from

$$\sigma_x^2(n) = \mathrm{E}\left\{|x(n) - E\{x(n)\}|^2\right\} = \mathrm{E}\{x^2(n)\} = \int_{-\infty}^{\infty} \xi^2 p_x(\xi)d\xi = \frac{2}{\lambda^2}.$$

Solution 7.7. Since

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{xy}(\xi,\zeta)d\xi d\zeta = 1$$

we have

$$\int_0^{\infty}\int_0^{\infty} k\xi e^{-\xi(\zeta+1)}d\xi d\zeta = \int_0^{\infty} k\xi e^{-\xi}d\xi \int_0^{\infty} e^{-\xi\zeta}d\zeta = \int_0^{\infty} k\xi e^{-\xi}\frac{1}{\xi}d\xi = \int_0^{\infty} ke^{-\xi}d\xi = k$$

and the value of constant k is 1.

Solution 7.8. Since the random signal $z(n)$ takes the greater of the values $x(n)$ and $y(n)$ the probability that $z(n) = \max\{x(n),y(n)\}$ is lower than or equal to an assumed χ is equal to the probability that both random samples $x(n)$ and $y(n)$ are lower than or equal to this assumed χ, that is

$$P\{z(n) \le \chi\} = P\{\max\{x(n),y(n)\} \le \chi\}$$
$$= P\{x(n) \le \chi \text{ and } y(n) \le \chi\} = P\{x(n) \le \chi\}P\{y(n) \le \chi\}.$$

Since

$$P\{x(n) \le \chi\} = F_{x(n)}(\chi) \quad \text{and} \quad P\{y(n) \le \chi\} = F_{y(n)}(\chi),$$

we get the probability distribution of the random variable $z(n)$ in the form

$$F_{z(n)}(\chi) = P\{z(n) \le \chi\} = F_{x(n)}(\chi)F_{y(n)}(\chi).$$

The probability density function follows as the derivative od the probability distribution,

$$p_{z(n)}(\xi) = \frac{dF_{z(n)}(\xi)}{d\xi} = p_{x(n)}(\xi)F_{y(n)}(\xi) + F_{x(n)}(\xi)p_{y(n)}(\xi).$$

Solution 7.9. There are 5 out of 10 black balls. The probability that $x(0) = 0$ is

$$P_0 = \frac{5}{10}.$$

If the first ball was 0, then we have 9 balls for the second draw, with 4 balls marked with 0. The probability that $x(1) = 0$, if $x(0) = 0$ is

$$P_1 = \frac{4}{9}.$$

If $x(0) = 0$ and $x(1) = 0$, then there are 8 remaining balls with 3 of them being marked with 0. The probability that $x(2) = 0$, with $x(0) = 0$ and $x(1) = 0$, is

$$P_2 = \frac{3}{8}.$$

The probability for $k = 0$ is

$$P(k = 0) = \frac{5}{10}\frac{4}{9}\frac{3}{8}\frac{2}{7}.$$

In general, if there were N balls, with an equal number of balls being marked with 1 (or white) and 0 (or black), and we considered M signal samples (drawings), the probability $P(k = 0)$ would be

$$P(k = 0) = \prod_{i=0}^{M-1} \frac{N/2 - i}{N - i}.$$

Solution 7.10. The variance of this random signal is defined by

$$\mathrm{Var}\{x(n)\} = \int_{-\infty}^{\infty} \xi^2 p_x(\xi) d\xi = \int_{-\infty}^{\infty} \xi^2 \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{\xi^2}{2\sigma_x^2}} d\xi.$$

Using the fact that

$$\begin{aligned}\mathrm{Var}\{x(n)\}\mathrm{Var}\{x(n)\} &= \int_{-\infty}^{\infty} \xi^2 \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{\xi^2}{2\sigma_x^2}} d\xi \int_{-\infty}^{\infty} \zeta^2 \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{\zeta^2}{2\sigma_x^2}} d\zeta \\ &= \frac{1}{\sigma_x^2 2\pi} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \xi^2\zeta^2 e^{-\frac{\xi^2+\zeta^2}{2\sigma_x^2}} d\xi d\zeta\end{aligned}$$

and the substitution of the variables to polar coordinates, $\xi = \sigma_x\rho\cos(\phi)$, $\zeta = \sigma_x\rho\sin(\phi)$, we get

$$\begin{aligned}\mathrm{Var}\{x(n)\}\mathrm{Var}\{x(n)\} &= \frac{1}{2\pi}\int_0^{\infty}\int_0^{2\pi} \rho^4\cos^2(\phi)\sin^2(\phi) e^{-\frac{\rho^2}{2}} \sigma_x^4 \rho d\rho d\phi \\ &= \frac{1}{8}\sigma_x^4 \int_0^{\infty} \rho^4 e^{-\frac{\rho^2}{2}} \rho d\rho = \frac{1}{2}\sigma_x^4 \int_0^{\infty} t^2 e^{-t} dt = \sigma_x^4,\end{aligned}$$

where $\cos^2(\phi)\sin^2(\phi) = \sin^2(2\phi)/4 = (1 - \cos(4\phi))/8$ and the substitution $\rho^2/2 = t$ are used. This means that $\mathrm{Var}\{x(n)\} = \sigma_x^2$.

Solution 7.11. For a random signal $x(n)$, with a probability density function $p_x(\xi)$, the median is defined as the value m_x such that

$$\int_{-\infty}^{m_x} p_x(\xi)d\xi = \int_{m_x}^{\infty} p_x(\xi)d\xi.$$

For the zero-mean Gaussian distributed random variable $m_x = 0$. For the random variable $|x(n)|$, the probability density function is

$$p_{|x|}(\xi) = 2p_x(\xi)u(\xi) = \frac{2}{\sigma_x\sqrt{2\pi}}e^{-\frac{\xi^2}{2\sigma_x^2}}u(\xi).$$

The median of $|x(n)|$ is obtained from

$$\frac{2}{\sigma_x\sqrt{2\pi}}\int_0^{m_x} e^{-\frac{\xi^2}{2\sigma_x^2}}d\xi = \frac{2}{\sigma_x\sqrt{2\pi}}\int_{m_x}^{\infty} e^{-\frac{\xi^2}{2\sigma_x^2}}d\xi = \frac{1}{2}$$

or

$$\frac{2}{\sigma_x\sqrt{2\pi}}\int_{m_x}^{\infty} e^{-\frac{\xi^2}{2\sigma_x^2}}d\xi = 1 - \text{erf}(\frac{m_x}{\sigma_x\sqrt{2}}) = \frac{1}{4}.$$

The solution is

$$m_x = 0.6745\sigma_x.$$

Solution 7.12. (a) When $\varepsilon(n)$ is a zero-mean Gaussian distributed random noise, with variance σ_ε^2, the variance of $y(n) = \varepsilon(n) - \varepsilon(n-1)$ is equal to

$$\sigma_y^2 = 2\sigma_\varepsilon^2.$$

Based on the result from Problem 7.11, the median of $|y(n)| = |\varepsilon(n) - \varepsilon(n-1)|$ is related to the standard deviation as

$$m_y = 0.6745\sigma_y = \sqrt{2}\ 0.6745\sigma_\varepsilon.$$

(b) For the signal $x(n)$ such that $|x(n) - x(n-1)| = |s(n) + \varepsilon(n) - s(n-1) - \varepsilon(n-1)| \approx |\varepsilon(n) - \varepsilon(n-1)|$ holds, we have

$$\underset{n=2,3,..,N}{\text{median}}\{\Big|x(n) - x(n-1)\Big|\} \approx \underset{n=2,3,..,N}{\text{median}}\{\Big|\varepsilon(n) - \varepsilon(n-1)\Big|\} \approx \sqrt{2}\ 0.6745\sigma_\varepsilon.$$

This means that the noise variance can be estimated by

$$\hat{\sigma}_\varepsilon = \frac{1}{\sqrt{2}\ 0.6745}\underset{n=2,3,..,N}{\text{median}}\{\Big|x(n) - x(n-1)\Big|\}.$$

Since the total signal energy can be calculated as

$$E_x = \sum_{n=1}^{N}|x(n)|^2 = E_s + N\sigma_\varepsilon^2,$$

the previous relation can be used to estimate the input SNR value.

(c) The true standard deviation of noise in Example 7.45 was $\sigma_\varepsilon = 4$. The value of variance, estimated using the previous relation, is

$$\hat{\sigma}_\varepsilon = \frac{1}{\sqrt{2}\ 0.6745}\underset{n=2,3,...,N}{\text{median}}\{\left|x(n)-x(n-1)\right|\} = 4.5.$$

The difference between $\hat{\sigma}_\varepsilon$ and σ_ε is due the fact that the signal variations $|s(n)-s(n-1)|$ are small, but not negligible. However, the estimation $\hat{\sigma}_\varepsilon$ is sufficiently accurate for the presented algorithm, since its slightly higher value than the true standard deviation will affect the confidence intervals only, by increasing their bounds and the corresponding probabilities (from the factor of 7.1 in Example 7.45 to the factor of 8, corresponding to the case as if the true standard deviation $\sigma_\varepsilon = 4$ were used with the confidence interval bounds defined by $2.82\sigma_{X_N}$, instead of the assumed bounds defined by $2.5\sigma_{X_N}$).

The standard deviation estimate could also be obtained using the variance definition for $x(n)-x(n-1)$, given by

$$\underset{n=2,3,...,N}{\text{mean}}\{\left|x(n)-x(n-1)\right|^2\} \approx \underset{n=2,3,...,N}{\text{mean}}\{\left|\varepsilon(n)-\varepsilon(n-1)\right|^2\} = 2\hat{\sigma}_\varepsilon^2.$$

In Example 7.45, this kind of estimation would produce $\hat{\sigma}_\varepsilon = 5.3$.

Solution 7.13. Let us find the probability that $x(n) < \xi$ for arbitrary ξ. Consider the case when $\xi < 0$,

$$P\{x(n) < \xi\} = \frac{0.2}{2}\left(1 + \text{erf}\left(\frac{\xi-3}{2}\right)\right).$$

It has been taken into account that the considered sample is Gaussian (with probability 0.2), along with the probability that the sample value is smaller than ξ.

For $\xi > 0$, we should take into account that the signal takes $x(n) = 0$ with the probability 80% as well as that in the remaining 20% cases, Gaussian random value could be smaller than ξ. So, we get

$$P\{x(n) < \xi\} = 0.8 + \frac{0.2}{2}\left(1 + \text{erf}\left(\frac{\xi-3}{2}\right)\right).$$

Now, we have

$$P\{x(n) < \xi\} = \begin{cases} \frac{0.2}{2}\left(1 + \text{erf}\left(\frac{\xi-3}{2}\right)\right) & \text{for} \xi < 0 \\ 0.8 + \frac{0.2}{2}\left(1 + \text{erf}\left(\frac{\xi-3}{2}\right)\right) & \text{for } \xi > 0. \end{cases}$$

This function has a discontinuity at $\xi = 0$. It is not differentiable at this point as well. The derivative of $P\{x(n) < \xi\}$ can be expressed in the form of the generalized functions (Dirac delta function) as

$$\frac{d}{d\xi}P\{x(n) < \xi\} = p_{y(n)}(\xi) = \frac{0.2}{2\sqrt{\pi}}e^{-\frac{(\xi-3)^2}{4}} + 0.8\delta(\xi).$$

The expected value and the variance are

$$\mu_{y(n)} = \int_{-\infty}^{\infty} \xi p_{y(n)}(\xi)d\xi = 0.2 \times 3 + 0.8 \times 0 = 0.6$$

$$\sigma^2_{y(n)} = \int_{-\infty}^{\infty} (\xi - 0.6)^2 p_{y(n)}(\xi)d\xi = 0.2 \times 7.76 + 0.8 \times (0.6)^2 = 1.84.$$

Solution 7.14. The probability that $|\varepsilon(n)| > A$ is

$$P\{|\varepsilon(n)| > A\} = P\{\varepsilon(n) < -A\} + P\{\varepsilon(n) > A\}$$
$$= \int_{-\infty}^{-A} \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} e^{-\frac{\zeta^2}{2\sigma_\varepsilon^2}} d\zeta + \int_{A}^{\infty} \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} e^{-\frac{\zeta^2}{2\sigma_\varepsilon^2}} d\zeta = 1 - \text{erf}\left(\frac{A}{\sqrt{2}\sigma_\varepsilon}\right).$$

For $A = 10$ and $\sigma_\varepsilon^2 = 2$, we get

$$P\{|\varepsilon(n)| > 10\} = 1 - \text{erf}(5) \approx 1.5 \times 10^{-12}.$$

For $N = 2000$, the expected number of samples with amplitude above A is $P\{|\varepsilon(n)| > 10\} \times 2000 \approx 3 \times 10^{-9} \approx 0$. This means that we do not expect any sample with amplitude higher than 10.

For $A = 4$, we have

$$P\{|\varepsilon(n)| > A\} = 1 - \text{erf}\,(2) \approx 4.7 \times 10^{-3}$$

with $2000 \times 4.7 \times 10^{-3} = 9.4 \approx 9$ samples among the considered 2000 assuming an amplitude higher than 4.

Solution 7.15. If we are in the position to use a reduced set of the signal samples for processing, then the ideal scenario would be to eliminate signal samples with the higher noise values and to keep for processing the samples with the lower noise values. For the case of N, the signal samples and signal processing based on M samples, we can find the interval of amplitudes A for the lowest M noisy samples. The probability that $|x(n)| < A\sigma_\varepsilon$ is equal to

$$P\{|x(n)| < A\sigma_\varepsilon\} = \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} \int_{-A\sigma_\varepsilon}^{A\sigma_\varepsilon} e^{-\xi^2/(2\sigma_\varepsilon^2)} d\xi.$$

Since we use M out of N samples, this probability should be equal to M/N,

$$\frac{1}{\sqrt{2\pi}} \int_{-A}^{A} e^{-\xi^2/2} d\xi = \text{erf}\left(\frac{A}{\sqrt{2}}\right) = \frac{M}{N}.$$

The calculation of A value is easily related to the inverse $\text{erf}(x)$ function, denoted by $\text{erfinv}(x)$. For a given M/N, the amplitude is $A = \sqrt{2}\text{erfinv}(\frac{M}{N})$. For example, for $M = N/2$ a half of the lowest noise samples will be within the interval $[-0.6745\sigma_\varepsilon, 0.6745\sigma_\varepsilon]$, since $A = \sqrt{2}\text{erfinv}(0.5) = 0.6745$.

The probability density function of the new noise is

$$p_y(\xi) = \begin{cases} \frac{k}{\sigma_\varepsilon\sqrt{2\pi}} e^{-\xi^2/(2\sigma_\varepsilon^2)} & \text{for} \quad |\xi| < A\sigma_\varepsilon \\ 0 & \text{for} \quad |\xi| \geq A\sigma_\varepsilon. \end{cases}$$

The constant k is obtained from the condition that $\int_{-\infty}^{\infty} p_y(\xi) d\xi = 1$. Its value is $k = N/M$.

The variance of this new noise, formed from the Gaussian noise after the largest $N - M$ values are removed, is much lower than the variance of the whole noise. It is given by

$$\sigma_y^2 = \frac{\frac{N}{M}}{\sigma_\varepsilon\sqrt{2\pi}} \int_{-\sqrt{2}\text{erfinv}(\frac{M}{N})\sigma_\varepsilon}^{\sqrt{2}\text{erfinv}(\frac{M}{N})\sigma_\varepsilon} \xi^2 e^{-\xi^2/(2\sigma_\varepsilon^2)} d\xi. \tag{7.157}$$

Solution 7.16. The probability density function for any sample $x(n)$, $n \neq n_0$, is

$$p_{x(n),n\neq n_0}(\xi) = \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} e^{-\xi^2/(2\sigma_\varepsilon^2)}.$$

The probability that any of these samples is smaller than a value of λ could be defined using (7.43)

$$\begin{aligned} P^-(\lambda) &= \text{Probability}\{x(n) < \lambda, n \neq n_0\} \\ &\text{Probability}\{x(n) < 0, n \neq n_0\} + \text{Probability}\{0 \leq x(n) < \lambda, n \neq n_0\} \\ &= 0.5 + 0.5\,\text{erf}(\lambda/(\sqrt{2}\sigma_\varepsilon)). \end{aligned}$$

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 λ is

$$\begin{aligned} P^-_{N-1}(\lambda) &= \text{Probability}\{\text{All } N-1 \text{ values of } x(n) < \lambda, n \neq n_0\} \\ &= \left[0.5 + 0.5\,\text{erf}(\lambda/(\sqrt{2}\sigma_\varepsilon))\right]^{N-1}. \end{aligned}$$

The probability density function of the sample $x(n_0)$ is a Gaussian function with the mean value A, that is

$$p_{x(n_0)}(\xi) = \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} e^{-(\xi-A)^2/(2\sigma_\varepsilon^2)}.$$

The probability that the random variable $x(n_0)$ takes a value around λ, $\lambda \leq x(n_0) < \lambda + d\lambda$, is

$$P^+_{n_0}(\lambda) = \text{Probability}\{\lambda \leq x(n_0) < \lambda + d\lambda\} = \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} e^{-(\xi-A)^2/(2\sigma_\varepsilon^2)} d\lambda \tag{7.158}$$

The probability that all values of $x(n), 0 \leq n \leq N-1, n \neq n_0$ are smaller than λ and that, at the same time, $\lambda \leq x(n_0) < \lambda + d\lambda$ is

$$P_A(\lambda) = P^-_{N-1}(\lambda)P^+_{n_0}(\lambda) = \left[0.5 + 0.5\,\text{erf}\left(\frac{\lambda}{\sqrt{2}\sigma_\varepsilon}\right)\right]^{N-1} \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} e^{-(\xi-A)^2/(2\sigma_\varepsilon^2)} d\lambda,$$

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 λ

$$P_A = \int_{-\infty}^{\infty} P_A(\lambda)d\lambda = \int_{-\infty}^{\infty} \left[0.5 + 0.5\,\text{erf}\left(\frac{\lambda}{\sqrt{2}\sigma_\varepsilon}\right)\right]^{N-1} \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} e^{-(\xi-A)^2/(2\sigma_\varepsilon^2)} d\lambda. \tag{7.159}$$

This integral cannot be simplified and should be evaluated numerically.

Solution 7.17. The probability density function for the sequence $y(n)$ is

$$p_{y(n)}(\zeta) = \begin{cases} B\frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(\zeta)^2}{2\sigma_x^2}} & \text{for } -A < \zeta \leq A \\ 0 & \text{otherwise.} \end{cases}$$

The constant B can be calculated from $\int_{-\infty}^{\infty} p_{y(n)}(\zeta)d\zeta = 1$. Its value is $B = 1/\operatorname{erf}\left(\frac{A}{\sigma_x\sqrt{2}}\right)$. Now, we have $\mu_{y(n)} = 0$ and

$$\sigma^2_{y(n)} = \int_{-A}^{A} \zeta^2 \frac{1}{\operatorname{erf}\left(\frac{A}{\sigma_x\sqrt{2}}\right)} \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(\zeta)^2}{2\sigma_x^2}} d\zeta = \sigma_x^2 \left(1 - \frac{A}{\sigma_x} \frac{\sqrt{2}e^{-\frac{A^2}{2\sigma_x^2}}}{\sqrt{\pi}\operatorname{erf}\left(\frac{A}{\sigma_x\sqrt{2}}\right)}\right).$$

By denoting $\beta = A/(\sqrt{2}\sigma_x)$, the variance $\sigma^2_{y(n)}$ can be written as

$$\sigma^2_{y(n)} = \sigma_x^2 \left(1 - 2\beta \frac{e^{-\beta^2}}{\sqrt{\pi}\operatorname{erf}(\beta)}\right).$$

Solution 7.18. False detection means that we make a wrong decision by classifying instant n into set $\mathbb{N}_x$. The probability is

$$P_F = P\{\varepsilon(n) > T\} = \frac{1}{2} - \frac{1}{2}\operatorname{erf}\left(\frac{T}{\sqrt{2}\sigma_\varepsilon}\right)$$

Now, we can find T as

$$T = \sqrt{2}\sigma_\varepsilon \operatorname{erfinv}(1 - 2P_F) \approx 2.33\sigma_\varepsilon$$

where $\operatorname{erfinv}(\cdot)$ is the inverse erf function. Note that the threshold does not depend on A.

Solution 7.19. The joint probability distribution is

$$p_{x(n),y(n)}(\xi,\zeta) = p_{x(n)}(\xi)p_{y(n)}(\zeta),$$

since signals are mutually independent. The probability that $x(n) > y(n)$ can be obtained by integrating $p_{x(n),y(n)}(\xi,\zeta)$ over the region $\xi > \zeta$,

$$P\{x(n) > y(n)\} = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\xi-5)^2}{2}} \int_{-\infty}^{\xi} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\zeta-1)^2}{2}} d\zeta d\xi \approx 0.99766.$$

For 1000 instants, we expect that $x(n) > y(n)$ is satisfied in about 998 instants.

Solution 7.20. Since the variable

$$z = \frac{1}{M}\sum_{n=1}^{M} x(n)y(n)$$

is also of zero-mean, then its variance is

$$\sigma_z^2 = \mathrm{E}[z^2] = \mathrm{E}\left[\frac{1}{M}\sum_{n=1}^{M} x(n)y(n)\frac{1}{M}\sum_{m=1}^{M} x(m)y(m)\right]$$

$$= \frac{1}{M^2}\sum_{n=1}^{M}\sum_{m=1}^{M} \mathrm{E}\left[x(n)y(n)x(m)y(m)\right] = \frac{1}{M^2}\sum_{n=1}^{M}\sum_{m=1}^{M} \mathrm{E}[x(n)x(m)]\,\mathrm{E}[y(n)y(m)]$$

$$= \frac{1}{M^2}\sum_{n=1}^{M} \mathrm{E}[x^2(n)]\,\mathrm{E}[y^2(n)] = \frac{1}{M^2}\sum_{n=1}^{M} \sigma_x^2\sigma_y^2 = \frac{1}{M}\sigma_x^2\sigma_y^2.$$

Solution 7.21. The moments of the Gaussian distributed random variable follow from the moment generating function, related to the Fourier transform of the Gaussian distribution (characteristics function), as

$$M_x(\theta) = \Phi_x(-j\theta) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(\xi-\mu)^2/(2\sigma^2)} e^{j(-j\theta)\xi} d\xi = e^{-\sigma^2(-j\theta)^2/2} e^{j(-j\theta)\mu} = e^{\sigma^2\theta^2/2} e^{\theta\mu}.$$

Expanding the moment generating function $M_x(\theta)$ into Taylor's series around $\theta = 0$ we get

$$\begin{aligned} e^{\theta\mu} e^{\sigma^2\theta^2/2} &= (1 + (\theta\mu) + \frac{1}{2!}(\theta\mu)^2 + \frac{1}{3!}(\theta\mu)^3 + \frac{1}{4!}(\theta\mu)^4 + \dots) \\ &\times (1 + (\sigma^2\theta^2/2) + \frac{1}{2!}(\sigma^2\theta^2/2)^2 + \frac{1}{3!}(\sigma^2\theta^2/2)^3 + \frac{1}{4!}(\sigma^2\theta^2/2)^4 + \dots) \\ &= 1 + \theta\mu + \frac{1}{2!}\theta^2(\mu^2 + \sigma^2) + \frac{1}{3!}\theta^3(\mu^3 + 3\mu\sigma^2) + \frac{1}{4!}\theta^4(\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4)\dots \end{aligned}$$

The moments of the Gaussian distributed random variable are given by (on the right for $\mu = 0$)

$$\begin{aligned} M_1 &= \mu, & M_1 &= 0, \\ M_2 &= \mu^2 + \sigma^2, & M_2 &= \sigma^2, \\ M_3 &= \mu^3 + 3\mu\sigma^2, & M_3 &= 0, \\ M_4 &= \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4, & M_4 &= 3\sigma^4. \end{aligned}$$

The cumulants K_i of the random variable $x(n)$ are obtained, by definition, from the Taylor's series around $\theta = 0$ of the logarithm of the moment generating function $\ln(M_x\theta))$. In the case of the Gaussian distribute variable

$$\ln(M_x(\theta)) = \theta\mu + \sigma^2\theta^2/2.$$

Obviously,

$$\begin{aligned} K_1 &= \mu, & K_1 &= 0, \\ K_2 &= \sigma^2, & K_2 &= \sigma^2, \\ K_3 &= 0, & K_3 &= 0, \\ K_4 &= 0, & K_4 &= 0, \end{aligned}$$

and $K_i = 0$ for $i > 2$. This is well-known criterion to test if a random variable is Gaussian distributed, since the cumulants of this distribution should be zero for $i > 2$. Since the third-oder moments are zero-valued for any even distribution function, the fourth-order cumulant is used to check if a random variable is Gaussian distributed. For an even distributed random variable, the fourth-order cumulant is related to the moments as $K_4 = M_4 - 3M_2^2$, and M_i is statistically estimated as $M_i = \text{mean}(x^i(n)))$.

The *kurtosis* is defined as the fourth-order moment of the centered and normalized random variable

$$\text{Kurt}_x = \text{E}\{\Big(\frac{x(n) - \mu_x}{\sigma_x}\Big)^4\}.$$

For the Gaussian random variable $\text{Kurt}_x = 3$. Any different value than $\text{Kurt}_x = 3$ produces the *excess kurtosis*,

$$\text{ExcessKurt}_x = \text{Kurt}_x - 3,$$

whose value is different than zero, and this value is an indicator of the distribution deviation from the Gaussian distribution.

Solution 7.22. The probability that the random variable is within $-\infty < \xi < \infty$ is equal to

$$1 = \int_{-\infty}^{\infty} p_{\varepsilon(n)}(\xi)d\xi = \int_{-\infty}^{\infty} \frac{a}{1+\xi^2}d\xi = a\arctan(\xi)|_{-\infty}^{\infty} = a\pi,$$

resulting in $a = 1/\pi$. The expected value is

$$\mu_\varepsilon = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\xi}{1+\xi^2}d\xi = 0,$$

while the variance

$$\sigma_\varepsilon^2 = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\xi^2}{1+\xi^2}d\xi \to \infty$$

does not exist. This noise belongs to the class of impulsive, heavy tailed, noises.

Solution 7.23. The expected value of the Poisson distributed random variable is

$$\mu_x = \sum_{k=0}^{\infty} kP(k) = \sum_{k=0}^{\infty} k\frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=1}^{\infty} k\frac{\lambda^k e^{-\lambda}}{k!} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}$$
$$= \lambda e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = \lambda e^{-\lambda}e^{\lambda} = \lambda.$$

The variance of this random variable is given by

$$\sigma_x^2 = \sum_{k=0}^{\infty}(k-\mu_x)^2P(k) = \sum_{k=0}^{\infty}(k-\lambda)^2\frac{\lambda^k e^{-\lambda}}{k!} = \Big(\sum_{k=0}^{\infty}k^2\frac{\lambda^k e^{-\lambda}}{k!}\Big) - \lambda^2$$
$$= \Big(\sum_{k=0}^{\infty}(k(k-1)+k)\frac{\lambda^k e^{-\lambda}}{k!}\Big) - \lambda^2 = \Big(\sum_{k=2}^{\infty}(k(k-1)\frac{\lambda^k e^{-\lambda}}{k!} + \sum_{k=1}^{\infty}k\frac{\lambda^k e^{-\lambda}}{k!})\Big) - \lambda^2$$
$$= \Big(e^{-\lambda}\lambda^2\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\Big) + \lambda - \lambda^2 = e^{-\lambda}\lambda^2 e^{\lambda} + \lambda - \lambda^2 = \lambda.$$

Notice that the variance is the mean-value dependent, $\sigma_x^2 = \mu_x = \lambda$. In the confidence interval calculation, this problem can be solved using the variance stabilization (see Section 7.4.7 and Fig. 7.27). The transformation that would produce a mean value independent estimate of the variance would be $g(\lambda) = \sqrt{\lambda}$.

Solution 7.24. The transfer function of the causal system is

$$H(z) = \frac{1}{1-0.5z^{-1}}.$$

The z-transform of the input signal $x(n)$ is

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n} = \sum_{n=-\infty}^{\infty} a\delta(n)z^{-n} = a.$$

The z-transform of the output signal is given by

$$Y(z) = H(z)X(z) = \frac{a}{1-0.5z^{-1}},\ |z| > 1/2.$$

Using the power series expansion of $Y(z)$ we can write

$$Y(z) = a\sum_{n=0}^{\infty}(1/2)^n z^{-n}.$$

The output signal is

$$y(n) = a\cdot 2^{-n}u(n).$$

It has been assumed that the random variable a is uniform within $[4,5]$. Its probability density function is defined by

$$p_a(\xi) = \begin{cases} 1, & \xi \in [4,5] \\ \\ 0, & \text{elsewhere.} \end{cases}$$

The expected value and the autocorrelation of the output signal $y(n)$ are

$$\mu_y(n) = E\{y(n)\} = \int_{-\infty}^{\infty} y(n)p(a)da = 9\cdot 2^{-(n+1)}u(n)$$

$$r_{yy}(n,m) = E\{y(n)y^*(m)\} = \frac{61}{3}2^{-(n+m)}u(n)u(m).$$

The output signal $y(n)$ is not WSS.

Solution 7.25. (a) The autocorrelation function of the input signal is

$$r_{xx}(n) = r_{\varepsilon\varepsilon}(n) = \delta(n).$$

Its z-transform and the power spectral density are

$$R_{xx}(z) = \sum_{n=-\infty}^{\infty} r_{xx}(n)z^{-n} = 1$$
$$S_{xx}(\omega) = 1.$$

The power spectral density of the output signal is

$$S_{yy}(\omega) = R_{yy}(e^{j\omega}) = S_{xx}(\omega)\left|H(e^{j\omega})\right|^2 = 1, \text{ for } \omega \neq 0.$$

The inverse Fourier transform produces the autocorrelation function

$$r_{yy}(n) = r_{\varepsilon_h\varepsilon_h}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{yy}(\omega)e^{j\omega n}d\omega = \delta(n).$$

(b) The z-transform of the cross-correlation of the input and the output signal $y(n) = \varepsilon(n) *_n h(n) = \varepsilon_h(n)$, is

$$R_{xy}(z) = R_{xx}(z)H(z).$$

For $z = e^{j\omega}$, we get

$$R_{\varepsilon\varepsilon_h}(e^{j\omega}) = S_{\varepsilon\varepsilon}(\omega)H(e^{j\omega}) = H(e^{j\omega}),$$

resulting in

$$r_{\varepsilon\varepsilon_h}(n) = h(n) = \begin{cases} \frac{2}{\pi}\frac{\sin^2(n\pi/2)}{n}, & n \neq 0 \\ 0, & n = 0. \end{cases}$$

It is easy to conclude that the cross-correlation function is antisymmetric $r_{xy}(-n) = -r_{xy}(n)$.

(c) The analytic part of the signal $x(n) = \varepsilon(n)$ is

$$x_a(n) = \varepsilon_a(n) = x(n) + jx_h(n) = x(n) + j\sum_{k=-\infty}^{\infty} h(k)x(n-k).$$

The Fourier transform of both sides produces

$$X_a(e^{j\omega}) = X(e^{j\omega}) + jH(e^{j\omega})X(e^{j\omega}).$$

If we divide both sides of the previous equation by $X(e^{j\omega})$ we get

$$\frac{X_a(e^{j\omega})}{X(e^{j\omega})} = H_a(e^{j\omega}) = 1 + jH(e^{j\omega}) = 1 + \text{sgn}(\omega) = \begin{cases} 2, & \omega > 0 \\ 1, & \omega = 0 \\ 0, & \omega < 0. \end{cases}$$

The power spectral density of the output signal is

$$S_{\varepsilon_a\varepsilon_a}(\omega) = \left|H_a(e^{j\omega})\right|^2 S_{\varepsilon\varepsilon}(\omega) = \left|H_a(e^{j\omega})\right|^2 = \begin{cases} 4, & \omega > 0 \\ 1, & \omega = 0 \\ 0, & \omega < 0 \end{cases}$$

with the autocorrelation function of $\varepsilon_a(n)$,

$$r_{\varepsilon_a\varepsilon_a}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{\varepsilon_a\varepsilon_a}(\omega)e^{-j\omega n}d\omega = \begin{cases} -\frac{4}{jn\pi}, & \text{for odd } n \\ 0, & \text{for even } n. \end{cases}$$

Solution 7.26. The power spectral density of the input signal is

$$S_{xx}(\omega) = \sum_{n=-\infty}^{\infty} r_{xx}(n)e^{-j\omega n} = \sum_{n=-\infty}^{\infty} r_{\varepsilon\varepsilon}(n)e^{-j\omega n} = \sigma_\varepsilon^2.$$

The transfer function is

$$H(z) = \frac{1}{1 - az^{-1}} = \frac{z}{z - a}$$

with the impulse response $h(n) = a^n u(n)$. The impulse response is real-valued. The z-transform of the output signal autocorrelation function is

$$R_{yy}(z) = H(z)H(1/z)R_{xx}(z) = \frac{z}{(z-a)(1-az)}\sigma_\varepsilon^2.$$

The inverse z-transform results in the autocorrelation function of $y(n)$

$$r_{yy}(n) = \frac{a^{|n|}}{1-a^2}\sigma_\varepsilon^2.$$

The power spectral density of the output signal is

$$S_{yy}(\omega) = R_{yy}(e^{j\omega}) = \frac{\sigma_\varepsilon^2}{(1-ae^{-j\omega})(1-ae^{j\omega})} = \frac{\sigma_\varepsilon^2}{1-2a\cos\omega+a^2}.$$

Solution 7.27. The expected value of $y(n)$ is

$$\mu_y(n) = \mathrm{E}\left\{\sum_{k=-\infty}^{\infty} h(k)x(n-k)\right\} = \sum_{k=0}^{\infty} a^k \mathrm{E}\{\varepsilon(n-k)\}u(n-k) = \sum_{k=0}^{n} a^k \mu_\varepsilon = \mu_\varepsilon \frac{1-a^{n+1}}{1-a}u(n).$$

The variance is

$$\begin{aligned}\sigma_y^2(n) &= \mathrm{E}\left\{\left(y(n)-\mu_y(n)\right)^2\right\} = \mathrm{E}\{y^2(n)\} - \mu_y^2(n) \\ &= \sum_{k_1=0}^{n}\sum_{k_2=0}^{n} a^{k_1}a^{k_2}\mathrm{E}\{\varepsilon(n-k_1)\varepsilon(n-k_2)\}u(n) - \left(\mu_\varepsilon \frac{1-a^{n+1}}{1-a}\right)^2 u(n).\end{aligned}$$

Since $\mathrm{E}\{\varepsilon(n-k_1)\varepsilon(n-k_2)\} = \sigma_\varepsilon^2\delta(k_1-k_2) + \mu_\varepsilon^2$, we get

$$\sigma_y^2(n) = \sigma_\varepsilon^2 \frac{1-a^{2(n+1)}}{1-a^2}u(n).$$

Solution 7.28. The expected value is

$$\mu_x = \mu_\varepsilon + \sum_{k=1}^{N} a_k \mathrm{E}\{e^{j(\omega_k n+\theta_k)}\} = \mu_\varepsilon,$$

since

$$\mathrm{E}\{e^{j(\omega_k n+\theta_k)}\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{j(\omega_k n+\theta_k)}d\theta_k = 0.$$

The autocorrelation is

$$r_{xx}(n) = \sigma_\varepsilon^2\delta(n) + \mu_\varepsilon^2 + \sum_{k=1}^{N} a_k^2 e^{j\omega_k n},$$

while the power spectral density for $-\pi < \omega \le \pi$ is

$$S_{xx}(e^{j\omega}) = \mathrm{FT}\{r_{xx}(n)\} = \sigma_\varepsilon^2 + 2\pi\mu_\varepsilon^2\delta(\omega) + 2\pi\sum_{k=1}^{N} a_k^2\delta(\omega-\omega_k).$$

Solution 7.29. For the optimal filtering $d(n) = s(n)$. The cross-correlation of the input signal and the desired signal is

$$r_{dx}(n) = \mathrm{E}\{d(k)x(k-n)\} = \mathrm{E}\{s(k)[s^*(k-n)+\varepsilon^*(k-n)]\} = r_{ss}(n) = 0.25^{|n|}.$$

Its z-transform is

$$R_{dx}(z) = R_{ss}(z) = \frac{-15z/4}{(z-1/4)(z-4)}.$$

The input signal autocorrelation is

$$r_{xx}(n) = r_{ss}(n) + r_{\varepsilon\varepsilon}(n) = 0.25^{|n|} + \delta(n),$$

with the z-transform

$$R_{xx}(z) = \frac{-15z/4}{(z-1/4)(z-4)} + 1 = \frac{z^2 - 8z + 1}{(z-1/4)(z-4)}.$$

The optimal filter transfer function is

$$H(z) = \frac{R_{dx}(z)}{R_{xx}(z)} = \frac{R_{ss}(z)}{R_{xx}(z)} = \frac{-15z/4}{z^2 - 8z + 1}.$$

A stable system requires the region of convergence $0.127 < |z| < 7.873$. This region of convergence does not correspond to a causal system.

Solution 7.30. By direct calculation, in the noisy case, we obtain

$$X(2) = 2009 + j204$$

and

$$X_R(k) = 8.$$

Note that the noise-free DFT value $X(2)$ is 8.

Solution 7.31. With the rectangular window, the spectrogram form is given by

$$S_x(n,k) = \left| \sum_{i=0}^{N-1} x(n+i) e^{-j\frac{2\pi}{N}ik} \right|^2 = \sum_{i_1=0}^{N-1} \sum_{i_2=0}^{N-1} x(n+i_1) x^*(n+i_2) e^{-j\frac{2\pi}{N}(i_1-i_2)k}.$$

a) The expected value of the spectrogram is

$$\mathrm{E}\{S_x(n,k)\} = \sum_{i_1=0}^{N-1} \sum_{i_2=0}^{N-1} \mathrm{E}\{x(n+i_1) x^*(n+i_2)\} e^{-j\frac{2\pi}{N}(i_1-i_2)k}.$$

Using the fact that the signal $s(n)$ is deterministic and the noise $\varepsilon(n)$ is zero-mean, we get

$$\begin{aligned}\mathrm{E}\{S_x(n,k)\} &= \sum_{i_1=0}^{N-1} \sum_{i_2=0}^{N-1} s(n+i_1) s^*(n+i_2) e^{-j\frac{2\pi}{N}(i_1-i_2)k} \\ &\quad + \sum_{i_1=0}^{N-1} \sum_{i_2=0}^{N-1} \mathrm{E}\{\varepsilon(n+i_1)\varepsilon^*(n+i_2)\} e^{-j\frac{2\pi}{N}(i_1-i_2)k}\end{aligned}$$

or

$$\begin{aligned}\mathrm{E}\{S_x(n,k)\} &= S_s(n,k) + \sigma_\varepsilon^2 \sum_{i_1=0}^{N-1} \sum_{i_2=0}^{N-1} \delta(i_1 - i_2) e^{-j\frac{2\pi}{N}(i_1-i_2)k} \\ &= S_s(n,k) + \sigma_\varepsilon^2 \sum_{i=0}^{N-1} 1 = S_s(n,k) + N\sigma_\varepsilon^2\end{aligned}$$

since for the noise holds

$$r_{\varepsilon\varepsilon}(i) = \mathrm{E}\{\varepsilon(n+i)\varepsilon^*(i)\} = \sigma_\varepsilon^2 \delta(i).$$

b) The variance of $S_x(n,k)$ is

$$\sigma^2 = \mathrm{E}\{S_x(n,k)S_x^*(n,k)\} - \mathrm{E}\{S_x(n,k)\}\mathrm{E}\{S_x^*(n,k)\}.$$

The first term can be written as

$$\mathrm{E}\{S_x(n,k)S_x^*(n,k)\} = \sum_{i_1=0}^{N-1}\sum_{i_2=0}^{N-1}\sum_{i_3=0}^{N-1}\sum_{i_4=0}^{N-1}\mathrm{E}\{x(n+i_1)x^*(n+i_2) \times x^*(n+i_3)x(n+i_4)\}e^{-j\frac{2\pi}{N}(i_1-i_2-i_3+i_4)k}$$

where

$$\begin{aligned}&\mathrm{E}\{x(n+i_1)x^*(n+i_2)x^*(n+i_3)x(n+i_4)\}\\&= s(n+i_1)s^*(n+i_2)s^*(n+i_3)s(n+i_4)\\&+ s(n+i_1)s^*(n+i_2)r_{\varepsilon\varepsilon}(i_4-i_3) + s(n+i_1)s^*(n+i_3)r_{\varepsilon\varepsilon}(i_4-i_2)\\&+ s^*(n+i_2)s(n+i_4)r_{\varepsilon\varepsilon}(i_1-i_3) + s^*(n+i_3)s(n+i_4)r_{\varepsilon\varepsilon}(i_1-i_2)\\&+ \mathrm{E}\{\varepsilon(n+i_1)\varepsilon^*(n+i_2)\varepsilon^*(n+i_3)\varepsilon(n+i_4)\}.\end{aligned}$$

The facts that odd order moments of a Gaussian zero-mean noise are zero and $r_{\varepsilon\varepsilon^*}(k) = r_{\varepsilon^*\varepsilon}(k) = 0$ for a complex-valued noise with i.i.d. are used. According to relation (7.151) from the note, holds

$$\mathrm{E}\{\varepsilon(n+i_1)\varepsilon^*(n+i_2)\varepsilon^*(n+i_3)\varepsilon(n+i_4)\} = r_{\varepsilon\varepsilon}(i_1-i_2)r_{\varepsilon\varepsilon}(i_4-i_3) + r_{\varepsilon\varepsilon}(i_1-i_3)r_{\varepsilon\varepsilon}(i_4-i_2). \tag{7.160}$$

After few straightforward transformations, we get

$$\begin{aligned}\mathrm{E}\{S_x(n,k)S_x^*(n,k)\} = S_s^2(n,k) + \sigma_\varepsilon^2\sum_{i_1=0}^{N-1}\sum_{i_2=0}^{N-1}\sum_{i_3=0}^{N-1}\sum_{i_4=0}^{N-1}[s(n+i_1)s^*(n+i_2)\delta(i_4-i_3)\\+s(n+i_1)s^*(n+i_3)\delta(i_4-i_2) + s(n+i_4)s^*(n+i_2)\delta(i_1-i_3)\\+s(n+i_4)s^*(n+i_3)\delta(i_1-i_2)]e^{-j\frac{2\pi}{N}(i_1-i_2-i_3+i_4)k}\\+\sigma_\varepsilon^4\sum_{i_1=0}^{N-1}\sum_{i_2=0}^{N-1}\sum_{i_3=0}^{N-1}\sum_{i_4=0}^{N-1}[\delta(i_1-i_2)\delta(i_4-i_3) + \delta(i_1-i_3)\delta(i_4-i_2)]e^{-j\frac{2\pi}{N}(i_1-i_2-i_3+i_4)k}.\end{aligned}$$

The final form of the variance is

$$\sigma^2 = S_s^2(n,k) + 4N\sigma_\varepsilon^2 S_s(n,k) + 2N^2\sigma_\varepsilon^4 - (S_s(n,k) + N\sigma_\varepsilon^2)^2 = 2N\sigma_\varepsilon^2 S_s(n,k) + N^2\sigma_\varepsilon^4. \tag{7.161}$$

The variance is proportional to the signal spectrogram values $S_s(n,k)$.

Solution 7.32. (a) The expected value of the Fourier transform of $x(n)$ is

$$\mathrm{E}\{W_x(n,\omega)\} = \sum_{k=-L}^{L}\mathrm{E}\{x(n+k)x^*(n-k)\}e^{-j2\omega k}.$$

the signal is deterministic and it is not correlated with the white noise $\varepsilon(n)$,

$$\mathrm{E}\{x(n+k)x^*(n-k)\} = s(n+k)s^*(n-k) + r_{\varepsilon\varepsilon}(2k),$$

where $r_{\varepsilon\varepsilon}(2k)$ is the autocorrelation function of the additive noise $\varepsilon(n)$. The noise variance is σ_ε^2. Then,

$$\mathrm{E}\{x(n+k)x^*(n-k)\} = s(n+k)s^*(n-k) + \sigma_\varepsilon^2\delta(2k).$$

Final form of the expected value of $W_x(n,\omega)$ is

$$\mathrm{E}\{W_x(n,\omega)\} = W_s(n,\omega) + \sigma_\varepsilon^2.$$

b) The variance of $W_{xx}(n,\omega)$ follows from

$$\sigma^2 = \mathrm{E}\{W_x(n,\omega)W_x^*(n,\omega)\} - \mathrm{E}\{W_x(n,\omega)\}\mathrm{E}\{W_x^*(n,\omega)\}.$$

The first term can be written as

$$\mathrm{E}\{W_x(n,\omega)W_x^*(n,\omega)\} = \sum_{k_1=-L}^{L}\sum_{k_2=-L}^{L}\mathrm{E}\{x(n+k_1)x^*(n-k_1)x^*(n+k_2)x(n-k_2)\}e^{-j2\omega(k_1-k_2)}.$$

In the case of a Gaussian zero-mean white stationary noise complex-valued noise with i.i.d. real and imaginary parts $r_{\varepsilon\varepsilon^*}(k) = r_{\varepsilon^*\varepsilon}(k) = 0$ and we can write

$$\begin{aligned}
&\mathrm{E}\{x(n+k_1)x^*(n-k_1)x^*(n+k_2)x(n-k_2)\}\\
&= s(n+k_1)s^*(n-k_1)s^*(n+k_2)s(n-k_2) + s(n+k_1)s^*(n-k_1)r_{\varepsilon\varepsilon}(-2k_2)\\
&+ s(n+k_1)s^*(n+k_2)r_{\varepsilon\varepsilon}(k_2-k_1) + s^*(n-k_1)s(n-k_2)r_{\varepsilon\varepsilon}(k_1-k_2)\\
&+ s^*(n+k_2)s(n-k_2)r_{\varepsilon\varepsilon}(2k_1) + r_{\varepsilon\varepsilon}(2k_1)r_{\varepsilon\varepsilon}(-2k_2) + r_{\varepsilon\varepsilon}^2(k_1-k_2).
\end{aligned}$$

The note from the previous problem is used. Since

$$\mathrm{E}\{W_x(n,\omega)\}\mathrm{E}\{W_x^*(n,\omega)\} = W_s^2(n,\omega) + 2\sigma_\varepsilon^2 W_s(n,\omega) + \sigma_\varepsilon^4$$

the Fourier transform variance is

$$\begin{aligned}
\sigma^2 = \sum_{k_1=-L}^{L}\sum_{k_2=-L}^{L}&[s(n+k_1)s^*(n+k_2)r_{\varepsilon\varepsilon}(k_2-k_1) + r_{\varepsilon\varepsilon}^2(k_1-k_2)\\
+s^*(n-k_1)s(n-k_2)r_{\varepsilon\varepsilon}(k_1-k_2)]e^{-j\omega(k_1-k_2)} &= \sigma_\varepsilon^2\sum_{k=-L}^{L}(2|s(n+k)|^2+\sigma_\varepsilon^2).
\end{aligned}$$

For an FM signal $\sigma^2 = \sigma_\varepsilon^2(2L+1)(2A^2+\sigma_\varepsilon^2)$. This variance is constant.

Solution 7.33. The signal $s(n)$ and the noise $\varepsilon(n)$ are not correlated. In this case,

$$\begin{aligned}
r_{xx}(n) &= r_{ss}(n) + r_{\varepsilon\varepsilon}(n) = 4(0.5)^{|n|} + 2\delta(n)\\
R_{xx}(z) &= 4\frac{1}{1-0.5z^{-1}} + 4\frac{0.5z}{1-0.5z} + 1 = \frac{3z}{(2z-1)(2-z)} + 1.
\end{aligned} \tag{7.162}$$

(a) For the optimal filtering $d(n) = s(n)$. The cross-correlation of the desired and input signal is

$$r_{dx}(n) = \mathrm{E}\{d(k)x(n-k)\} = \mathrm{E}\{s(k)[s(k-n)+\varepsilon(k-n)]\} = r_{ss}(n) = 4(0.5)^{|n|}.$$

The optimal filter transfer function is

$$H(z) = \frac{R_{dx}(z)}{R_{xx}(z)} = \frac{\frac{3z}{(2z-1)(2-z)}}{\frac{3z}{(2z-1)(2-z)}+1} = \frac{3z}{-2z^2+8z-2}.$$

(b) For the optimal smoothing $d(n) = s(n-1)$, with

$$r_{dx}(n) = \mathrm{E}\{d(k)x(n-k)\} = \mathrm{E}\{s(k-1)[s(k-n)+\varepsilon(k-n)]\} = r_{ss}(n-1)$$

and

$$R_{dx}(z) = \sum_{n=-\infty}^{\infty} 4(0.5)^{|n-1|z^{-n}} = zR_{ss}(z) = \frac{3z^2}{(2z-1)(2-z)}$$

follows

$$H(z) = \frac{3z^2}{-2z^2+8z-2}.$$

(c) In the case of prediction $d(n) = s(n+1)$ and

$$r_{dx}(n) = \mathrm{E}\{d(k)x(n-k)\} = \mathrm{E}\{s(k+1)[s(k-n)+\varepsilon(k-n)]\} = r_{ss}(n+1),$$

$$R_{dx}(z) = \sum_{n=-\infty}^{\infty} 4(0.5)^{|n+1|z^{-n}} = z^{-1}R_{ss}(z) = \frac{3}{(2z-1)(2-z)}$$

with

$$H(z) = \frac{3}{-2z^2+8z-2}.$$

Solution 7.34. For the optimal filter the desired signal is $d(n) = s(n)$. The correlation functions are

$$\begin{aligned} r_{xx}(n) &= \mathrm{E}\{x(k)x^*(k-n)\} = \mathrm{E}\{(s(k)+\varepsilon(k))(s^*(k-n)+\varepsilon^*(k-n))\} \\ &= r_{ss}(n) + 2r_{s\varepsilon}(n) + r_{\varepsilon\varepsilon}(n) = 3(0.9)^{|n|} + 8\delta(n) \end{aligned}$$

and

$$r_{dx}(n) = \mathrm{E}\{s(k)[s^*(k-n)+\varepsilon^*(k-n)]\} = r_{ss}(n) + r_{s\varepsilon}(n) = 3(0.9)^{|n|} + 2\delta(n).$$

Calculation of the z-transforms and the filter transfer function is left to the reader.

Solution 7.35. The power spectral densities of the signal and the input noise are

$$S_{dd}(e^{j\omega}) = \begin{cases} 1-|\omega/2| & \text{for} \quad |\omega/2|<1 \\ 0 & \text{elsewhere} \end{cases}$$

and

$$S_{\varepsilon\varepsilon}(e^{j\omega}) = \begin{cases} 1-||\omega|-2| & \text{for} \quad |\omega-2|<1 \\ 0 & \text{elsewhere.} \end{cases}.$$

The optimal filter frequency response is

$$H(e^{j\omega}) = \frac{S_{dd}(e^{j\omega})}{S_{dd}(e^{j\omega}) + S_{\varepsilon\varepsilon}(e^{j\omega})}.$$

For $0 \le \omega < \pi$, we get

$$H(e^{j\omega}) = \begin{cases} 1 & \text{for} \quad \omega \le 1 \\ \frac{2-\omega}{\omega} & \text{for} \quad 1 < \omega \le 2 \\ 0 & \text{for} \quad 2 < \omega \le 3 \\ 1 & \text{for} \quad 3 < \omega < \pi \end{cases}$$

since for $1 < \omega \leq 2$ holds

$$H(e^{j\omega}) = \frac{1-\omega/2}{1-\omega/2+(1-|\omega-2|)} = \frac{1-\omega/2}{1-\omega/2+(1+(\omega-2))} = \frac{2-\omega}{\omega}.$$

The result for $-\pi \leq \omega < 0$ is symmetric and is shown in Fig. 7.38(bottom). The input SNR is

$$SNR_i = \frac{E_s}{E_\varepsilon} = \frac{2}{2} = 1$$

or 0 [dB]. The output SNR is

$$SNR_o = \frac{\frac{1}{2\pi}\int_{-\pi}^{\pi} S_{dd}(e^{j\omega})\left|H(e^{j\omega})\right|^2 d\omega}{\frac{1}{2\pi}\int_{-\pi}^{\pi} S_{\varepsilon\varepsilon}(e^{j\omega})^2\left|H(e^{j\omega})\right|^2 d\omega} = \frac{3/2+2\int_1^2(1-\frac{\omega}{2})\left|\frac{2-\omega}{\omega}\right|^2 d\omega}{2\int_1^2(1+(\omega-2))\left|\frac{2-\omega}{\omega}\right|^2 d\omega}$$

$$= \frac{10-12\ln 2}{16\ln 2-11} = 18.6181$$

or 12.7 [dB].

Solution 7.36. The calculation model is given by

$$\widehat{W_x}(n,k) = \sum_{m=0}^{N-1}\left[x_Q(n+m)x_Q(n-m)+e(n+m,n-m)\right]e^{-j2\pi mk/N}$$

$$= \sum_{m=0}^{N-1}\left\{\left[x(n+m)+e(n+m)\right]\left[(x(n-m)+e(n-m)\right]+e(n+m,n-m)\right\}e^{-j2\pi mk/N}.$$

The expected value is calculated from

$$\mathrm{E}\{\widehat{W_x}(n,k)\} = \sum_{m=0}^{N-1} x(n+m)x(n-m)e^{-j2\pi mk/N}$$

$$+\mathrm{E}\{\sum_{m=0}^{N-1} e(n+m)e(n-m)e^{-j2\pi mk/N}\} = W_x(n,k)+\frac{1}{12}\Delta^2.$$

It has been assumed that the errors in two different signal samples are not correlated $\mathrm{E}\{e(n+m)e(n-m)\}=0$ for $m \neq 0$ and that the signal and the error are not correlated, $\mathrm{E}\{x(n+m)e(n-m)\}=0$ for any m and n.

Part IV

Adaptive Systems and Neural Networks

Chapter 8

Adaptive Systems

8.1 INTRODUCTION

Classic systems for signal processing are designed to satisfy properties defined in advance. Their parameters are time-invariant. Adaptive systems change their parameters or form, in order to achieve the best possible performance. These systems are characterized by ability to observe variations in the input signal behavior and to react to these changes by adapting their parameters in order to improve the desired performance of the output signal. Adaptive systems have the ability to "learn" so that they can appropriately adapt the performance when the system environment is changed. By definition the adaptive systems are time-variant. These systems are often nonlinear as well. These two facts make the design and analysis of adaptive systems more difficult than in the case of classical time-invariant systems. Adaptive systems are the topic of this chapter.

Consider an adaptive system with one input and one output signal, as in Figure 8.1. In addition to the algorithm that transforms the input signal to the output signal, the adaptive system have a part that tracks the system performance and implements appropriate system changes. This control system takes into account the input signal, the output signal, and some additional information that can help in making a decision on how the system parameters should change.

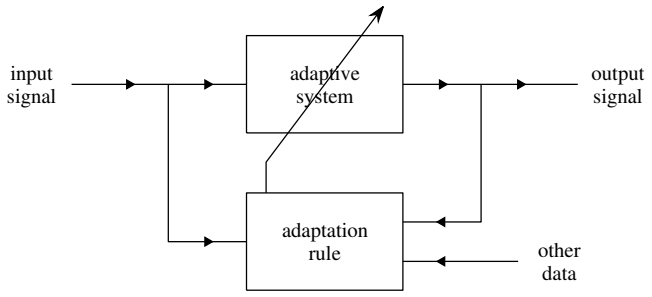


Figure 8.1 General adaptive system

Authors: *Ljubiša Stanković, Miloš Daković*

Part V

Time-Frequency Analysis

Chapter 10

Linear Time-Frequency Representations

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, that is, 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, that is, 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.

10.1 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.10.1.

Authors: *Ljubiša Stanković, Miloš Daković, Thayaparan Thayananthan*

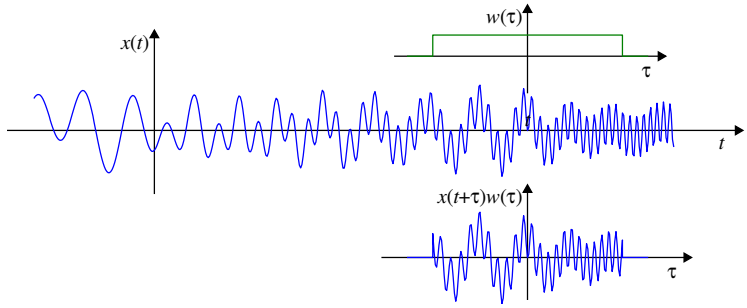


Figure 10.1 Illustration of the signal localization in the STFT calculation.

Analytic formulation of the STFT is

$$STFT(t,\Omega) = \int_{-\infty}^{\infty} x(t+\tau)\, w(\tau)\, e^{-j\Omega\tau}\, d\tau. \tag{10.1}$$

From (10.1) 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. 10.1). From the definition, it is clear that the STFT satisfies properties inherited from the Fourier transform (e.g., linearity).

By denoting $x_t(\tau) = x(t+\tau)$ we can conclude that the STFT is the Fourier transform of the signal $x_t(\tau)w(\tau)$,

$$STFT(t,\Omega) = \mathrm{FT}_\tau\{x_t(\tau)w(\tau)\}.$$

Another form of the STFT, with the same time-frequency performance, is

$$STFT_{II}(t,\Omega) = \int_{-\infty}^{\infty} x(\tau)w^*(\tau-t)e^{-j\Omega\tau}d\tau \tag{10.2}$$

where $w^*(t)$ denotes the conjugated window function.

It is obvious that definitions (10.1) and (10.2) differ only in phase, that is, $STFT_{II}(t,\Omega) = e^{-j\Omega t}STFT(t,\Omega)$ for real valued windows $w(\tau)$. We will mainly use the first STFT form.

Example 10.1. To illustrate the STFT application, let us perform time-frequency analysis of the following signal

$$x(t) = \delta(t-t_1) + \delta(t-t_2) + e^{j\Omega_1 t} + e^{j\Omega_2 t}. \tag{10.3}$$

The STFT of this signal equals

$$\begin{aligned} STFT(t,\Omega) &= w(t_1-t)e^{-j\Omega(t_1-t)} + w(t_2-t)e^{-j\Omega(t_2-t)} \\ &\quad + W(\Omega-\Omega_1)e^{j\Omega_1 t} + W(\Omega-\Omega_2)e^{j\Omega_2 t}, \end{aligned} \tag{10.4}$$

where $W(\Omega)$ is the Fourier transform of the used window. The STFT is depicted in Fig. 10.2 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.

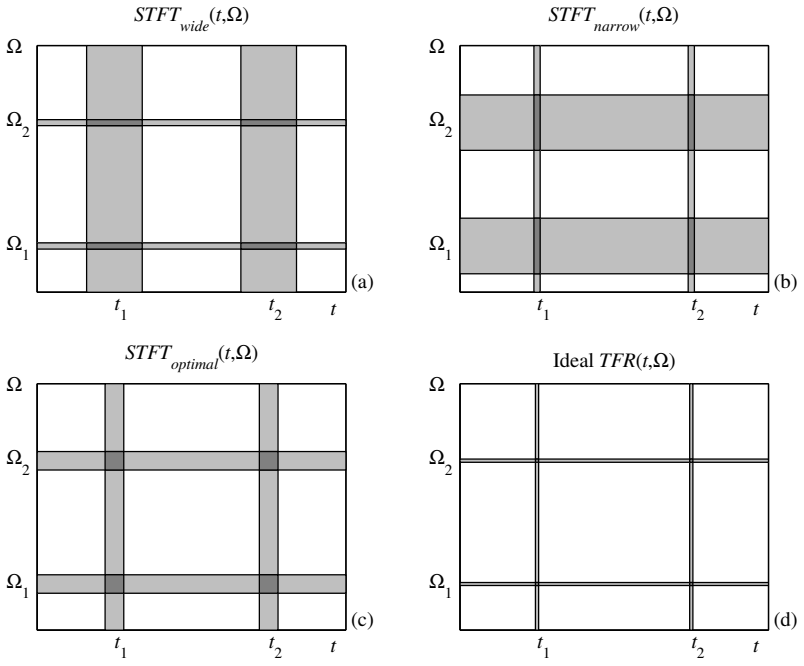


Figure 10.2 Time-frequency representation of a sum of two delta pulses and two sinusoids obtained using: (a) a wide window, (b) a narrow window, (c) a medium width window, and (d) an ideal time-frequency representation.

Example 10.2. The STFT of the signal

$$x(t) = e^{jat^2} \tag{10.5}$$

can be approximately calculated for a large a, 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 $|\tau| < T$.

★ Applying the stationary phase method (1.69), we get

$$STFT(t,\Omega) = \int_{-\infty}^{\infty} e^{ja(t+\tau)^2} w(\tau) e^{-j\Omega\tau} d\tau = \int_{-T}^{T} e^{ja(t+\tau)^2} w(\tau) e^{-j\Omega\tau} d\tau$$

$$\simeq e^{jat^2} e^{j(2at-\Omega)\tau_0} e^{ja\tau_0^2} w(\tau_0) \sqrt{\frac{2\pi j}{2a}} = e^{jat^2} e^{-j(2at-\Omega)^2/4a} w\left(\frac{\Omega - 2at}{2a}\right) \sqrt{\frac{\pi j}{a}} \tag{10.6}$$

since

$$2a(t+\tau_0) = \Omega.$$

Note that the STFT absolute value reduces to

$$|STFT(t,\Omega)| \simeq \left| w\left(\frac{\Omega - 2at}{2a}\right)\right| \sqrt{\frac{\pi}{a}}. \tag{10.7}$$

In this case, the width of $|STFT(t,\Omega)|$ along frequency does not decrease with an 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,

$$D_o = 4aT + \frac{2c}{T}, \tag{10.8}$$

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.10.3.

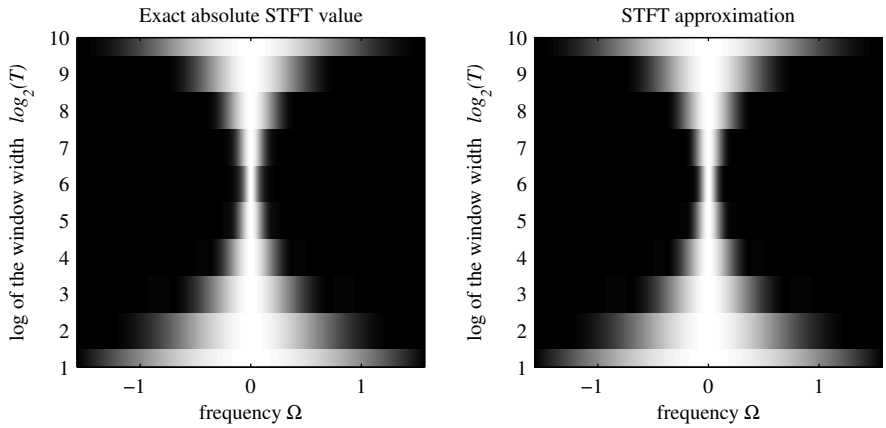


Figure 10.3 Exact absolute STFT value of a linear FM signal at $t = 0$ for various window widths $T = 2,4,8,16,..,1024$ (left) and its approximation calculated as an appropriately scaled convolution of the time and frequency domain window $w(\tau)$ (right).

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

$$T_o = \sqrt{\frac{c}{2a}}. \tag{10.9}$$

As expected, for a sinusoid, $a \to 0$, $T_o \to \infty$. 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 $2 - \frac{2c}{T^2} = 0$ 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

$$\begin{aligned} STFT(t,\Omega) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} X(\theta)\, e^{j(t+\tau)\theta}\, w(\tau)\, e^{-j\Omega\tau}\, d\theta\, d\tau \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\theta) W(\Omega-\theta)\, e^{jt\theta} d\theta = \left[X(\Omega)e^{jt\Omega}\right] *_{\Omega} W(\Omega), \end{aligned} \tag{10.10}$$

where $*_{\Omega}$ denotes convolution in Ω. It may be interpreted as an inverse Fourier transform of the frequency localized version of $X(\Omega)$, with localization window $W(\Omega) = \mathrm{FT}\{w(\tau)\}$.

The energetic version of the STFT, called the spectrogram, is defined by

$$SPEC(t,\Omega) = |\, STFT(t,\Omega)\,|^2 = \left|\int_{-\infty}^{\infty} x(\tau) w^*(\tau - t) e^{-j\Omega\tau} d\tau\right|^2 = \left|\int_{-\infty}^{\infty} x(t+\tau) w^*(\tau) e^{-j\Omega\tau} d\tau\right|^2.$$

Obviously, linearity property is lost in the spectrogram.

Example 10.3. For illustration consider two different signals $x_1(t)$ and $x_2(t)$ producing the same amplitude of the Fourier transform, Fig. 10.4,

$$\begin{aligned} x_1(t) &= \sin\left(122\pi\frac{t}{128}\right) - \cos\left(42\pi\frac{t}{128} - \frac{16}{11}\pi\left(\frac{t-128}{64}\right)^2\right) \\ &\quad - 1.2\cos\left(94\pi\frac{t}{128} - 2\pi\left(\frac{t-128}{64}\right)^2 - \pi\left(\frac{t-120}{64}\right)^3\right) e^{-\left(\frac{t-140}{75}\right)^2} \\ &\quad - 1.6\cos\left(15\pi\frac{t}{128} - 2\pi\left(\frac{t-50}{64}\right)^2\right) e^{-\left(\frac{t-50}{16}\right)^2} \\ x_2(t) &= x_1(255-t). \end{aligned} \tag{10.11}$$

Their spectrograms are shown in Fig.10.5. 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 decreasing frequency as time progresses, and in the second signal with increasing frequency), and two chirps (one appearing at different time instants and the other having different frequency variations).

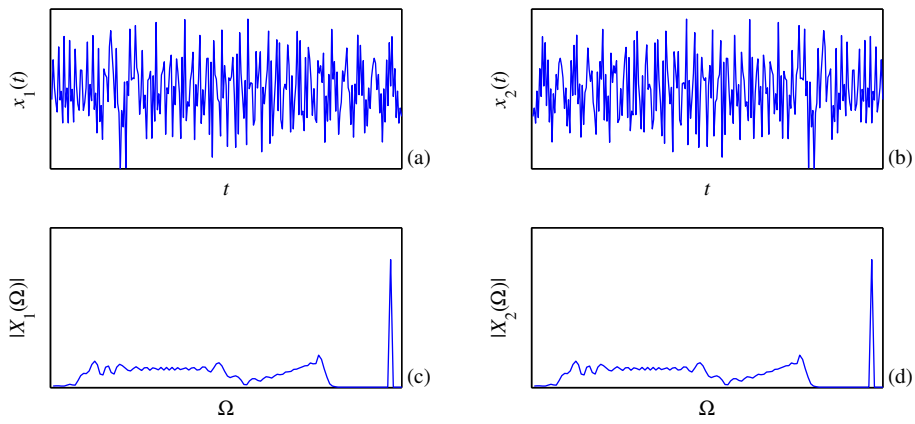


Figure 10.4 Two different signals $x_1(t) \neq x_2(t)$ with the same amplitudes of their Fourier transforms, $|X_1(\Omega)| = |X_2(\Omega)|$.

10.2 STFT INVERSION

The signal can be obtained from the STFT calculated at an instant t, $STFT(t,\Omega)$, as its inverse Fourier transform

$$x(t+\tau)w(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} STFT(t,\Omega)e^{-j\Omega\tau}\,d\Omega.$$

This relation can be theoretically used for the signal within the region $w(\tau) \neq 0$,

$$x(t+\tau) = \frac{1}{w(\tau)}\frac{1}{2\pi}\int_{-\infty}^{\infty} STFT(t,\Omega)e^{-j\Omega\tau}\,d\Omega.$$

In practice it is used within the region of significant window $w(\tau)$ values.

If the STFT is calculated at a set of the discrete-time instants $t = iR$, shifted for R for each next STFT calculation, then a set of the STFT values and its inverse transforms

$$x(iR+\tau)w(\tau) = \int_{-\infty}^{\infty} STFT(iR,\Omega)e^{-j\Omega\tau}\,d\tau$$

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.

For the correct reconstruction, the segments $r_i(\tau) = x(iR+\tau)w(\tau)$, whose positions on the time axis are illustrated in Fig. 10.1, should be properly re-positioned in terms of t-axis, using $\tau = t - iR$, and summed

$$\sum_i r_i(t-iR) = \sum_i x(iR+t-iR)w(t-iR) = x(t)\sum_i w(t-iR). \tag{10.12}$$

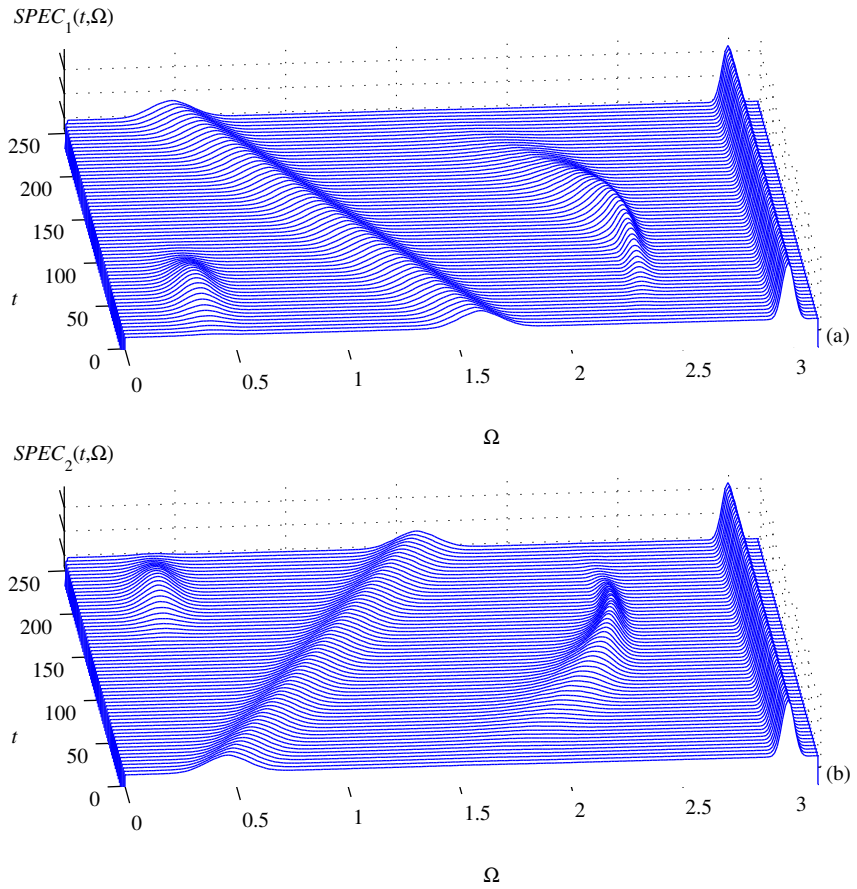


Figure 10.5 Spectrograms of the signals presented in Fig.10.4.

If the sum of shifted versions of the windows is constant (without loss of generality assume equal to 1),

$$\sum_i w(\tau - iR) = 1, \tag{10.13}$$

then

$$x(t) = \sum_i \frac{1}{2\pi} \int_{-\infty}^{\infty} STFT(iR,\Omega) e^{-j\Omega(t-iR)}\, d\Omega.$$

The condition $\sum_i w(\tau - iR) = 1$ is equivalent to the requirement that the periodic extension of the window, with the period R, is constant (see Fig. 10.6). The periodic extension of a continuous signal corresponds to the sampling of the window Fourier transform at $\Omega = 2\pi k/R$ in the Fourier domain, (1.66). This means that

$$W(\frac{2\pi}{R}k) = 0$$

for an integer value of k, $k \neq 0$, when $\sum_i w(\lambda - iR) = 1$.

The inversion of the STFT, defined in (10.2), gives

$$x(\tau)w^*(\tau - t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} STFT(t,\Omega)e^{j\Omega\tau}d\tau. \tag{10.14}$$

If this STFT is available at a discrete set of instants, $t = iR$ (or any other set of discrete instants t_i), then the summation of $STFT(t,\Omega)$ over all values calculated at different instants t is

$$\sum_i \frac{1}{2\pi}\int_{-\infty}^{\infty} STFT(iR,\Omega)e^{j\Omega\tau}d\Omega = \sum_i x(\tau)w^*(\tau - iR) = x(\tau) \tag{10.15}$$

if the condition in (10.13) holds for the window $w^*(\tau)$.

This kind of inversion is called the constant overlap-add (COLA) method. Many window forms can easily satisfy the condition in (10.13). The reconstruction of discrete-time signals will be considered in Section 10.5.

The other signal reconstruction approach, called the weighted overlap-add (WOLA) method uses a weighted inversion of the STFT. Namely, the signal $x(\tau)w^*(\tau - t)$ inverted in (10.14), commonly with a real-valued window, is additionally weighted by the synthesis window, unusually the same as the analysis window, to produce

$$r_i(\tau) = [x(\tau)w^*(\tau - iR)]w(\tau - iR) = x(\tau)|w(\tau - iR)|^2.$$

The inversion relation (for a real-valued window $w(\tau)$) is now

$$\sum_i \Big(\frac{1}{2\pi}\int_{-\infty}^{\infty} STFT(iR,\Omega)e^{j\Omega\tau}d\Omega \;\; w(\tau - iR)\Big) = \sum_i x(\tau)w^2(\tau - iR) = x(\tau) \tag{10.16}$$

if the condition

$$\sum_i w^2(\tau - iR) = 1, \tag{10.17}$$

holds. The same condition can be derived from the following analysis.

The STFT can be considered as a projection (inner product) of the signal, $x(\tau)$, to the time-frequency kernel function

$$h_{t,\Omega}(\tau) = w(\tau - t)e^{j\Omega\tau},$$

that is

$$STFT(t,\Omega) = \Big\langle x(\tau), h_{t,\Omega}(\tau)\Big\rangle_\tau = \Big\langle x(\tau), w(\tau - t)e^{j\Omega\tau}\Big\rangle_\tau = \int_{-\infty}^{\infty} x(\tau)w^*(\tau - t)e^{-j\Omega\tau}d\tau. \tag{10.18}$$

The back-projection of the STFT to the same (conjugate) kernel is $\Big\langle STFT(t,\Omega), h^*_{t,\Omega}(\tau)\Big\rangle_{t,\Omega}$ or

$$\Big\langle STFT(t,\Omega), w^*(\tau - t)e^{-j\Omega\tau}\Big\rangle_{t,\Omega} = \sum_{t_i}\frac{1}{2\pi}\int_{-\infty}^{\infty} STFT(t_i,\Omega)w(\tau - t_i)e^{j\Omega\tau}d\Omega. \tag{10.19}$$

With $t = t_i = iR$, relation (10.19) reduces to (10.16), with the same reconstruction condition.

10.3 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.

10.3.1 Rectangular Window

The simplest window is the rectangular one, defined by

$$w(\tau) = \begin{cases} 1 & \text{for } |\tau| < T \\ 0 & \text{elsewhere} \end{cases} \tag{10.20}$$

whose Fourier transform is

$$W_R(\Omega) = \int_{-T}^{T} e^{-j\Omega\tau} d\tau = \frac{2\sin(\Omega T)}{\Omega}. \tag{10.21}$$

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 Ω 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

$$W(e^{j\omega}) = \sum_{n=-N/2}^{N/2-1} e^{-j\omega n} = \frac{\sin(\omega N/2)}{\sin(\omega/2)}.$$

10.3.2 Triangular (Bartlett) Window

The triangular window is defined by

$$w(\tau) = \begin{cases} 1 - |\tau/T| & \text{for } |\tau| < T \\ 0 & \text{elsewhere.} \end{cases} \tag{10.22}$$

This window could be considered as a self convolution of the rectangular window of the duration T, since

$$\begin{aligned} &[u(t + T/2) - u(t - T/2)] *_t [u(t + T/2) - u(t - T/2)] \\ &= (1 - |\tau/T|)\,[u(t + T) - u(t - T)]. \end{aligned}$$

The Fourier transform of the triangular window is a product of two Fourier transforms of the rectangular window of the width T,

$$W_T(\Omega) = \frac{4\sin^2(\Omega T/2)}{\Omega^2}. \tag{10.23}$$

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

$$w(n) = \left[1 - \frac{2|n|}{N}\right][u(n+N/2) - u(n-N/2)].$$

In the frequency domain its form is

$$W(e^{j\omega}) = \sum_{n=-N/2}^{N/2-1} \left[1 - \frac{2|n|}{N}\right] e^{-j\omega n} = \frac{\sin^2(\omega N/4)}{\sin^2(\omega/2)}.$$

10.3.3 Hann(ing) Window

The Hann(ing) window is of the form

$$w(\tau) = \begin{cases} 0.5(1+\cos(\pi\tau/T)) = \cos^2(\pi\tau/(2T)), & \text{for } |\tau| < T \\ 0, & \text{elsewhere.} \end{cases} \tag{10.24}$$

Since $\cos(\pi\tau/T) = [\exp(j\pi\tau/T) + \exp(-j\pi\tau/T)]/2$, the Fourier transform of this window is related to the Fourier transform of the rectangular window of the same width as

$$\begin{aligned} W_H(\Omega) &= \frac{1}{2}W_R(\Omega) + \frac{1}{4}W_R(\Omega - \pi/T) + \frac{1}{4}W_R(\Omega + \pi/T) \\ &= \frac{\pi^2 \sin(\Omega T)}{\Omega(\pi^2 - \Omega^2 T^2)}. \end{aligned} \tag{10.25}$$

The function $W_H(\Omega)$ decays in frequency as Ω^3, much faster than $W_R(\Omega)$.

The discrete-time domain form is

$$w(n) = 0.5\left[1 + \cos\left(\frac{2\pi n}{N}\right)\right][u(n+N/2) - u(n-N/2)],$$

with the DFT of the form

$$W(k) = \frac{N}{2}\delta(k) + \frac{N}{4}\delta(k+1) + \frac{N}{4}\delta(k-1).$$

When the window is used on the data set from 0 to $N-1$ then

$$\begin{aligned} w(n) &= 0.5\left[1 - \cos\left(\frac{2\pi n}{N}\right)\right][u(n) - u(n-N)] \\ W(k) &= \frac{N}{2}\delta(k) - \frac{N}{4}\delta(k+1) - \frac{N}{4}\delta(k-1). \end{aligned}$$

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 the windowed signal is a moving average (smoothed) form of the original signal,

$$\begin{aligned} \text{DFT}\{x(n)w(n)\} &= \frac{1}{N}\text{DFT}\{x(n)\} *_k \text{DFT}\{w(n)\} \\ &= \frac{1}{4}X(k+1) + \frac{1}{2}X(k) + \frac{1}{4}X(k-1). \end{aligned}$$

Example 10.4. Find the window that will correspond to the frequency smoothing $(X(k+1)+X(k)+X(k-1))/3$, that is, to

$$\begin{aligned}\mathrm{DFT}\{x(n)w(n)\} &= \frac{1}{N}\mathrm{DFT}\{x(n)\} *_k \mathrm{DFT}\{w(n)\} \\ &= \frac{1}{3}X(k+1)+\frac{1}{3}X(k)+\frac{1}{3}X(k-1).\end{aligned}$$

★ 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 10.5. 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

$$STFT(t,\Omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\theta)W(\Omega-\theta)e^{jt\theta}d\theta$$

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

$$STFT_H(t,\Omega) = \frac{1}{2}STFT_R(t,\Omega)+\frac{1}{4}STFT_R\left(t,\Omega-\frac{\pi}{T}\right)+\frac{1}{4}STFT_R\left(t,\Omega+\frac{\pi}{T}\right). \qquad (10.26)$$

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. 10.2.

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_Ω in the frequency domain. This essentially prevents us from achieving the ideal resolution ($d_t = 0$ 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.

The Hann(ing) window satisfies the constant overlap-add (COLA) reconstruction condition, $\sum_i w(\tau - iR) = 1$, with $R = T$, as shown in Fig. 10.6. This property follows from $\cos^2(\pi\tau/(2T)) + \cos^2(\pi(\tau \pm T)/(2T)) = \cos^2(\pi\tau/(2T)) + \sin^2(\pi\tau/(2T)) = 1$.

The same condition can be satisfied with $R = T/2$, $R = T/4$, ..., after an appropriate scaling of the window amplitude.

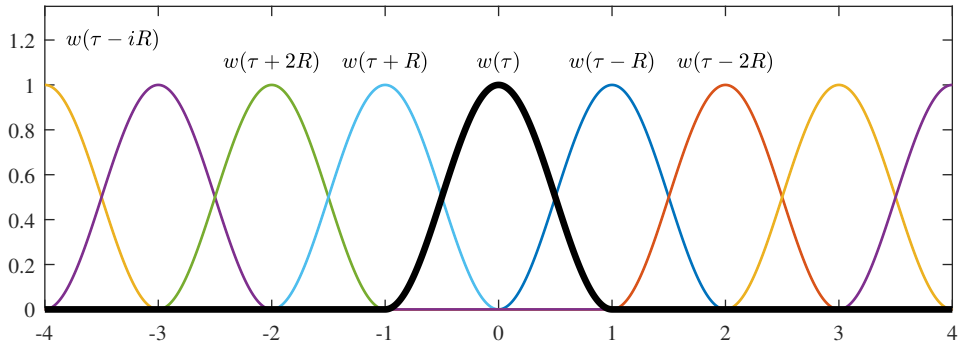


Figure 10.6 Hann(ing) window and its shifted versions, that satisfy the constant overlap-add (COLA) reconstruction condition $\sum_i w(\tau - iR) = 1$, with $R = 1$.

10.3.4 Hamming Window

This window has the form

$$w(\tau) = \begin{cases} 0.54 + 0.46\cos(\pi\tau/T)) & \text{for } |\tau| < T \\ 0 & \text{elsewhere.} \end{cases} \tag{10.27}$$

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 (that is, 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

$$a = 25/46 \cong 0.54. \tag{10.28}$$

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

$$W(k) = 0.54N\delta(k) + 0.23N\delta(k+1) + 0.23N\delta(k-1).$$

10.3.5 Blackman and Kaiser Windows

In some applications, it is crucial that the sidelobes are suppressed, as much as possible. This is achieved using windows of more complicated forms, like the Blackman window. It is defined by

$$w(\tau) = \begin{cases} 0.42 + 0.5\cos(\pi\tau/T) + 0.08\cos(2\pi\tau/T) & \text{for } |\tau| < T \\ 0 & \text{elsewhere.} \end{cases} \tag{10.29}$$

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 (that is, 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

$$w(n) = \left[0.42 + 0.5\cos\left(\frac{2\pi n}{N}\right) + 0.08\cos\left(\frac{4\pi n}{N}\right)\right][u(n+\frac{N}{2}) - u(n-\frac{N}{2})]$$
$$W(k) = [0.42\delta(k) + 0.25(\delta(k+1) + \delta(k-1)) + 0.04(\delta(k+2) + \delta(k-2))]N.$$

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 as -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. 10.7.

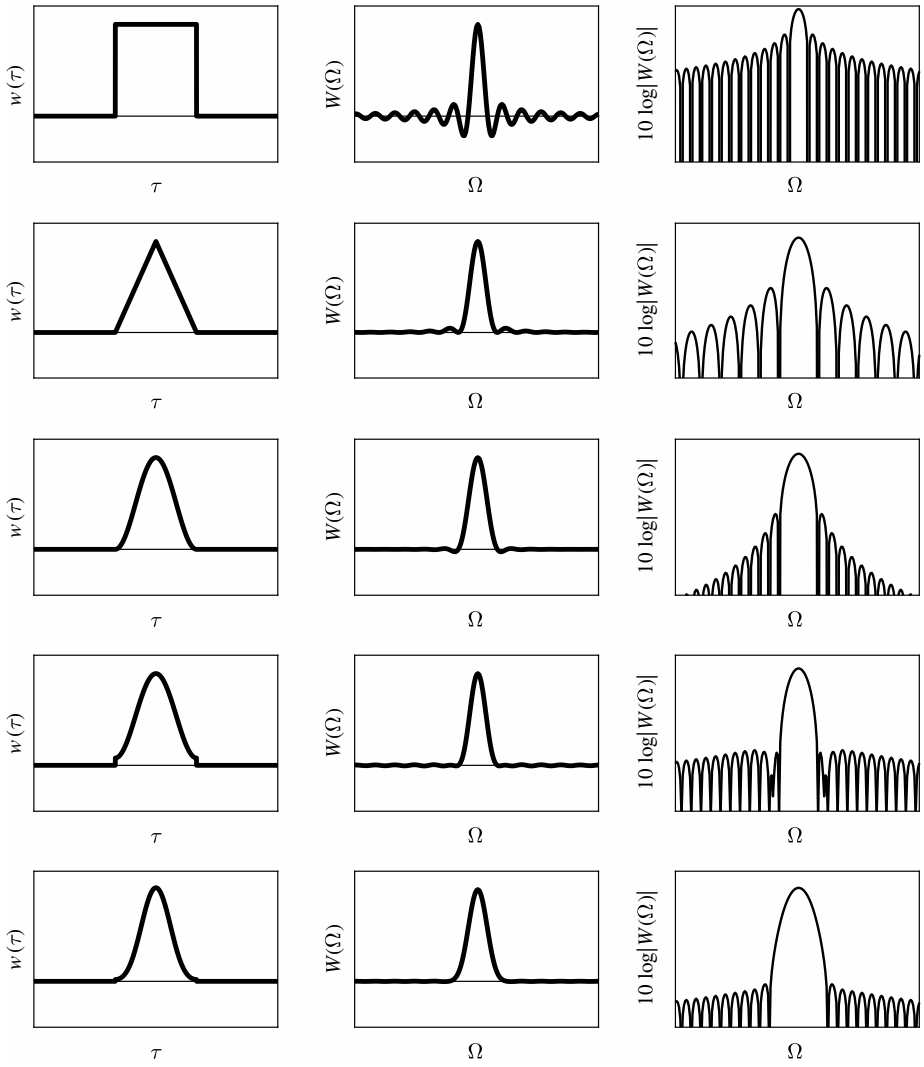


Figure 10.7 Windows in the time and frequency domains: rectangular window (first row), triangular (Bartlett) window (second row), Hann(ing) window (third row), Hamming window (fourth row), and Blackman window (fifth row).

Example 10.6. Calculate the STFT of the signals $x_1(t) = 2\cos(4\pi t/T) + 2\cos(12\pi t/T)$ and $x_2(t) = 2\cos(4\pi t/T) + 0.001\cos(64\pi t/T)$ at $t = 0$. Use a Hamming and a Blackman window with $T = 128$ and $\Delta t = 1$. Comment the results.

★ The STFT at $t = 0$ is shown in Fig.10.8. The resolution of close components in $x_1(t)$ is better when the Hamming window is used, since the main lobe of the Blackman window is wider. Small

signal component in $x_2(t)$ is visible in the STFT with the Blackman window since its side-lobes are lower.

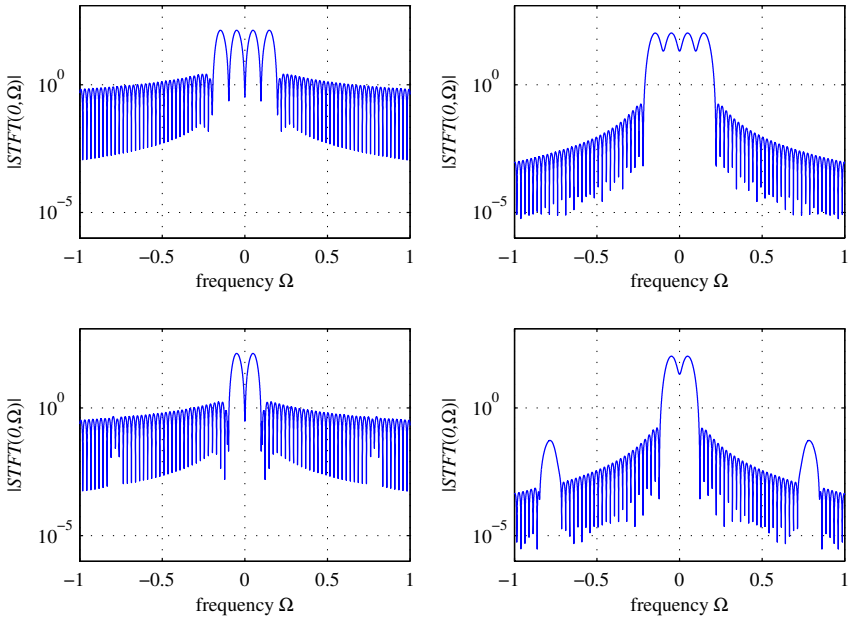


Figure 10.8 The STFT at $n = 0$ calculated using the Hamming window (left) and the Blackman window (right) of the signals $x_1(n)$ (top) and $x_2(n)$ (bottom).

10.4 REALIZATIONS OF THE STFT

Discretization and realizations of the STFT will be discussed in this section. Recursive realization which is appropriate for the on-line implementation of the STFT will be presented, along with the filter bank form of the STFT.

10.4.1 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 Δt we get

$$STFT(t,\Omega) = \int_{-\infty}^{\infty} x(t+\tau)w(\tau)e^{-j\Omega\tau}d\tau \simeq \sum_{m=-\infty}^{\infty} x((n+m)\Delta t)w(m\Delta t)e^{-jm\Delta t\Omega}\Delta t.$$

By denoting

$$x(n) = x(n\Delta t)\Delta t$$

and normalizing the frequency Ω by Δt, $\omega = \Delta t\Omega$, we get the time-discrete form of the STFT as

$$STFT(n,\omega) = \sum_{m=-\infty}^{\infty} w(m)x(n+m)e^{-jm\omega}. \tag{10.30}$$

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 τ 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 2π. The relation between the analog and the discrete-time form is

$$STFT(n,\omega) = \sum_{k=-\infty}^{\infty} STFT(n\Delta t, \Omega + 2k\Omega_0) \text{ with } \omega = \Delta t\Omega.$$

The sampling interval Δt is related to the period in frequency as $\Delta t = \pi/\Omega_0$. 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 Ω_m is the maximum 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 maximum 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 Ω_m, that is, for $|\Omega| > \Omega_m$, can be neglected, and that overlapping of the frequency content above Ω_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

$$STFT(n,k) = STFT(n,\omega)|_{\omega=\frac{2\pi}{N}k} = \sum_{m=-N/2}^{N/2-1} w(m)x(n+m)e^{-j2\pi mk/N} \tag{10.31}$$

and it can be efficiently calculated using the fast DFT routines

$$STFT(n,k) = \text{DFT}_m\{w(m)x(n+m)\},$$

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 10.7. 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 (10.31) with $N=4$, when $k=-2,-1,0,1$, for given instant n, the following matrix notation can be used

$$\begin{bmatrix} STFT(n,-2) \\ STFT(n,-1) \\ STFT(n,0) \\ STFT(n,1) \end{bmatrix} = \begin{bmatrix} W_4^4 & W_4^2 & 1 & W_4^{-2} \\ W_4^2 & W_4^1 & 1 & W_4^{-1} \\ 1 & 1 & 1 & 1 \\ W_4^{-2} & W_4^{-1} & 1 & W_4^1 \end{bmatrix} \begin{bmatrix} x(n-2) \\ x(n-1) \\ x(n) \\ x(n+1) \end{bmatrix}$$

or

$$\mathbf{STFT}(n) = \mathbf{W}_4 \mathbf{x}(n)$$

with $\mathbf{STFT}(n) = [STFT(n,-2)\ STFT(n,-1)\ STFT(n,0)\ STFT(n,1)]^T$, $\mathbf{x}(n) = [x(n-2)\ x(n-1)\ x(n)\ x(n+1)]^T$, and $\mathbf{W}_4$ is the DFT matrix of order four with elements $W_4^{mk} = \exp(-j2\pi mk/N)$. Here, a rectangular window is assumed. Including the window function, the previous relation can be written as

$$\mathbf{STFT}(n) = \mathbf{W}_4 \mathbf{H_4} \mathbf{x}(n),$$

with

$$\mathbf{H_4} = \begin{bmatrix} w(-2) & 0 & 0 & 0 \\ 0 & w(-1) & 0 & 0 \\ 0 & 0 & w(0) & 0 \\ 0 & 0 & 0 & w(1) \end{bmatrix}$$

being a diagonal matrix whose elements are the window values $w(m)$, $\mathbf{H_4} = \text{diag}(w(m))$, $m=-2,-1,0,1$ and

$$\mathbf{W}_4\mathbf{H_4} = \begin{bmatrix} w(-2)W_4^4 & w(-1)W_4^2 & w(0) & w(1)W_4^{-2} \\ w(-2)W_4^2 & w(-1)W_4^1 & w(0) & w(1)W_4^{-1} \\ w(-2) & w(-1) & w(0) & w(1) \\ w(-2)W_4^{-2} & w(-1)W_4^{-1} & w(0) & w(1)W_4^1 \end{bmatrix}.$$

All STFT values for the nonoverlapping case are obtained as

$$\mathbf{STFT} = \mathbf{W}_4\mathbf{H_4} \begin{bmatrix} x(0) & x(4) & x(8) & x(12) \\ x(1) & x(5) & x(9) & x(13) \\ x(2) & x(6) & x(10) & x(14) \\ x(3) & x(7) & x(11) & x(15) \end{bmatrix} = \mathbf{W}_4\mathbf{H_4}\mathbf{X}_{4,4},$$

where **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

$$\begin{aligned} \mathbf{STFT} &= \begin{bmatrix} \mathbf{STFT}_M(0) & \mathbf{STFT}_M(M) & \dots & \mathbf{STFT}_M(N-M) \end{bmatrix} \\ &= \begin{bmatrix} STFT(2,-2) & STFT(6,-2) & STFT(10,-2) & STFT(14,-2) \\ STFT(2,-1) & STFT(6,-1) & STFT(10,-1) & STFT(14,-1) \\ STFT(2,0) & STFT(6,0) & STFT(10,0) & STFT(14,0) \\ STFT(2,1) & STFT(6,1) & STFT(10,1) & STFT(14,1) \end{bmatrix}. \end{aligned}$$

The matrix $\mathbf{X}_{4,4}$ is formed of four successive signal values in each column. Notation $\mathbf{X}_{N,R}$ 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

$$\mathbf{STFT} = \mathbf{H_4 W_4} \begin{bmatrix} x(0)\, x(1)\, x(2) \ldots x(10)\, x(11)\, x(12) \\ x(1)\, x(2)\, x(3) \ldots x(11)\, x(12)\, x(13) \\ x(2)\, x(3)\, x(4) \ldots x(12)\, x(13)\, x(14) \\ x(3)\, x(4)\, x(5) \ldots x(13)\, x(14)\, x(15) \end{bmatrix}$$

$$\mathbf{STFT} = \mathbf{W_4 H_4 X}_{4,1}.$$

The step R defines the difference of the 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 nonoverlapped calculation). In the second example difference is $R = 1 < 4$, meaning the overlapped STFT calculation. Note that the window function $\mathbf{H_N}$ and the DFT matrix $\mathbf{W_N}$ remain the same for both cases.

Example 10.8. Consider the 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 maximum 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 $d_t = 2T = N\Delta t = 32\Delta t$. The width of the mainlobe in frequency domain d_w is defined by $32 d_w \Delta t = 4\pi$, 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 $\pi/(\Omega_m + d_w) = 1/256$.

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, $w(\tau)e^{-(t+\tau)^2} \cong w(\tau)e^{-t^2}$ for $-1/16 < \tau \leq 1/16$. Then, according to the stationary phase method, we can write the STFT approximation,

$$|STFT(t,\Omega)|^2 = \tfrac{1}{6} e^{-2t^2} w^2\left(\tfrac{\Omega + 12\pi t + 32\pi}{12\pi}\right) + \tfrac{1}{32} e^{-8(t-1)^2} w^2\left(\tfrac{\Omega - 32\pi t - 160\pi}{32\pi}\right)$$

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 $-128\pi \leq \Omega < 128\pi$. 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 -256π, 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

$$x_r(t) = e^{-t^2} e^{-j6\pi t^2 - j32\pi t} + e^{-4(t-1)^2} e^{j16\pi t^2 + j(160-256)\pi t},$$

with approximation,

$$|STFT(t,\Omega)|^2 = \tfrac{1}{6} e^{-2t^2} w^2\left(\tfrac{\Omega+12\pi t+32\pi}{12\pi}\right) + \tfrac{1}{32} e^{-8(t-1)^2} w^2\left(\tfrac{\Omega-32\pi t-96\pi}{32\pi}\right), \qquad (10.32)$$

with $t = n/128$ and $\Omega = 128\omega$ within $-\pi \leq \omega < \pi$ or $-128\pi \leq \Omega < 128\pi$.

10.4.2 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 (10.31).

For other window forms, the STFT can be obtained from the STFT obtained by using the rectangular window. For example, according to (10.26) 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).

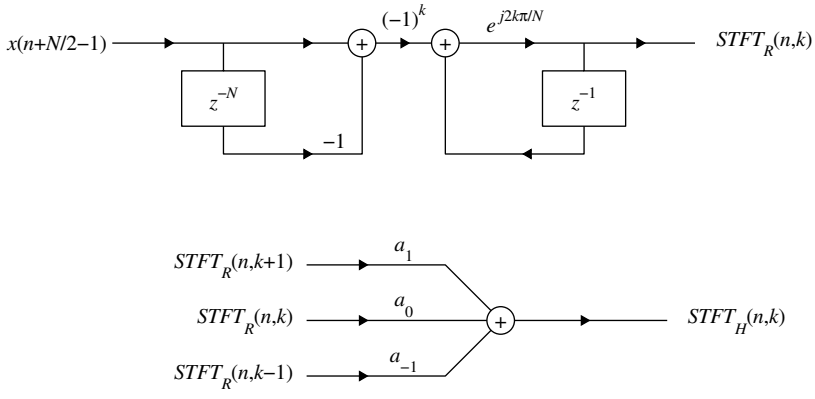


Figure 10.9 Recursive implementation of the STFT for the rectangular and other windows.

A system for the recursive implementation of the STFT is shown in Fig. 10.9. The STFT obtained using the rectangular window is denoted by $STFT_R(n,k)$, Fig.10.9, while the values of coefficients are

$$(a_{-1},a_0,a_1) = (\frac{1}{4},\frac{1}{2},\frac{1}{4}),$$
$$(a_{-1},a_0,a_1) = (0.23,0.54,0.23),$$
$$(a_{-2},a_{-1},a_0,a_1,a_2) = (0.04,0.25,0.42,0.25,0.04)$$

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 $\mathbf{STFT}_H = 0.5\mathbf{STFT}+0.25\mathbf{STFT}_{\downarrow} + 0.25\mathbf{STFT}_{\uparrow}$, where $\mathbf{STFT}_{\downarrow}$ and $\mathbf{STFT}_{\uparrow}$ are the STFT matrices with circularly shifted rows down and up for one position, respectively.

10.4.3 Filter Bank STFT Implementation

According to (10.1), the STFT can be written as a convolution

$$STFT(t,\Omega) = \int_{-\infty}^{\infty} x(t+\tau)\, w(\tau)\, e^{-j\Omega\tau}\, d\tau = \int_{-\infty}^{\infty} x(t-\tau) w(\tau) e^{j\Omega\tau} d\tau = x(t) *_t \left[w(t) e^{j\Omega t} \right]$$

where an even, real valued, window function is assumed, $w(\tau) = w(-\tau)$. For a discrete set of frequencies $\Omega_k = k\Delta\Omega = 2\pi k/(N\Delta t)$, $k = 0,1,2,\ldots,N-1$, and discrete values of signal, we get that the discrete STFT, (10.31), is an output of the filter bank with impulse responses

$$STFT(n,k) = x(n) *_n \left[w(n) e^{j2\pi kn/N} \right] = x(n) *_n h_k(n)$$
$$h_k(n) = w(n) e^{j2\pi kn/N}$$
$$k = 0,1,\ldots,N-1$$

what is illustrated in Fig.10.10. 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.

10.4.3.1 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

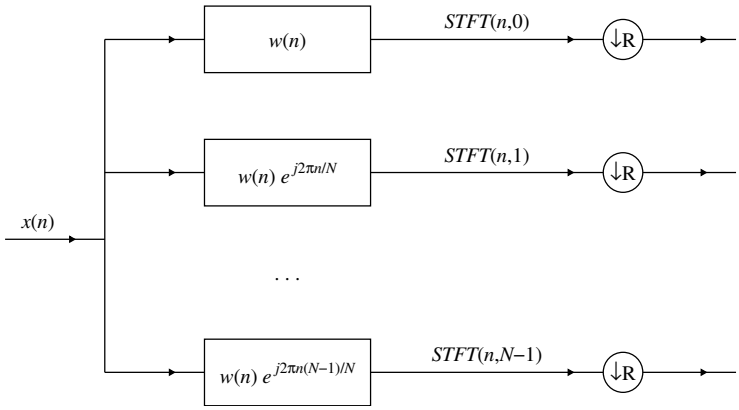


Figure 10.10 Filter bank realization of the STFT

calculated with a rectangular window into a STFT with, for example, the Hann(ing), Hamming, or Blackman window, is presented in Fig.10.9.

If a signal $x(n)$ is of duration M, with $0 \leq n \leq M-1$, 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

$$STFT_N(n,k) = \sum_{m=-N/2}^{N/2-1} w(m)x(n+m)e^{-j2\pi mk/M}$$
$$k = -M/2, -M/2+1, \ldots, -1, 0, 1, \ldots, M/2-1.$$

Example 10.9. 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.10.11. Its values are $STFT_8(n,k)$ at $n = 4,12,20\ldots,60$ and $-4 \leq k \leq 3$. The nonoverlapping STFT values obtained using the rectangular window are shifted in frequency, scaled, and added up, Fig. 10.12, to produce the STFT with a Hamming window, Fig. 10.13.

The STFT calculation for the same linear FM signal will be repeated for the overlapping STFT with step $R = 1$, when $n = 0,1,2,3,5,\ldots,63$ is used. Here, it has been assumed that the linear FM signal is available for all $-N/2 \leq m+n \leq M-1+N/2-1$. Results for the rectangular and Hamming window (obtained by a simple matrix calculation from the rectangular window case) are presented in Fig.10.14. Three window widths are used here.

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.10.15. Note that regarding to the amount of information all these figures do not differ from the basic time-frequency representation presented in Fig.10.11.

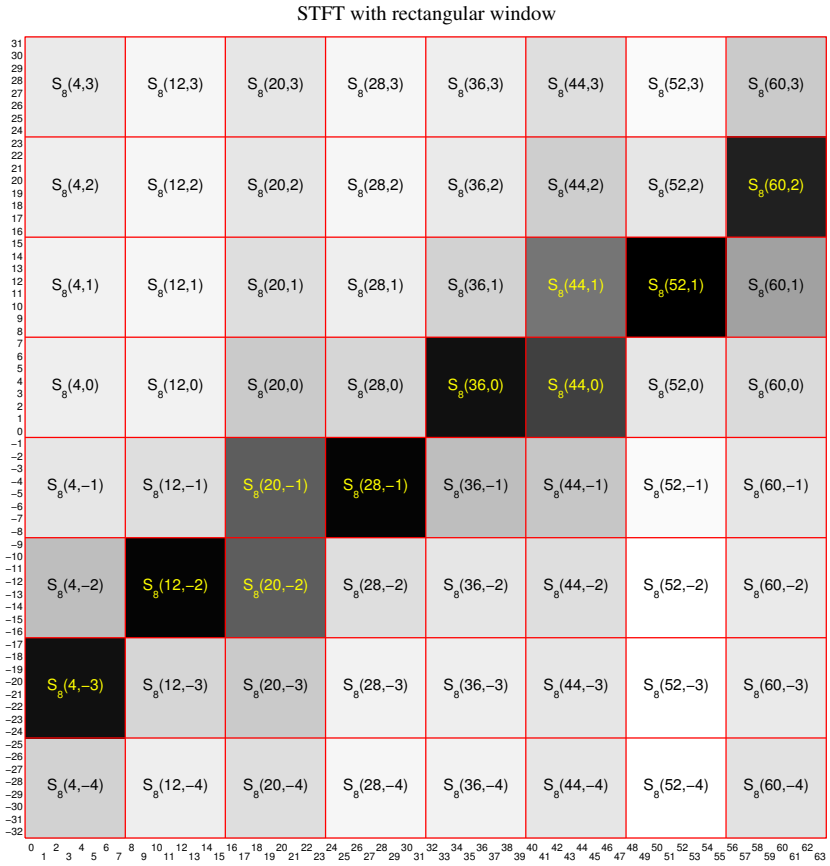


Figure 10.11 The STFT of a linear FM signal $x(n)$ calculated using a rectangular window of the width $N = 8$.

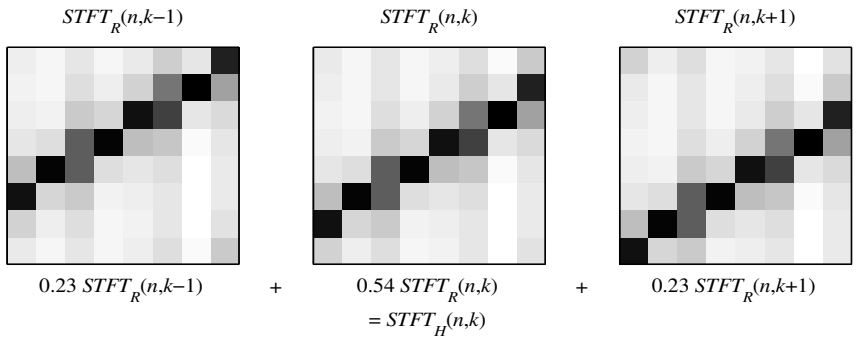


Figure 10.12 The STFT of a linear FM signal calculated using a rectangular window (from the previous figure), along with its frequency shifted versions $STFT_R(n,k-1)$ and $STFT_R(n,k+1)$. Their weighted sum produces the STFT of the same signal with a Hamming window $STFT_H(n,k)$.

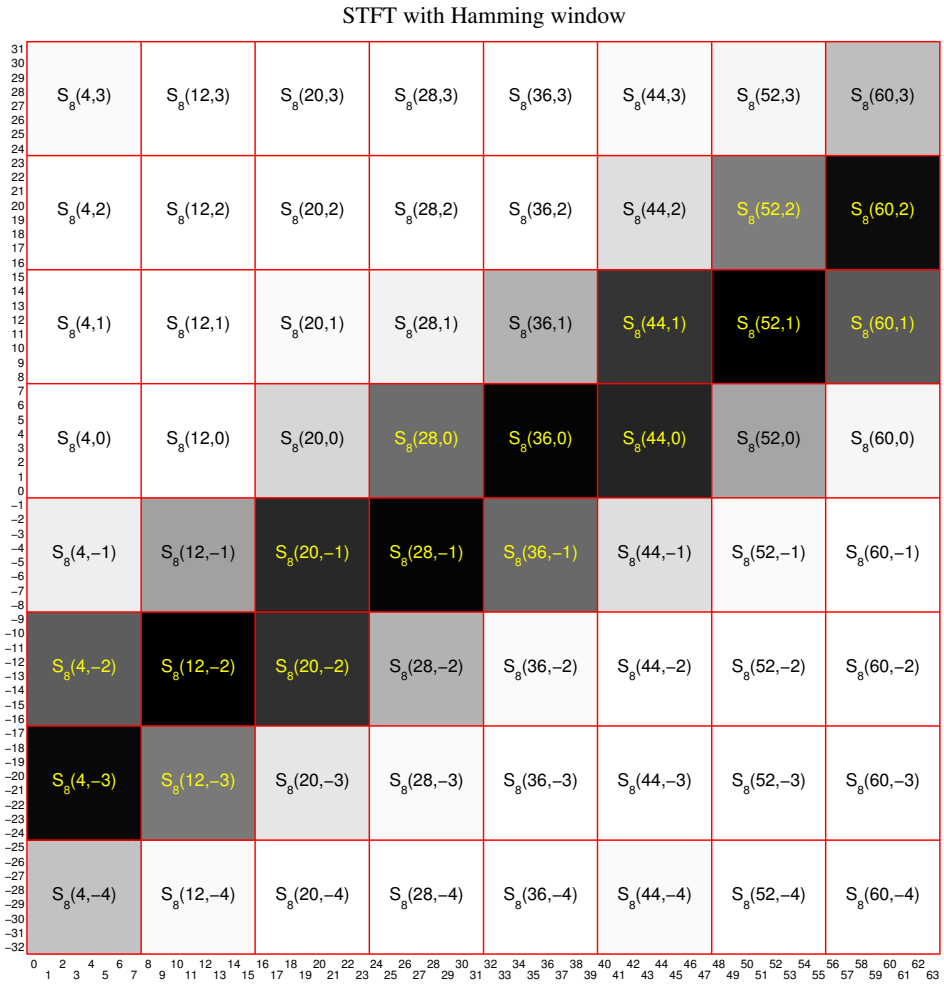


Figure 10.13 The STFT of a linear FM signal $x(n)$ calculated using the Hamming window with $N = 8$. Calculation is illustrated in the previous figure.

10.5 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.10.16. Windowed signal values are reconstructed from the STFTs by a simple inversion of each STFT

$$\mathbf{STFT}(n) = \mathbf{W}_N \mathbf{H}_w \mathbf{x}(n)$$
$$\mathbf{H}_w \mathbf{x}(n) = \text{IDFT}\{\mathbf{STFT}(n)\} = \mathbf{W}_N^{-1} \mathbf{STFT}(n)$$

where $\mathbf{H}_w$ is a diagonal matrix with the window values as its elements, $\mathbf{H}_w = \text{diag}(w(m))$.

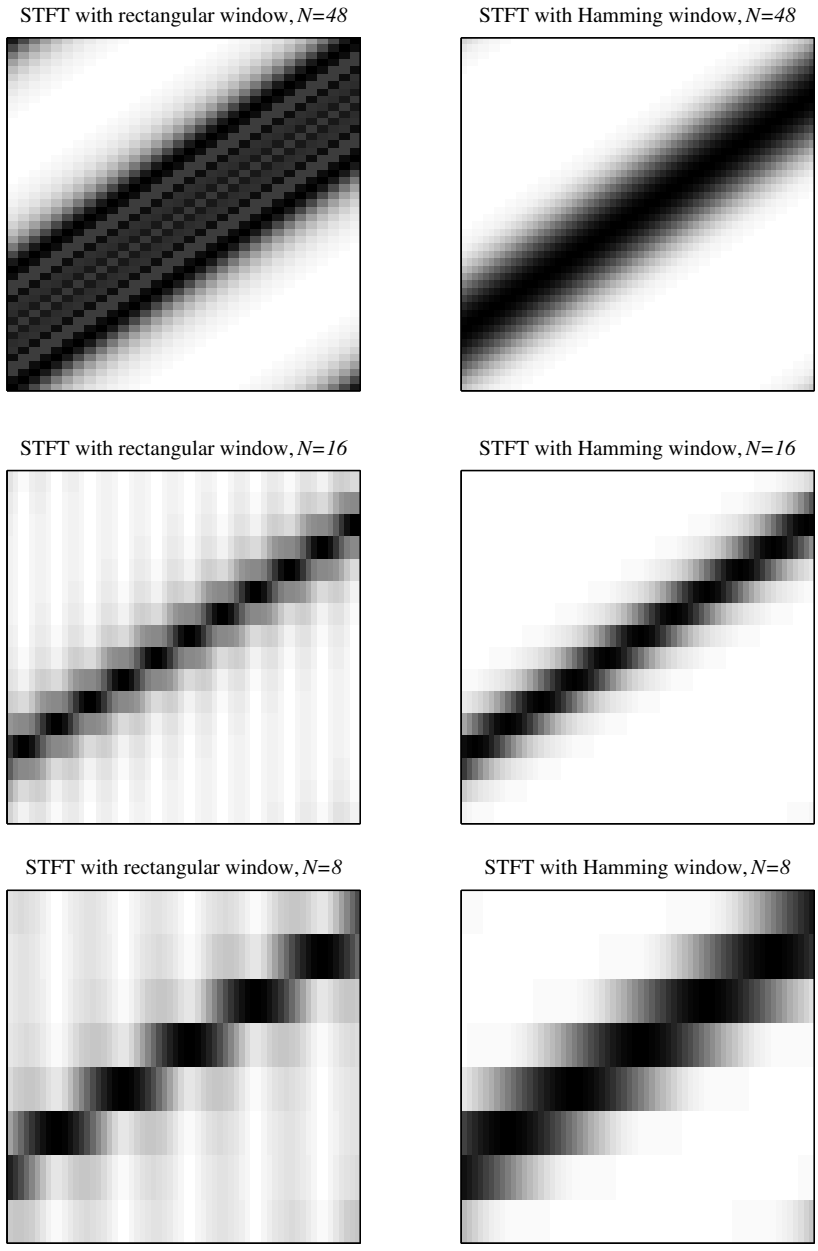


Figure 10.14 Time-frequency analysis of a linear frequency modulated signal with overlapping windows of various widths. Time step in the STFT calculation is $R = 1$.

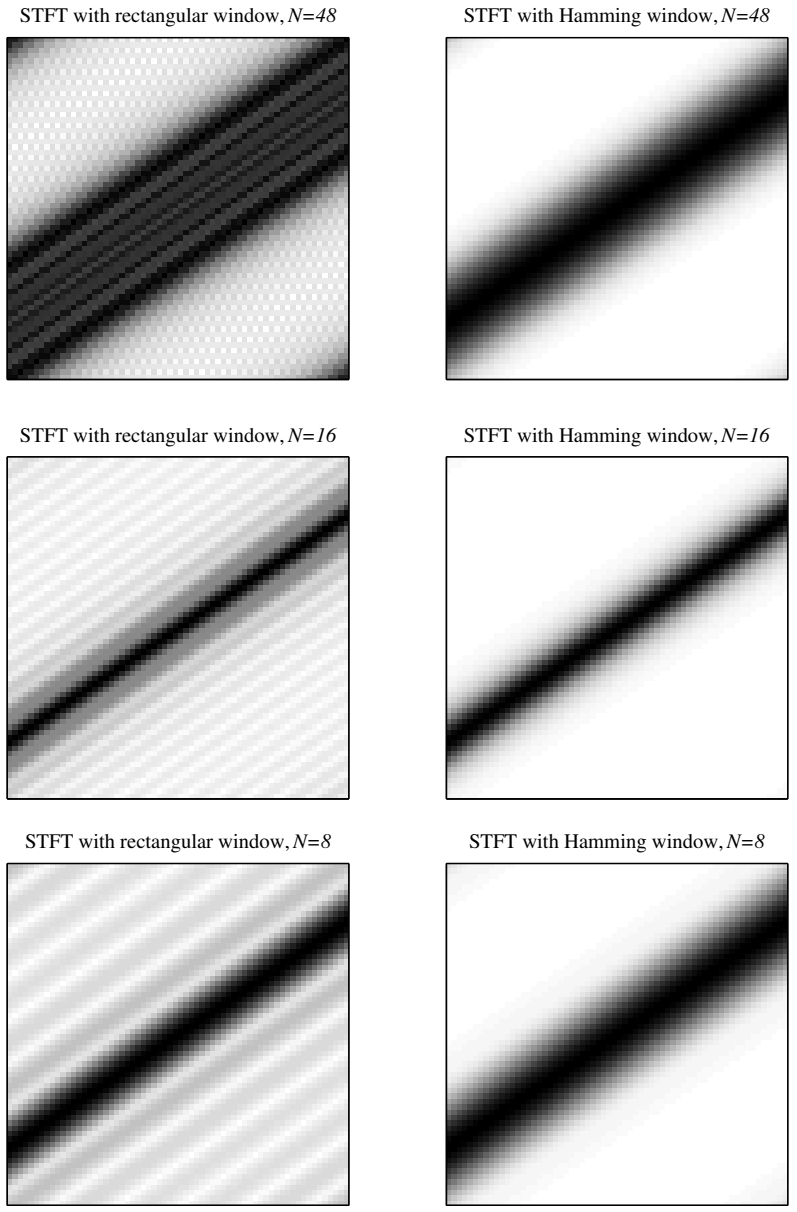


Figure 10.15 Time-frequency analysis of a linear frequency modulated signal with overlapping windows of various widths. Time step in the STFT calculation is $R = 1$. For each window width the frequency axis is interpolated (signal in time is zero padded) up to the total number of available signal samples $M = 64$.

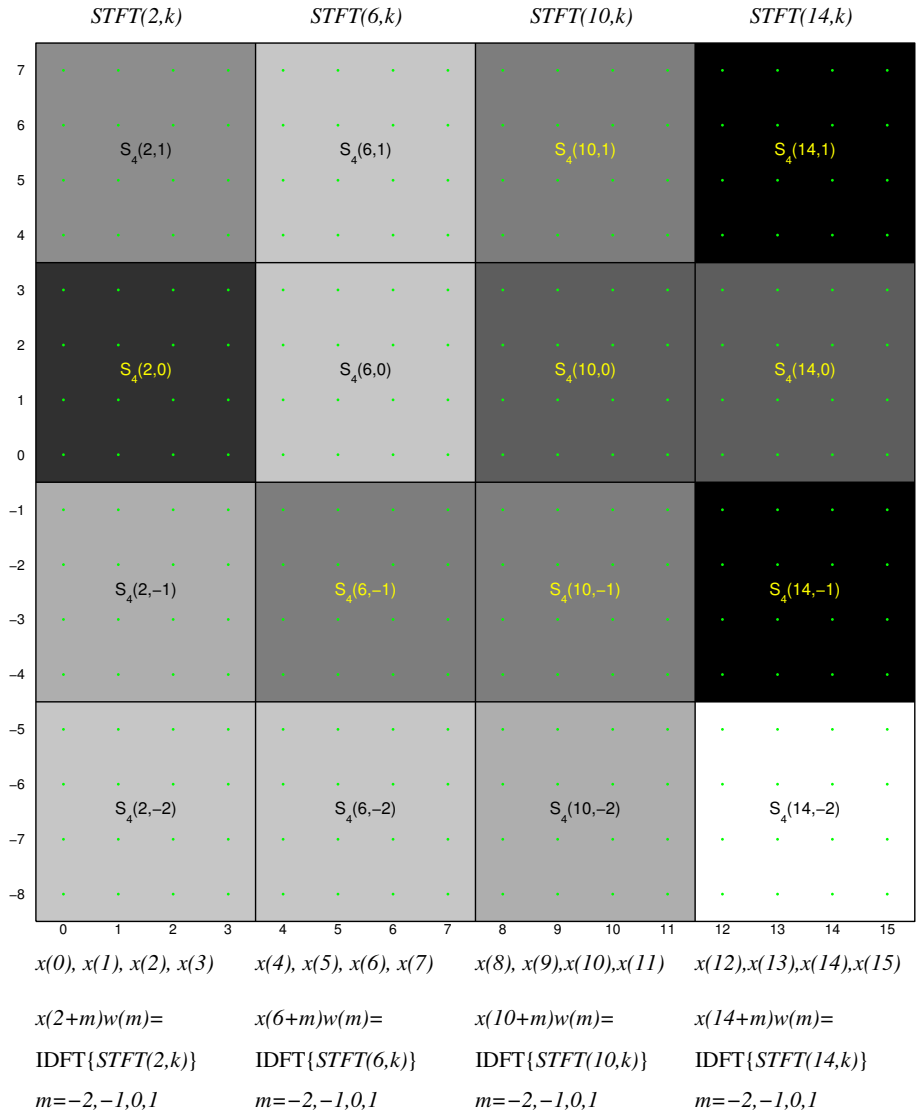


Figure 10.16 Illustration of the signal reconstruction from the STFT with nonoverlapping windows.

Example 10.10. Consider a signal with $M = 16$ samples, $x(0), x(1), \ldots, x(15)$. Write a matrix form for the signal inversion using a four-sample STFT ($N = 4$) calculated with the rectangular and a Hann(ing) window: (a) Without overlapping, $R = 4$. (b) With a time step in the STFT calculation of $R = 2$.

★ (a) For the nonoverlapping case the STFT calculation is done according to:

$$\mathbf{STFT} = \mathbf{W_4 H_4} \begin{bmatrix} x(0)\, x(4) & x(8) & x(12) \\ x(1)\, x(5) & x(9) & x(13) \\ x(2)\, x(6) & x(10) & x(14) \\ x(3)\, x(7) & x(11) & x(15) \end{bmatrix}.$$

with $\mathbf{H_4}$ =diag($[w(-2)\ w(-1)\ w(0)\ w(1)]$) and $\mathbf{W}_4$ is the corresponding four sample DFT matrix.

The inversion relation is

$$\begin{bmatrix} x(0)\, x(4) & x(8) & x(12) \\ x(1)\, x(5) & x(9) & x(13) \\ x(2)\, x(6) & x(10) & x(14) \\ x(3)\, x(7) & x(11) & x(15) \end{bmatrix} = \mathbf{H_4^{-1} W_4^{-1} STFT}$$

where the elements of diagonal matrix $\mathbf{H_4^{-1}}$ are proportional to $1/w(m)$, $\mathbf{H_4^{-1}}$ =diag($[1/w(-2)$ $1/w(-1)\ 1/w(0)\ 1/w(1)]$). If a rectangular window is used in the STFT calculation then $\mathbf{H_4^{-1}} = \mathbf{I}_4$ is unity matrix and this kind of 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 zero-valued, so $1/w(m)$ does not exist.

(b) The STFT calculation is done with overlapping with step $R = 2$, Fig.10.17. For $N = 4$ and calculation step $R = 2$ the STFT calculation corresponds to

$$\mathbf{STFT} = \mathbf{W_4 H_4} \begin{bmatrix} 0 & x(0) & x(2) & x(4) & x(6) & x(8) & x(10) & x(12) & x(14) \\ 0 & x(1) & x(3) & x(5) & x(7) & x(9) & x(11) & x(13) & x(15) \\ x(0) & x(2) & x(4) & x(6) & x(8) & x(10) & x(12) & x(14) & 0 \\ x(1) & x(3) & x(5) & x(7) & x(9) & x(11) & x(13) & x(15) & 0 \end{bmatrix}$$

The inversion is

$$\mathbf{W}_4^{-1}\mathbf{STFT} = \mathbf{H_4 X} =$$
$$\begin{bmatrix} 0 & x(0)w(-2) & x(2)w(-2) & x(4)w(-2) & \dots & x(14)w(-2) \\ 0 & x(1)w(-1) & x(3)w(-1) & x(5)w(-1) & \dots & x(15)w(-1) \\ x(0)w(0) & x(2)w(0) & x(4)w(0) & x(6)w(0) & \dots & 0 \\ x(1)w(1) & x(3)w(1) & x(5)w(1) & x(7)w(1) & \dots & 0 \end{bmatrix},$$

where $\mathbf{X}$ 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 $\mathbf{W}_4^{-1}\mathbf{STFT}$ 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,

$$\begin{aligned} x(0)w(0) + x(0)w(-2) &= x(0) \\ x(1)w(1) + x(1)w(-1) &= x(1) \\ x(2)w(0) + x(2)w(-2) &= x(2) \\ &\vdots \\ x(15)w(1) + x(15)w(-1) &= x(15). \end{aligned}$$

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.10.17. Its inversion produces $x(n_0+m)w(m)=\text{IDFT}\{STFT_N(n_0,k)\}$. Consider the previous STFT value in the case of nonoverlapping windows. It would be $STFT_N(n_0-N,k)$. Its inverse

$$\text{IDFT}\{STFT_N(n_0-N,k)\}=x(n_0-N+m)w(m)$$

is also presented in Fig.10.17. As it can be seen, by combining these two inverse transforms we would get signal with very low values around $n=n_0-N/\dot{2}$. 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.10.17, 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

$$\begin{array}{c}\cdots\\ \mathbf{STFT}(n_0-2R),\\ \mathbf{STFT}(n_0-R),\\ \mathbf{STFT}(n_0),\\ \mathbf{STFT}(n_0+R),\\ \mathbf{STFT}(n_0+2R),\\ \cdots\end{array} \tag{10.33}$$

Based on the available STFT values (10.33), the windowed signal values can be reconstructed as

$$\mathbf{H}_w\mathbf{x}(n_0+iR)=\mathbf{W}_N^{-1}\mathbf{STFT}(n_0+iR),\quad i=\cdots-2,-1,0,1,2,\ldots$$

For $m=-N/2,-N/2+1,\ldots,N/2-1$ we get the signal values $x(n_0+iR+m)$

$$w(m)x(n_0+iR+m)=\frac{1}{N}\sum_{k=-N/2}^{N/2-1}STFT(n_0+iR,k)e^{j2\pi mk/N}. \tag{10.34}$$

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 re-index the reconstructed signal values (10.34) by substitution $m=l-iR$, as in (10.12),

$$w(l-iR)x(n_0+l)=\frac{1}{N}\sum_{k=-N/2}^{N/2-1}STFT(n_0+iR,k)e^{j2\pi lk/N}e^{-j2\pi iRk/N}$$
$$-N/2\leq l-iR\leq N/2-1.$$

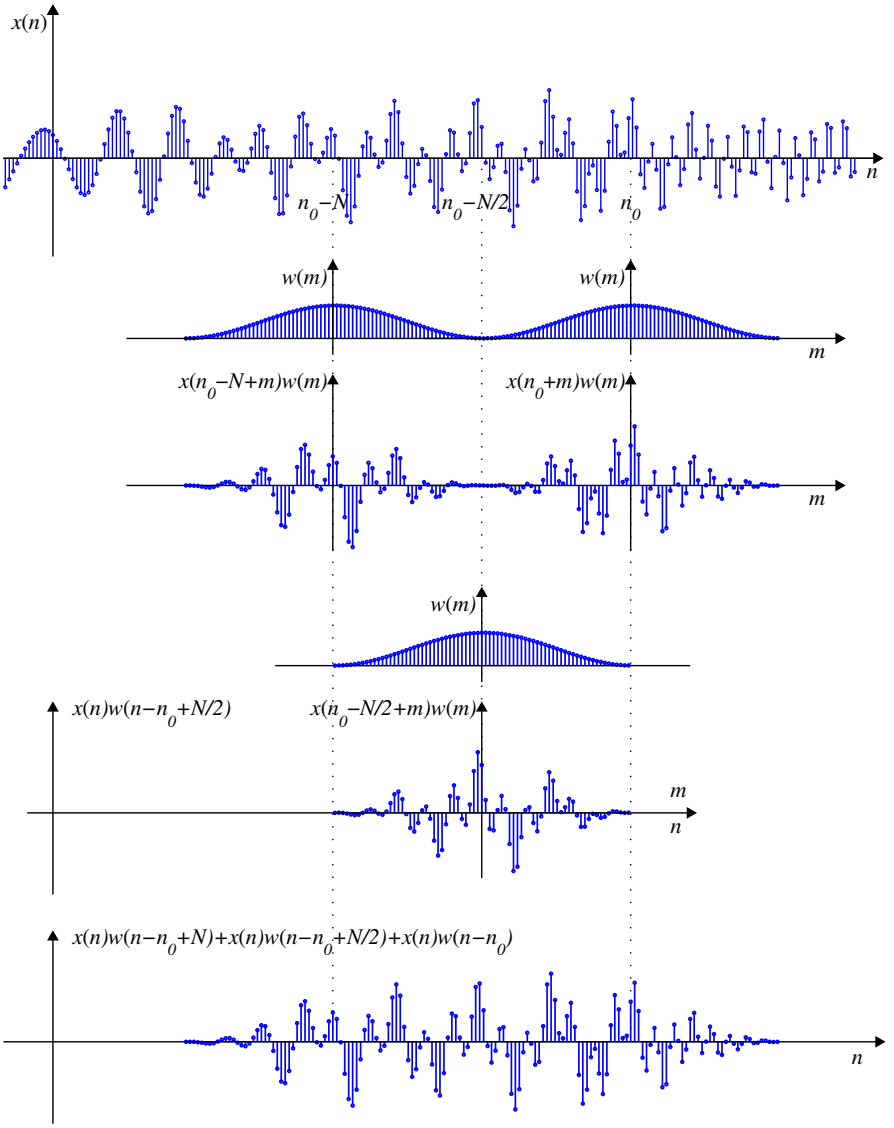


Figure 10.17 Illustration of the STFT calculation with windows overlapping in order to produce an inverse STFT whose sum will give the original signal within $n_0 - N \leq n < n_0$.

If $R < N$, then a value of the signal $x(n_0 + l)$ will be obtained by inverting the STFT

$$w(l)x(n_0 + l) = \frac{1}{N} \sum_{k=-N/2}^{N/2-1} STFT(n_0,k) e^{j2\pi lk/N}$$

The same signal value, $x(n_0+l)$, will be obtained within the other overlapping inversions

$$\begin{aligned}
&\vdots\\
w(l-2R)x(n_0+l) &= \frac{1}{N}\sum_{k=-N/2}^{N/2-1} STFT(n_0+2R,k)e^{j2\pi lk/N}e^{-j2\pi 2Rk/N}\\
w(l-R)x(n_0+l) &= \frac{1}{N}\sum_{k=-N/2}^{N/2-1} STFT(n_0+R,k)e^{j2\pi lk/N}e^{-j2\pi Rk/N}\\
w(l+R)x(n_0+l) &= \frac{1}{N}\sum_{k=-N/2}^{N/2-1} STFT(n_0-R,k)e^{j2\pi lk/N}e^{j2\pi Rk/N}\\
w(l+2R)x(n_0+l) &= \frac{1}{N}\sum_{k=-N/2}^{N/2-1} STFT(n_0-2R,k)e^{j2\pi lk/N}e^{j2\pi 2Rk/N}\\
&\vdots
\end{aligned}$$

as far as $w(l-2iR)$, for $i=0,\pm1,\pm2,\dots$ is within

$$-N/2 \le l-2iR < N/2.$$

By summing all the reconstructions over i satisfying $-N/2 \le l-iR \le N/2-1$ we get the final reconstructed signal $x(n_0+l)$. Obviously, this sum produces the exact, up to a constant undistorted signal value, if

$$\sum_i w(l-iR) = 1 \tag{10.35}$$

or

$$c(l) = \sum_i w(l-iR) = \text{const.} = \mathrm{C} \tag{10.36}$$

since

$$\sum_i w(l-iR)x(n_0+l) = \mathrm{C}x(n_0+l)$$

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 (10.36) is equivalent to

$$W(e^{j2\pi k/R}) = \mathrm{C}N\delta(k) \text{ for } k = 0,1,\dots,R-1.$$

Special cases:

1. For $R=N$ (nonoverlapping), relation (10.36) is satisfied for the rectangular window, only.

2. For a half of the overlapping period, $R=N/2$, condition (10.36) 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.10.18. 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.

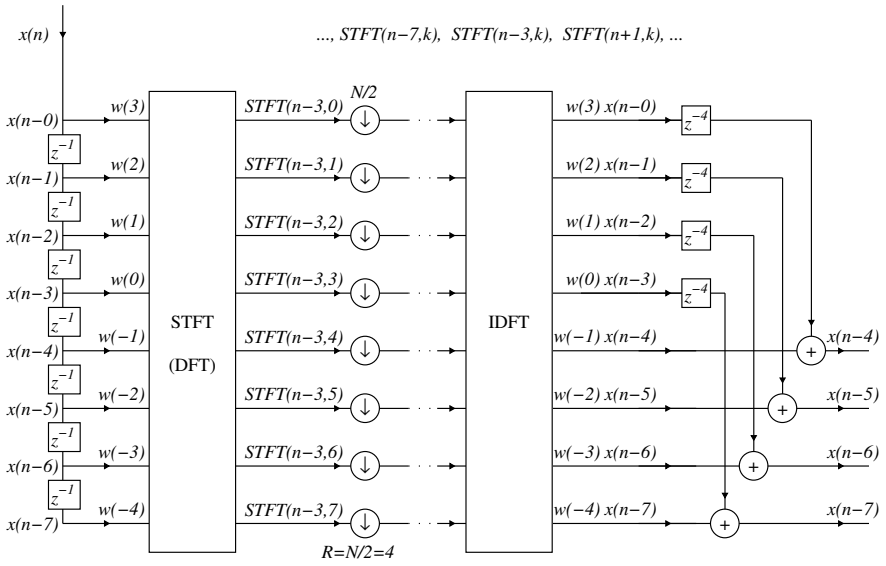


Figure 10.18 Signal reconstruction from the STFT for the case $N = 8$, when the STFT is calculated with step $R = N/2 = 4$ and the window satisfies $w(m) + w(m - N/2) = 1$. This is the case for the rectangular, Hann(ing), Blackman and triangular windows. The same holds for the Hamming window up to a constant scaling factor of 1.08.

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.10.19

$$\frac{1}{N}\sum_{k=-N/2}^{N/2-1} STFT(n,k) = \frac{1}{N}\sum_{m=-N/2}^{N/2-1}\left(w(m)x(n+m)\sum_{k=-N/2}^{N/2-1} e^{-j2\pi mk/N}\right)$$
$$= w(0)x(n).$$

Very efficient realizations, for this case, are the recursive ones, instead of the direct DFT calculation, Fig.10.9.

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.

If the reconstructed signal is weighted by the same analysis window, that is $w(l - iR)x(n_0 + l)$ is multiplied by $w(l - iR)$, then the reconstruction condition for the weighted overlap-add method is

$$\sum_i w^2(l - iR) = 1. \tag{10.37}$$

For more details on this from and its kernel framework interpretation see Section 10.2.

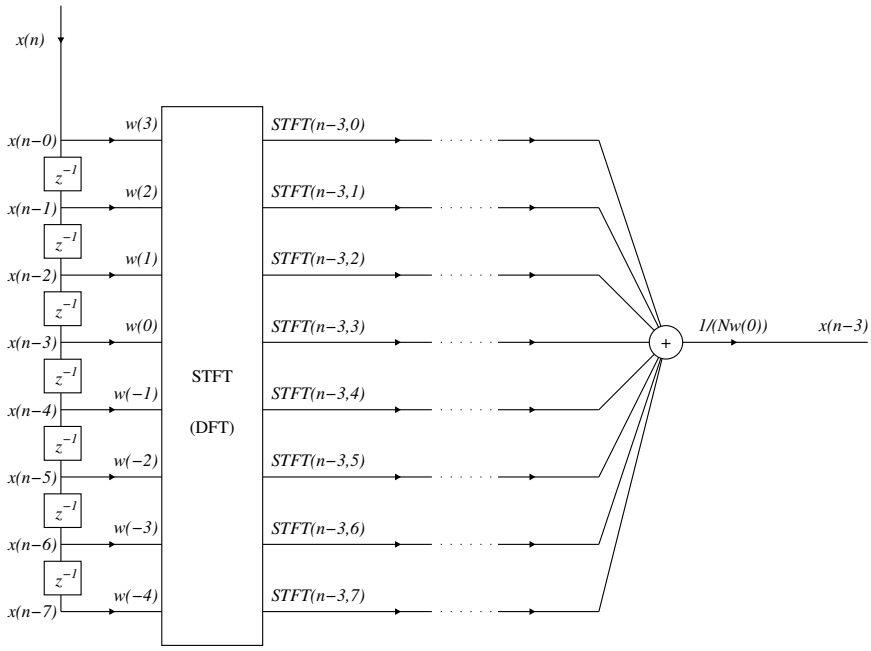


Figure 10.19 Signal reconstruction when the STFT is calculated with step $R = 1$.

10.6 VARYING WINDOWS IN THE STFT

Window with and form can be varying for different time instants, frequency bands, or can be time-frequency varying. These forms of the windows will be presented next.

10.6.1 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,

$$STFT_{N_i}(n_i,k) = \sum_{m=-N_i/2}^{N_i/2-1} x(n_i+m)e^{-j\frac{2\pi}{N_i}mk} \tag{10.38}$$

Notation $STFT_{N_i}(n_i,k)$ 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 π. 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

$$\mathbf{STFT}_{N_i}(n_i) = \mathbf{W}_{N_i}\mathbf{x}_{N_i}(n_i),$$

where $\mathbf{STFT}_{N_i}(n_i)$ and $\mathbf{x}_{N_i}(n_i)$ are column vectors. Their elements are $STFT_{N_i}(n_i,k)$, $k = -N_i/2,\ldots, N_i/2-1$ and $x(n_i+m)$, $m = -N_i/2,dots, N_i/2-1$, respectively

$$\mathbf{STFT}_{N_i}(n_i) = [STFT_{N_i}(n_i,-N_i/2)\ \ldots STFT_{N_i}(n_i,N_i/2-1)]^T$$
$$\mathbf{x}_{N_i}(n_i) = [x(n_i-N_i/2)\ \ldots x(n_i+N_i/2-1)]^T.$$

Matrix $\mathbf{W}_{N_i}$ 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 $p = -M/2,\ldots,M/2-1$ corresponding to the DFT frequencies $2\pi p/M$ (dots in Fig.10.20). With respect to this axis labeling, the block $STFT_{N_i}(n_i,k)$ will be positioned at the frequency $2\pi k/N_i = 2\pi(kM/N_i)/M$, that is, 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.10.20 where $M = 16$.

Example 10.11. 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:

$$STFT_4(2,1) = \sum_{m=-2}^{1} x(2+m)e^{-j\frac{2\pi}{4}m}.$$

It uses discrete-time samples of $x(n)$ within

$$-2 \le 2+m < 1$$
$$0 \le n \le 3.$$

The frequency term is $exp(-j2\pi m/4)$. For the DFT of a signal with $M = 16$

$$X(k) = \sum_{m=0}^{15} x(m)e^{-j\frac{2\pi}{16}mk}$$
$$k = -8,-7,\cdots -1,0,1,\ldots,6,7$$

this frequency would correspond to the term $exp(-j2\pi 4m/16)$. Therefore $k = 1$ corresponds to the frequency $p = 4$ in the DFT. Since the whole frequency range $-\pi \le \omega < \pi$ 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

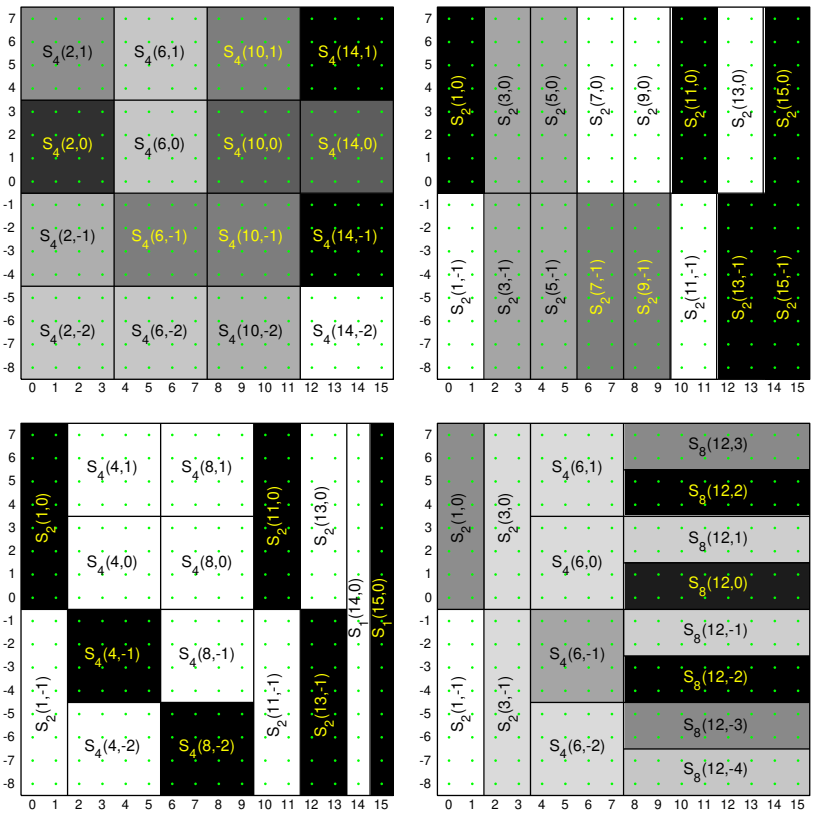


Figure 10.20 The nonoverlapping STFTs with: (a) constant window of the width $N = 4$, (b) constant window of the width $N = 2$, (c)-(d) time-varying windows. Time index is presented on the horizontal axis, while the DFT frequency index is shown on the vertical axis (the $STFT$ is denoted by S for notation simplicity).

4 DFT values,

$$k = -2, \text{ corresponds to } p = -8,-7,-6,-5$$
$$k = -1, \text{ corresponds to } p = -4,-3,-2,-1$$
$$k = 0, \text{ corresponds to } p = 0,1,2,3$$
$$k = 1, \text{ corresponds to } p = 4,5,6,7.$$

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.10.20(a).

In a nonoverlapping STFT, covering all signal samples

$$\mathbf{x} = [x(0), x(1), \ldots, x(M-1)]^T$$

with $\mathbf{STFT}_{N_i}(n_i)$, the STFT should be calculated at $n_0 = N_0/2$, $n_1 = N_0 + N_1/2$, $n_2 = N_0 + N_1 + N_2/2, \ldots, n_K = M - N_K/2$. A matrix form for all STFT values is

$$\mathbf{STFT} = \begin{bmatrix} \mathbf{W}_{N_0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_{N_1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{W}_{N_K} \end{bmatrix} \mathbf{x}$$

$$\mathbf{STFT} = \tilde{\mathbf{W}}\mathbf{x} = \tilde{\mathbf{W}}\mathbf{W}_M^{-1}\mathbf{X}, \tag{10.39}$$

where $\mathbf{STFT}$ is a column vector containing all STFT vectors $\mathbf{STFT}_{N_i}(n_i)$, $i = 0,1,\ldots,K$, $\mathbf{X} = \mathbf{W}_M\mathbf{x}$ is a DFT of the whole signal $x(n)$, while $\tilde{\mathbf{W}}$ is a block matrix ($M \times M$) formed from the smaller DFT matrices $\mathbf{W}_{N_0}$, $\mathbf{W}_{N_1}$, ... ,$\mathbf{W}_{N_K}$, as in (10.38). 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.10.20(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 10.12. Consider a signal $x(n)$ with $M = 16$ samples, whose values are $x = [0.5,\ 0.5,\ -0.25,\ j0.25,\ 0.25,\ -j0.25,\ -0.25,\ 0.25,\ -0.25,\ 0.25,\ 0.5,\ 0.5,\ -j0.5,\ j0.5,\ 0,\ -1]$. Some of its nonoverlapping STFTs are calculated according to (10.38) and shown in Fig.10.20. Different representations can be compared based on the concentration measures, for example,

$$\mu[STFT_N(n,k)] = \sum_n \sum_k |STFT_N(n,k)| = \|\mathbf{STFT}\|_1 .$$

The best STFT representation, in this sense, would be the one with the smallest $\mu[STFT_N(n,k)]$. For the considered signal and its four representations shown in Fig.10.20 the best representation, according to this criterion, is the one shown in Fig.10.20(b).

Example 10.13. 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

$$STFT_N(n,k) = \sum_{m=-N/2}^{N/2-1} x(n+m)e^{-j2\pi mk/N}.$$

For an odd N, the summation limits are $\pm(N-1)/2$. Calculate $STFT_1(n,k)$ for $n = 0,1,2,3,4,5,6,7$, then $STFT_2(n,k)$ for $n = 1,3,5,7$, then $STFT_4(n,k)$ for $n = 2,6$ and $STFT_8(n,k)$ 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 of $\mu[STFT(n,k)]^{1/2}$ 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 $\mu[STFT(n,k)]^{1/2}$?

★ (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 $N = 2$

$$\begin{aligned}
&STFT_2(n,0) = x(n) + x(n-1)\\
&STFT_2(1,0) = 1,\\
&STFT_2(3,0) = 0,\\
&STFT_2(5,0) = (1-j)/4,\\
&STFT_2(7,0) = (-1+j)/4
\end{aligned}$$

$$\begin{aligned}
&STFT_2(n,1) = x(n) - x(n-1)\\
&STFT_2(1,1) = 1,\\
&STFT_2(3,1) = -1,\\
&STFT_2(5,1) = (-1-j)/4,\\
&STFT_2(7,1) = (1+j)/4
\end{aligned}$$

– for $N = 4$ and $n = 2$, 6

$$\begin{aligned}
&STFT_4(n,0) = x(n-2) + x(n-1) + x(n) + x(n+1)\\
&STFT_4(2,0) = 1,\\
&STFT_4(6,0) = 0
\end{aligned}$$

$$\begin{aligned}
&STFT_4(n,1) = -x(n-2) + jx(n-1) + x(n) - jx(n+1)\\
&STFT_4(2,1) = (1+3j)/2,\\
&STFT_4(6,1) = 0
\end{aligned}$$

$$\begin{aligned}
&STFT_4(n,2) = x(n-2) - x(n-1) + x(n) - x(n+1)\\
&STFT_4(2,2) = 0,\\
&STFT_4(6,2) = 0,
\end{aligned}$$

$$\begin{aligned}
&STFT_4(n,3) = -x(n-2) - jx(n-1) + x(n) + jx(n+1)\\
&STFT_4(2,3) = (1-3j)/2,\\
&STFT_4(6,3) = -1
\end{aligned}$$

(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. 10.21, along with measure

$$\mu[STFT(n,k)] = \sum_n \sum_k |STFT(n,k)|^{1/2}$$

for each case.

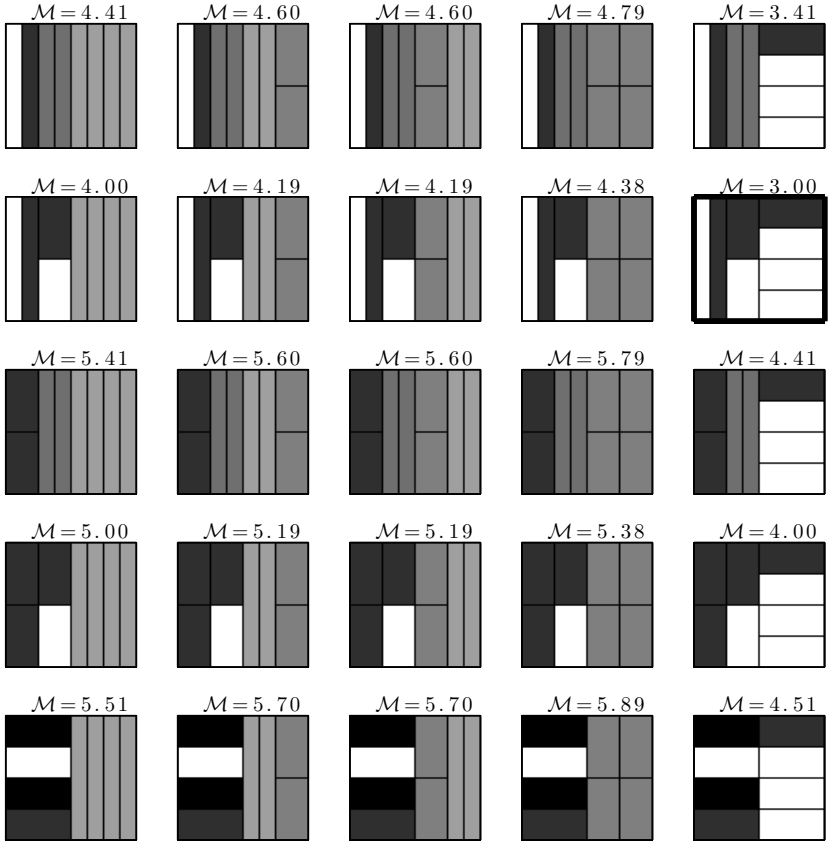


Figure 10.21 Time-frequency representation in various lattices (grid-lines are shown), with concentration measure $\mathcal{M} = \mu[SPEC(n,k)]^{1/2}$ value. The optimal representation, with respect to this measure, is presented with thicker gridlines. Time axis is $n = 0,1,2,3,4,5,6,7$ and the frequency axis is $k = 0,1,2,3,4,5,6,7$.

(c) By measuring the concentration for all of them, we will get that the optimal combination, to cover the time-frequency plane, is

$$
\begin{aligned}
&STFT_1(0,0) = x(0) = 0\\
&STFT_1(1,0) = x(1) = 1\\
&STFT_2(3,1) = x(3) - x(2) = -1\\
&STFT_2(3,0) = x(3) + x(2) = 0\\
&STFT_4(6,0) = x(4) + x(5) + x(6) + x(7) = 0\\
&STFT_4(6,1) = -x(4) + jx(5) + x(6) - jx(7) = 0\\
&STFT_4(6,2) = x(4) - x(5) + x(6) - x(7) = 0\\
&STFT_4(6,3) = -x(4) - jx(5) + x(6) + jx(7) = -1
\end{aligned}
$$

with just three nonzero transformation coefficients. It corresponds to the minimum of $\mu[SPEC(n,k)]$.

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. 10.21, corresponding to the constant window widths of $N = 1$, $N = 2$, and $N = 4$ samples.

Example 10.14. Discrete signal $x(n)$ for $n = 0,1,2,3,4,5$ is considered. Time-frequency plane is divided as presented in Fig. 10.22.

(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 $\mathbf{T}$.

(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

$$STFT_2(1,0) = 4, \quad STFT_2(1,1) = 0$$
$$STFT_1(2,0) = 1, \quad STFT_3(4,0) = 0$$
$$STFT_3(4,1) = 3, \quad STFT_3(4,2) = 3$$

find the signal samples $y(n)$.

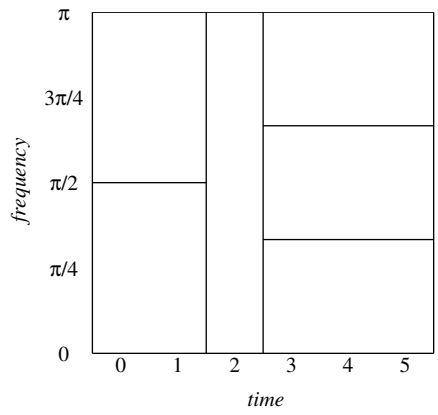


Figure 10.22 Areas in the time-frequency plane.

★ (a) Denoted areas are presented in Fig. 10.23.

(b) The STFT values are obtained using

$$STFT_N(n,k) = \sum_{m=-(N-1)/2}^{(N-1)/2-1} x(n+m)e^{-j2\pi mk/N} \quad \text{or}$$

$$STFT_N(n,k) = \sum_{m=-N/2}^{N/2-1} x(n+m)e^{-j2\pi mk/N}$$

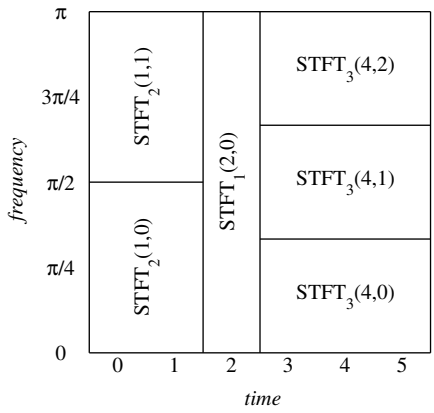


Figure 10.23 Denoted areas in the time-frequency plane.

for and odd and even number of samples N, respectively. The elements are

$$
\begin{aligned}
&STFT_2(1,0) = x(0) + x(1)\\
&STFT_2(1,1) = -x(0) + x(1)\\
&STFT_1(2,0) = x(2)\\
&STFT_3(4,0) = x(3) + x(4) + x(5)\\
&STFT_3(4,1) = \frac{-1+j\sqrt{3}}{2}x(3) + x(4) + \frac{-1-j\sqrt{3}}{2}x(5)\\
&STFT_3(4,2) = \frac{-1-j\sqrt{3}}{2}x(3) + x(4) + \frac{-1+j\sqrt{3}}{2}x(5).
\end{aligned}
$$

The transformation matrix (where the STFT elements are arranged into column vector $\mathbf{S}$) is

$$
\mathbf{T} = \begin{bmatrix}
1 & 1 & 0 & 0 & 0 & 0\\
-1 & 1 & 0 & 0 & 0 & 0\\
0 & 0 & 1 & 0 & 0 & 0\\
0 & 0 & 0 & 1 & 1 & 1\\
0 & 0 & 0 & \frac{-1+j\sqrt{3}}{2} & 1 & \frac{-1-j\sqrt{3}}{2}\\
0 & 0 & 0 & \frac{-1-j\sqrt{3}}{2} & 1 & \frac{-1+j\sqrt{3}}{2}
\end{bmatrix}.
$$

(c) The STFT elements are

$$
\mathbf{S} = \begin{bmatrix}
1 & 1 & 0 & 0 & 0 & 0\\
-1 & 1 & 0 & 0 & 0 & 0\\
0 & 0 & 1 & 0 & 0 & 0\\
0 & 0 & 0 & 1 & 1 & 1\\
0 & 0 & 0 & \frac{-1+j\sqrt{3}}{2} & 1 & \frac{-1-j\sqrt{3}}{2}\\
0 & 0 & 0 & \frac{-1-j\sqrt{3}}{2} & 1 & \frac{-1+j\sqrt{3}}{2}
\end{bmatrix}
\begin{bmatrix}
2\\ -2\\ 4\\ \sqrt{3}\\ -\sqrt{3}\\ 0
\end{bmatrix}
=
\begin{bmatrix}
0\\ -4\\ 4\\ 0\\ -\frac{3+j3\sqrt{3}}{2}\\ \frac{3-j3\sqrt{3}}{2}
\end{bmatrix}.
$$

(d) The signal samples $y(n)$ are obtained as $\mathbf{T}^{-1}\mathbf{S}$ resulting in

$$\left[y(5)\ y(4)\ y(3)\ y(2)\ y(1)\ y(0)\right]^T = \left[2\ \ 2\ \ 1\ \ -1\ \ \ 2\ \ -1\right]^T.$$

Example 10.15. 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

$$F(M) = F(M-2^0) + F(M-2^1) + F(M-2^2) + \cdots + F(M-2^m) = \sum_{m=0}^{\lfloor \log_2 M \rfloor} F(M-2^m),$$

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

$$F(2) = F(1) + F(0) = 2,$$
$$F(3) = F(2) + F(1) = 3.$$
$$F(4) = F(3) + F(2) + F(0) = 6, \quad \ldots$$

The results are presented in the table

M	1	2	3	4	5	6	7	8
$F(M)$	1	2	3	6	10	18	31	56

M	9	10	11	12	13	14	15	16
$F(M)$	98	174	306	542	956	1690	2983	5272

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.

10.6.1.1 Time-Varying Hann(ing) Windows

The continuous time-varying Hann(ing) window, positioned at $\tau = b_k$, can be defined by

$$w_k(\tau - b_k) = \begin{cases} \sin^2\left(\frac{\pi}{2}\frac{a_k}{b_k - a_k}\left(\frac{\tau}{a_k} - 1\right)\right), & \text{for } a_k < \tau \leq b_k \\ \cos^2\left(\frac{\pi}{2}\frac{b_k}{c_k - b_k}\left(\frac{\tau}{b_k} - 1\right)\right), & \text{for } b_k < \tau \leq c_k \\ 0, & \text{elsewhere,} \end{cases} \tag{10.40}$$

where $(a_k, b_k]$ and $(b_k, c_k]$, define the width of $w_k(\tau - b_k)$. For consecutive intervals the relations $a_{k+1} = b_k$, $b_{k+1} = c_k$ hold, as shown in Fig. 10.24. These windows satisfy the *constant overlap-add* relation

$$\sum_{k=0}^{K-1} w(\tau - b_k) = 1, \tag{10.41}$$

since the squared sine and cosine sum up in two consecutive windows with the same parameters.

The initial window function, $w_0(\tau - b_0)$, is defined

$$w_0(\tau - b_0) = \begin{cases} 1, & \text{for } a_0 = 0 < \tau \leq b_0 \\ \cos^2\left(\frac{\pi}{2}\frac{b_0}{c_0 - b_0}\left(\frac{\tau}{b_0} - 1\right)\right), & \text{for } b_0 < \tau \leq c_0 \\ 0, & \text{elsewhere,} \end{cases} \tag{10.42}$$

while the last window function, $w_{K-1}(\tau - b_{K-1})$, is defined

$$w_{K-1}(\tau - b_{K-1}) = \begin{cases} \sin^2\left(\frac{\pi}{2}\frac{a_{K-1}}{b_{K-1} - a_{K-1}}\left(\frac{\tau}{a_{K-1}} - 1\right)\right), & \text{for } a_{K-1} < \tau \leq b_{K-1} = t_{\max} \\ 0, & \text{elsewhere,} \end{cases} \tag{10.43}$$

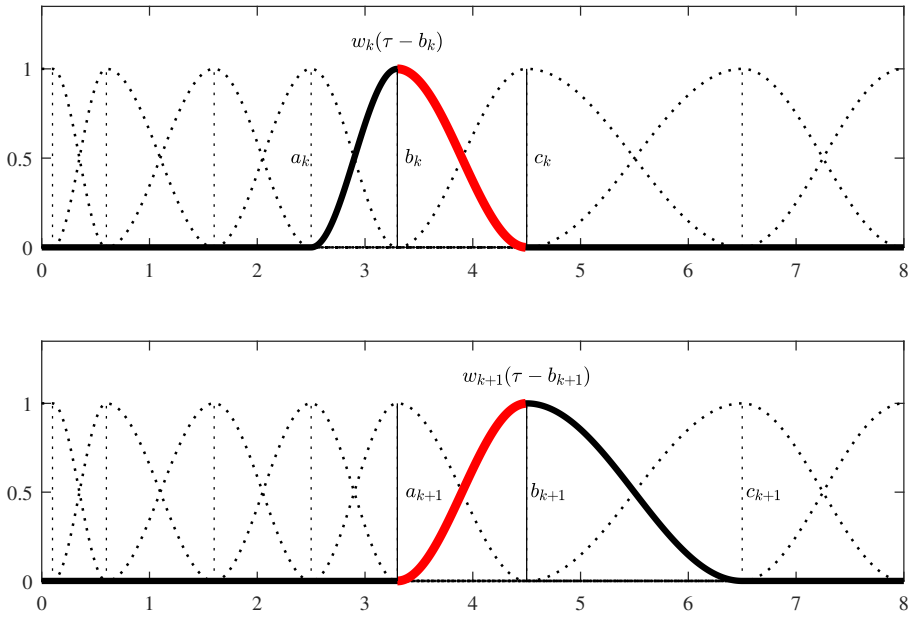


Figure 10.24 (a) Time-varying asymmetric Hann(ing) windows.

An example of eight time-varying Hann(ing) windows is shown in Fig. 10.25.

For uniform widths, as in Fig. 10.6(a), the intervals can be defined by

$$a_k = a_{k-1} + \frac{t_{\max}}{K-1}, \quad b_k = a_k + \frac{t_{\max}}{K-1}, \quad c_k = a_k + 2\frac{t_{\max}}{K-1} \tag{10.44}$$

with $a_1 = 0$ and $\lim_{\tau\to 0}(a_1/\tau) = 1$.

10.6.1.2 Time-Varying Sine Windows

For the *weighted overlap-add* method, the reconstruction condition is

$$\sum_{k=0}^{K-1} w^2(\tau - b_k) = 1. \tag{10.45}$$

A simple way to construct a window (function) for the weighted overlap-add method is to take the square root of the constant overlap-add window (for, example, the *sine window* as the square root of the Hann window). In this case, the window functions become

$$w_k(\tau - b_k) = \begin{cases} \sin\left(\frac{\pi}{2}\frac{a_k}{b_k - a_k}\left(\frac{\tau}{a_k} - 1\right)\right), & \text{for } a_k < \tau \le b_k \\ \cos\left(\frac{\pi}{2}\frac{b_k}{c_k - b_k}\left(\frac{\tau}{b_k} - 1\right)\right), & \text{for } b_k < \tau \le c_k \\ 0, \quad \text{elsewhere,} \end{cases} \tag{10.46}$$

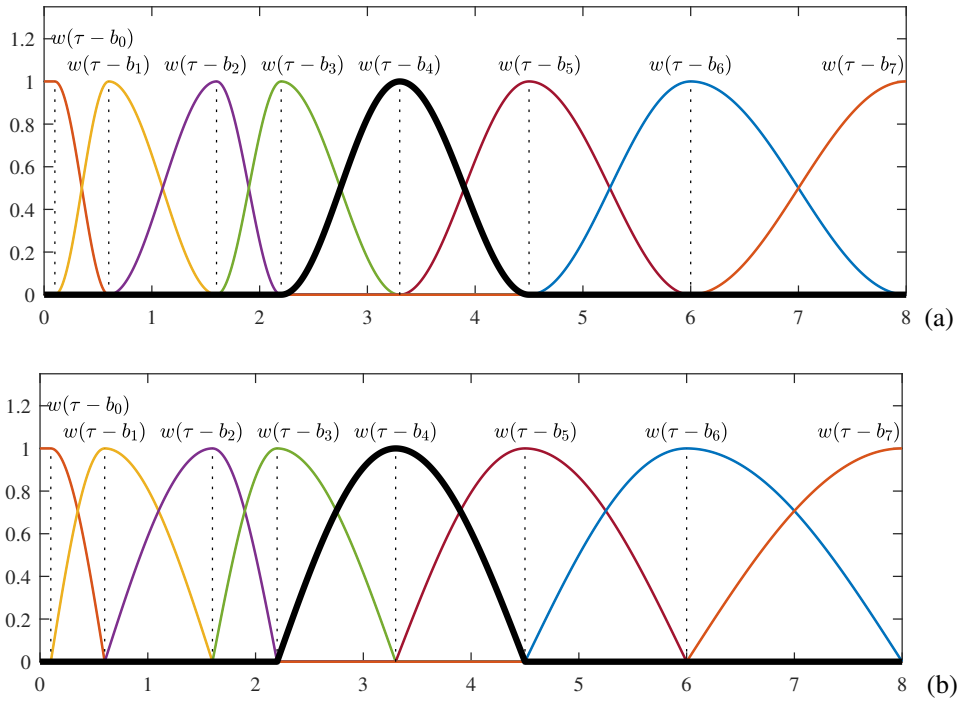


Figure 10.25 (a) Time-varying asymmetric Hann(ing) windows, $0 \leq t \leq 8$, that satisfy the constant overlap-add (COLA) reconstruction condition $\sum_k w(\tau - b_k) = 1$, with $b_k \in \{0.1,\ 0.6,\ 1.6,\ 2.2,\ 3.3,\ 4.5,\ 6.0,\ 8.0\}$, $k = 0,1,2,3,4,5,6,7$. (b) Time-varying asymmetric square root of the Hann(ing) windows (sine window), $0 \leq t \leq 8$, that satisfy the weighted overlap-add (WOLA) reconstruction condition $\sum_k w^2(\tau - b_k) = 1$, with $b_k \in \{0.1,\ 0.6,\ 1.6,\ 2.2,\ 3.3,\ 4.5,\ 6.0,\ 8.0\}$, $k = 0,1,2,3,4,5,6,7$.

with $a_{k+1} = b_k$, $b_{k+1} = c_k$ and the initial and the final intervals defined as in (10.42) and (10.43). The problem of this window is in its differentiability at the interval ending points, as can be seen from Fig. 10.25(b), causing slow frequency domain convergence. This problem will be addressed later, within the continuous wavelet transform analysis.

10.6.2 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.10.26

$$STFT_{N_i}(n,k_i) = \sum_{m=-N_i/2}^{N_i/2-1} w(m)x(n+m)e^{-j\frac{2\pi}{N_i}mk_i}.$$

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 10.26(left).

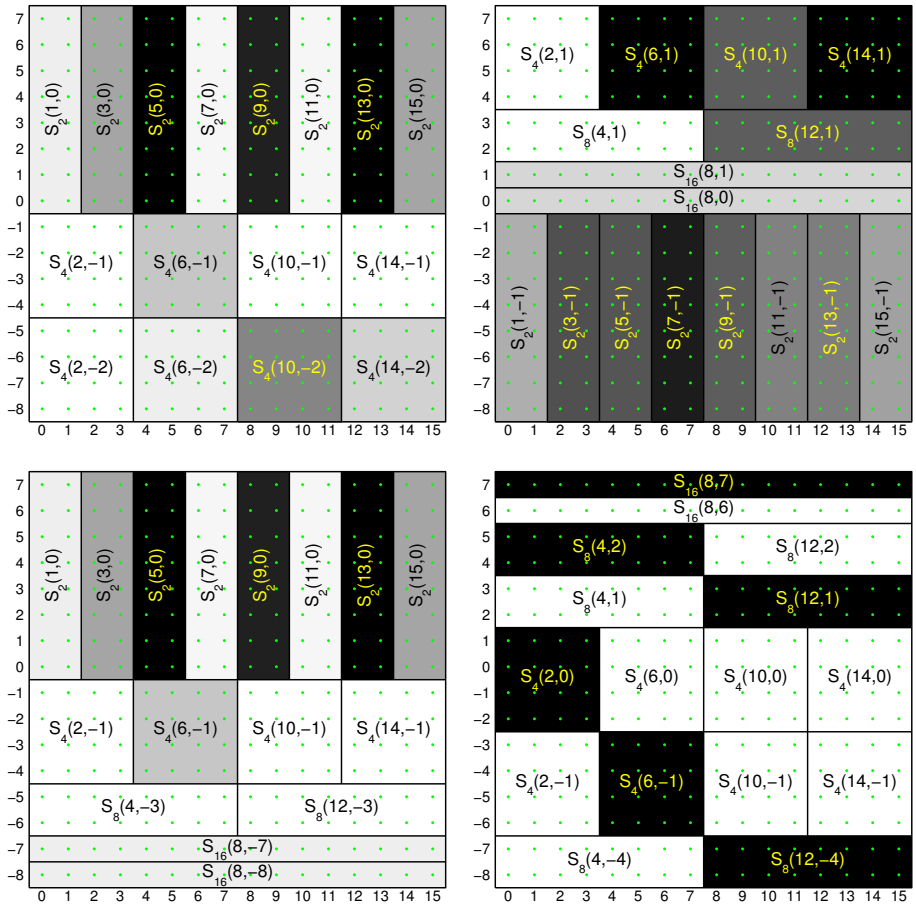


Figure 10.26 Time-frequency analysis with the STFT using frequency-varying windows.

For the signal used to illustrate the frequency-varying STFT in 10.26, 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 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

$$STFT(n,k) = \frac{1}{M}\sum_{i=0}^{M-1} P(i)X(k+i)e^{j2\pi in/M}, \tag{10.47}$$

$$\mathbf{STFT}_M(k) = \mathbf{W}_M^{-1}\mathbf{P}_M^{-1}\mathbf{X}(k).$$

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

$$\mathbf{STFT} = \begin{bmatrix} \mathbf{W}_{N_0}^{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_{N_1}^{-1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{W}_{N_K}^{-1} \end{bmatrix} \mathbf{X}, \tag{10.48}$$

where $\mathbf{X} = [X(0),\ X(1),\ \ldots, X(M-1)]^T$ is the DFT vector. A specific form of the STFT with the frequency-varying windows is called the wavelet transform and will be considered later in the book.

10.6.3 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 $STFT_{N_{(i,l)}}(n_i, k_l)$,

$$STFT_{N_{(i,l)}}(n_i, k_l) = \sum_{m=-N_{(i,l)}/2}^{N_{(i,l)}/2-1} w_{(i,l)}(m) x(n_i + m) e^{-j\frac{2\pi}{N_{(i,l)}} m k_l} \tag{10.49}$$

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.10.27, for each STFT value an expression based on (10.49) 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.

10.7 LOCAL POLYNOMIAL FOURIER TRANSFORM

After the presentation of the wavelet transform we will shift back our attention to the frequency of the signal, rather than to its amplitude values. There are signals whose instantaneous frequency variations are known up to an unknown set of parameters. For example, many signals could be expressed as polynomial-phase signals

$$x(t) = Ae^{j(\Omega_0 t + a_1 t^2 + a_2 t^3 + \cdots + a_N t^{N+1})}$$

where the parameters $\Omega_0, a_1, a_2, \ldots, a_N$ are unknown. For nonstationary signals, this approach may be used if the nonstationary signal could be considered as a polynomial phase signal within the analysis window. In that case, the local polynomial Fourier transform (LPFT) may be used. It is defined as

$$LPFT_{\Omega_1,\Omega_2,\ldots,\Omega_N}(t,\Omega) = \int_{-\infty}^{\infty} x(t+\tau) w(\tau) e^{-j(\Omega\tau + \Omega_1\tau^2 + \Omega_2\tau^3 + \cdots + \Omega_N\tau^{N+1})} d\tau. \tag{10.50}$$

In general, parameters $\Omega_1, \Omega_2, \ldots, \Omega_N$ could be time dependent, that is, for each time instant t, the set of optimal parameters could be different.

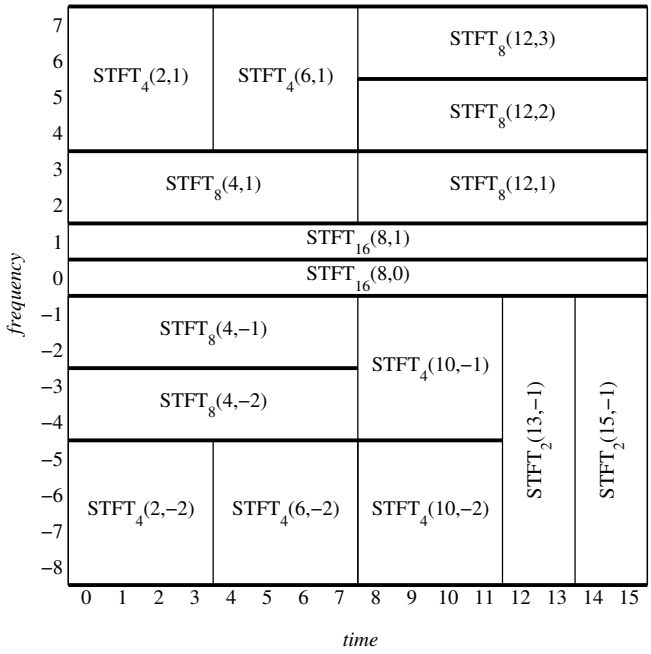


Figure 10.27 A time-frequency varying grid in the STFT calculation.

Realization of the LPFT reduces to the local signal $x(t+\tau)$ demodulation by $e^{-j(\Omega_1\tau^2+\Omega_2\tau^3+\cdots+\Omega_N\tau^{N+1})}$ followed by the STFT calculation.

Example 10.16. Consider the second-order polynomial-phase signal

$$x(t) = e^{j(\Omega_0 t + a_1 t^2)}.$$

Show that its LPFT could be completely concentrated along the instantaneous frequency.

★ Its LPFT has the form

$$\begin{aligned} LPFT_{\Omega_1}(t,\Omega) &= \int_{-\infty}^{\infty} x(t+\tau)w(\tau)e^{-j(\Omega\tau+\Omega_1\tau^2)}d\tau \\ &= e^{j(\Omega_0 t+a_1 t^2)}\int_{-\infty}^{\infty} w(\tau)e^{-j(\Omega-\Omega_0-2a_1 t)\tau}e^{-j(\Omega_1-a_1)\tau^2}d\tau. \end{aligned} \tag{10.51}$$

For $\Omega_1 = a_1$, the second-order phase term does not introduce any distortion to the local polynomial spectrogram,

$$\left|LPFT_{\Omega_1=a_1}(t,\Omega)\right|^2 = |W(\Omega-\Omega_0-2a_1 t)|^2,$$

with respect to the spectrogram of a sinusoid with constant frequency. For a wide window $w(\tau)$, like in the case of the STFT of a pure sinusoid, we achieve high concentration.

The LPFT could be considered as the Fourier transform of windowed signal demodulated with $\exp(-j(\Omega_1\tau^2+\Omega_2\tau^3+\cdots+\Omega_N\tau^{N+1}))$. Thus, if we are interested in signal filtering, we can find the coefficients $\Omega_1,\Omega_2,\ldots,\Omega_N$, demodulate the signal by multiplying it with $\exp(-j(\Omega_1\tau^2+\Omega_2\tau^3+\cdots+\Omega_N\tau^{N+1}))$ and use a standard filter for almost a pure sinusoid. In general, we can extend this approach to any signal $x(t)=e^{j\phi(t)}$ by estimating its phase $\phi(t)$ with $\widehat{\phi}(t)$ (using the instantaneous frequency estimation that will be discussed later) and filtering demodulated signal $x(t)\exp(-j\widehat{\phi}(t))$ by a lowpass filter. The resulting signal is obtained when the filtered signal is returned back to the original frequencies, by modulation with $\exp(j\widehat{\phi}(t))$.

Example 10.17. Consider the first-order LPFT of a signal $x(t)$. Show that the second-order moments of the LPFT could be calculated based on the windowed signal moment, windowed signal's Fourier transform moment and one more LPFT moment for any Ω_1 in (10.50), for example for $\Omega_1=1$.

★ The second-order moment of the first-order LPFT,

$$LPFT_{\Omega_1}(t,\Omega)=\int_{-\infty}^{\infty}x_t(\tau)e^{-j(\Omega\tau+\Omega_1\tau^2)}d\tau,$$

defined by

$$M_{\Omega_1}=\frac{1}{2\pi}\int_{-\infty}^{\infty}\Omega^2\left|LPFT_{\Omega_1}(t,\Omega)\right|^2d\Omega \tag{10.52}$$

is equal to

$$M_{\Omega_1}=\int_{-\infty}^{\infty}\left|\frac{d\left(x_t(\tau)e^{-j\Omega_1\tau^2}\right)}{d\tau}\right|^2d\tau,$$

since the LPFT could be considered as the Fourier transform of $x_t(\tau)e^{-j\Omega_1\tau^2}$, that is, $LPFT_{\Omega_1}(t,\Omega)=\mathrm{FT}\{x_t(\tau)e^{-j\Omega_1\tau^2}\}$, and the Parseval's theorem is used. After the derivative calculation

$$M_{\Omega_1}=\int_{-\infty}^{\infty}\left|\frac{dx_t(\tau)}{d\tau}-j2\Omega_1\tau x_t(\tau)\right|^2d\tau=$$

$$\int_{-\infty}^{\infty}(\left|\frac{dx_t(\tau)}{d\tau}\right|^2+j2\Omega_1\tau x_t^*(\tau)\frac{dx_t(\tau)}{d\tau}-j2\Omega_1\tau x_t(\tau)\frac{dx_t^*(\tau)}{d\tau}+|2\Omega_1\tau x_t(\tau)|^2)d\tau.$$

We can recognize some of the terms in the last line, as

$$M_0=\int_{-\infty}^{\infty}\left|\frac{dx_t(\tau)}{d\tau}\right|^2d\tau=\frac{1}{2\pi}\int_{-\infty}^{\infty}\Omega^2\left|LPFT_{\Omega_1=0}(t,\Omega)\right|^2d\Omega.$$

This is the moment of $X_t(\Omega) = \text{FT}\{x_t(\tau)\}$, since the integral of $|dx_t(\tau)/d\tau|^2$ over τ is equal to the integral of $|j\Omega X_t(\Omega)|^2$ over Ω, according to Parseval's theorem. Also, we can see that the last term in M_{Ω_1} contains the signal moment,

$$m_x = \int_{-\infty}^{\infty} \tau^2 |x_t(\tau)|^2 d\tau, \tag{10.53}$$

multiplied by $4\Omega_1^2$. Then, it is easy to conclude that

$$M_{\Omega_1} - M_0 - 4m_x\Omega_1^2 = \Omega_1 \int_{-\infty}^{\infty} \left(j2\tau x_t^*(\tau) \frac{d[x_t(\tau)]}{d\tau} - j2\tau x_t(\tau) \frac{d[x_t^*(\tau)]}{d\tau} \right) d\tau.$$

Note that the last integral does not depend on parameter Ω_1. Thus, the relation among the LPFT moments at any two Ω_1, for example, $\Omega_1 = a$ and an arbitrary Ω_1, easily follows as the ratio

$$\frac{M_{\Omega_1=a} - M_0 - 4a^2 m_x}{M_{\Omega_1} - M_0 - 4\Omega_1^2 m_x} = \frac{a}{\Omega_1}. \tag{10.54}$$

With $a = 1$, by leaving the notation for an arbitrary Ω_1 unchanged, we get

$$\frac{M_1 - M_0 - 4m_x}{M_{\Omega_1} - M_0 - 4\Omega_1^2 m_x} = \frac{1}{\Omega_1}, \tag{10.55}$$

with $M_1 = M_{\Omega_1=1}$.

Obviously, the second-order moment, for any Ω_1, can be expressed as a function of other three moments. In this case the relation reads

$$M_{\Omega_1} = 4\Omega_1^2 m_x + \Omega_1(M_1 - M_0 - 4m_x) + M_0.$$

Example 10.18. Find the position and the value of the second-order moment minimum of the LPFT, based on the windowed signal moment, the windowed signal's Fourier transform moment, and the LPFT moment for $\Omega_1 = 1$.

★ The minimum value of the second-order moment (meaning the best concentrated LPFT in the sense of the duration measures) could be calculated from

$$\frac{dM_{\Omega_1}}{d\Omega_1} = 0$$

as

$$\Omega_1 = -\frac{M_1 - M_0 - 4m_x}{8m_x}.$$

Since $m_x > 0$ this is a minimum of the function M_{Ω_1}. Thus, in general, there is no need for a direct search for the best concentrated LPFT over all possible values of Ω_1. It can be found based on three moments.

The value of M_{Ω_1} is

$$M_{\Omega_1} = M_0 - \frac{(M_1 - M_0 - 4m_x)^2}{16m_x}. \tag{10.56}$$

Note that any two moments, instead of M_0 and M_1, could be used in the derivation.

The fractional Fourier transform easily reduces to the first-order LPFT.

10.7.1 Fractional Fourier Transform with Relation to the LPFT

The fractional Fourier transform (FRFT) for an angle α ($\alpha \neq k\pi$) is defined as

$$X_\alpha(u) = \int_{-\infty}^{\infty} x(\tau) K_\alpha(u,\tau) d\tau, \tag{10.57}$$

where

$$K_\alpha(u,\tau) = \sqrt{\frac{1 - j\cot\alpha}{2\pi}} e^{j(u^2/2)\cot\alpha} e^{j(\tau^2/2)\cot\alpha} e^{-ju\tau\csc\alpha}. \tag{10.58}$$

It can be considered as a rotation of signal in the time-frequency plane for an angle α. Its inverse can be considered as a rotation for angle $-\alpha$

$$x(t) = \int_{-\infty}^{\infty} X_\alpha(u) K_{-\alpha}(u,t) du.$$

Special cases of the FRFT reduce to: $X_0(u) = x(u)$ and $X_{\pi/2}(u) = X(u)/\sqrt{2\pi}$, that is, the signal and its Fourier transform.

The windowed FRFT is

$$X_{w,\alpha}(t,u) = \sqrt{\frac{1-j\cot\alpha}{2\pi}} e^{j(u^2/2)\cot\alpha} \int_{-\infty}^{\infty} x(t+\tau) w(\tau) e^{j(\tau^2/2)\cot\alpha} e^{-ju\tau\csc\alpha} d\tau. \tag{10.59}$$

Relation between the windowed FRFT and the first-order LPFT is

$$X_{w,\alpha}(t,u) = \sqrt{\frac{1 - j\cot\alpha}{2\pi}} e^{j(u^2/2)\cot\alpha} LPFT_{\Omega_1}(t,\Omega) \tag{10.60}$$

where $\Omega_1 = \cot(\alpha)/2$ and $\Omega = u\csc(\alpha)$. Thus, all results can be easily converted from the first-order LPFT to the windowed FRFT, and vice versa. That is the reason why we will not present a detailed analysis for this transform after the LPFT has been presented.

By using a window, local forms of the FRFT are introduced as:

$$STFT_\alpha(u,v) = \int_{-\infty}^{\infty} X_\alpha(u+\tau) w(\tau) e^{-jv\tau} d\tau \tag{10.61}$$

$$STFT_\alpha(u,v) = \int_{-\infty}^{\infty} x(t+\tau) w(\tau) K_\alpha(u,\tau) d\tau \tag{10.62}$$

meaning that the lag truncation could be applied after signal rotation or prior to the rotation. Results are similar. A similar relation for the moments, like (10.55) in the case of LPFT, could be derived here. It states that any FRFT moment can be calculated if we know just any three of its moments.

10.8 HIGH-RESOLUTION STFT

High-resolution techniques are developed for efficient processing and separation of very close sinusoidal signals (in array signal processing, separation of sources with very close DOAs). Among these techniques the most widely used are Capon's method, MUSIC, and ESPRIT. The formulation of high-resolution techniques could be extended to the time-frequency representations. Here we will present a simple formulation of the STFT and the LPFT within Capon's method framework.

10.8.1 Capon's STFT

Here we will present the STFT formulation in a common array signal-processing notation. The STFT of a discrete time signal $x(n)$ in (causal) notation

$$STFT(\omega,n) = \frac{1}{N}\sum_{n=0}^{N-1} x(n+m)e^{-j\omega m}$$

can be written as

$$\begin{aligned} STFT(\omega,n) = \hat{s}_\omega(n) &= \mathbf{h}^H\mathbf{x}(n) = \frac{1}{N}\mathbf{a}^H(\omega)\mathbf{x}(n) \\ \mathbf{a}^H(\omega) &= [1\ e^{-i\omega}\ e^{-i\omega 2}\ldots e^{-i\omega(N-1)}] \\ \mathbf{x}(n) &= [x(n)\ x(n+1)\ x(n+2)\ldots\ x(n+N-1)]^T, \end{aligned} \tag{10.63}$$

where ^{T} denotes the transpose operation, and ^{H} denotes the conjugate and transpose (Hermitian) operation. Normalization of the STFT with N is done, as in the robust signal analysis.

The average power of the output signal $\hat{s}_\omega(n)$, over M samples (ergodicity over M samples around n is assumed), for a frequency ω, is

$$\begin{aligned} P(\omega) &= \frac{1}{M}\sum_n |\hat{s}_\omega(n)|^2 \\ &= \frac{1}{N^2}\mathbf{a}^H(\omega)\frac{1}{M}\sum_n[\mathbf{x}(n)\mathbf{x}^H(n)]\mathbf{a}(\omega) = \frac{1}{N^2}\mathbf{a}^H(\omega)\hat{\mathbf{R}}_x\mathbf{a}(\omega), \end{aligned} \tag{10.64}$$

where $\hat{\mathbf{R}}_x$ is the matrix defined by

$$\hat{\mathbf{R}}_x = \frac{1}{M}\sum_n \mathbf{x}(n)\mathbf{x}^H(n).$$

The standard STFT (10.63) can be derived based on the following consideration. Find $\mathbf{h}$ as a solution of the problem

$$\min_{\mathbf{h}}\{\mathbf{h}^H\mathbf{h}\} \quad \text{subject to } \mathbf{h}^H\mathbf{a}(\omega) = 1. \tag{10.65}$$

This minimization problem will be explained through the next example.

Example 10.19. Show that the output power of the filter producing $s(n) = \mathbf{h}^H \mathbf{x}(n)$ is minimized for the input $\mathbf{x}(n) = A\mathbf{a}(\omega) + \varepsilon(n)$, with respect the input white noise $\varepsilon(n)$, whose autocorrelation function is $\hat{\mathbf{R}}_\varepsilon = \rho \mathbf{I}$ if $\mathbf{h}^H \mathbf{h}$ is minimum subject to $\mathbf{h}^H \mathbf{a}(\omega) = 1$.

★ The output for the noise only is $s_\varepsilon(n) = \mathbf{h}^H \varepsilon(n)$, while its average power is

$$\frac{1}{M}\sum_n |\mathbf{h}^H \varepsilon(n)|^2 = \frac{1}{M}\sum_n \mathbf{h}^H \varepsilon(n)\varepsilon^H(n)\mathbf{h}$$
$$= \mathbf{h}^H \left(\frac{1}{M}\sum_n \varepsilon(n)\varepsilon^H(n) \right) \mathbf{h} = \rho\, \mathbf{h}^H \mathbf{h}.$$

Minimization of $\mathbf{h}^H \mathbf{h}$ is therefore equivalent to the output white noise power minimization.

The condition $\mathbf{h}^H \mathbf{a}(\omega) = 1$ means that the input in form of a sinusoid $A\mathbf{a}(\omega)$, at frequency ω, should not be changed, that is, if $\mathbf{x}(n) = A\mathbf{a}(\omega)$, then

$$\mathbf{h}^H \mathbf{x}(n) = \mathbf{h}^H A\mathbf{a}(\omega) = A.$$

Thus, the condition $\mathbf{h}^H \mathbf{a}(\omega) = 1$ means that the estimate is unbiased with respect to input sinusoidal signal with amplitude A.

The solution of minimization problem (10.65) is

$$\frac{\partial}{\partial \mathbf{h}^H}\{\mathbf{h}^H \mathbf{h} + \lambda(\mathbf{h}^H \mathbf{a}(\omega) - 1)\} = 0 \quad \text{subject to } \mathbf{h}^H \mathbf{a}(\omega) = 1,$$
$$2\mathbf{h} = -\lambda \mathbf{a}(\omega) \quad \text{subject to } \mathbf{h}^H \mathbf{a}(\omega) = 1$$

resulting in

$$\mathbf{h} = \frac{\mathbf{a}(\omega)}{\mathbf{a}^H(\omega)\mathbf{a}(\omega)} = \frac{1}{N}\mathbf{a}(\omega) \tag{10.66}$$

and the estimate (10.63), which is the standard STFT, follows.

Consider now a different optimization problem, defined by

$$\min_{\mathbf{h}}\{\frac{1}{M}\sum_n |\mathbf{h}^H \mathbf{x}(n)|^2\} \quad \text{subject to } \mathbf{h}^H \mathbf{a}(\omega) = 1. \tag{10.67}$$

Two points are emphasized in this optimization problem. First, the weights are selected to minimize the average power $\frac{1}{M}\sum_n |\mathbf{h}^H \mathbf{x}(n)|^2$ of the output signal of the filter. It means that the filter should give the best possible suppression of all components of signals-plus-noise components of the observations as well as a suppression of the components of the desired signal for all time-instants (minimization of the power of $y(n)$). Second, by setting the condition $\mathbf{h}^H \mathbf{a}(\omega) = 1$, in the considered time instant n the signal amplitude is preserved at the output.

The optimization problem can be rewritten in the form

$$\min_{\mathbf{h}}\{\frac{1}{M}\sum_n \mathbf{h}^H \mathbf{x}(n)\mathbf{x}^H(n)\mathbf{h}\} \quad \text{subject } \mathbf{h}^H \mathbf{a}(\omega) = 1.$$

By denoting

$$\hat{\mathbf{R}}_\mathbf{x} = \frac{1}{M}\sum_n \mathbf{x}(n)\mathbf{x}^H(n),$$

we get

$$\min_{\mathbf{h}}\{\mathbf{h}^H\hat{\mathbf{R}}_{\mathbf{x}}\mathbf{h}\} \quad \text{subject to } \mathbf{h}^H\mathbf{a}(\omega)=1.$$

The constrained minimization

$$\frac{\partial}{\partial \mathbf{h}^H}\{\mathbf{h}^H\hat{\mathbf{R}}_{\mathbf{x}}\mathbf{h}+\lambda(\mathbf{h}^H\mathbf{a}(\omega)-1)\}=0 \quad \text{subject to } \mathbf{h}^H\mathbf{a}(\omega)=1.$$

gives the solution

$$\mathbf{h}=-\hat{\mathbf{R}}_{\mathbf{x}}^{-1}\frac{\lambda\mathbf{a}(\omega)}{2} \quad \text{subject to } \mathbf{h}^H\mathbf{a}(\omega)=1. \tag{10.68}$$

The solution can be written in the form

$$\hat{\mathbf{h}}=\frac{\hat{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{a}(\omega)}{\mathbf{a}^H(\omega)\hat{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{a}(\omega)}, \tag{10.69}$$

where

$$\hat{\mathbf{R}}_{\mathbf{x}}=\frac{1}{M}\sum_n \mathbf{x}(n)\mathbf{x}^H(n). \tag{10.70}$$

The output signal power, in these cases, corresponds to Capon's form of the STFT, defined by

$$S_{\text{Capon}}(\omega)=\frac{1}{M}\sum_n|\mathbf{h}^H\mathbf{x}(n)|^2=\mathbf{h}^H\hat{\mathbf{R}}_{\mathbf{x}}\mathbf{h} \tag{10.71}$$

$$=\left(\frac{\hat{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{a}(\omega)}{\mathbf{a}^H(\omega)\hat{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{a}(\omega)}\right)^H\hat{\mathbf{R}}_{\mathbf{x}}\frac{\hat{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{a}(\omega)}{\mathbf{a}^H(\omega)\hat{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{a}(\omega)} \tag{10.72}$$

$$=\frac{1}{\mathbf{a}^H(\omega)\hat{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{a}(\omega)}. \tag{10.73}$$

Note that $\mathbf{a}^H(\omega)\hat{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{a}(\omega)$ is a real valued scalar. Along with (10.70), we can use a sliding window estimate of the autocorrelation matrix in the form

$$\hat{\mathbf{R}}_{\mathbf{x}}(n)=\frac{1}{K+1}\sum_{p=n-K/2}^{n+K/2}\mathbf{x}(p)\mathbf{x}^H(p), \tag{10.74}$$

where K is a parameter defining the width of a symmetric sliding window. Inserting $\hat{\mathbf{R}}_{\mathbf{x}}(n,K)$ instead of $\hat{\mathbf{R}}_{\mathbf{x}}$ in (10.71) gives the STFT with weights minimizing the output power in (10.67), for the observations in the neighborhood of the time instant of interest n.

The mean value of this power function, calculated in the neighborhood of the time n over the window used in (10.74), gives an averaged Capon's STFT as follows

$$S_{\text{Capon}}(n,\omega)=\frac{1}{\mathbf{a}^H(\omega)\hat{\mathbf{R}}_{\mathbf{x}}^{-1}(n)\mathbf{a}(\omega)}. \tag{10.75}$$

where n indicates the time instant of the interest and the mean is calculated over the observations $\mathbf{y}(n)$ in the corresponding window.

In the realization the autocorrelation function is regularized by a unity matrix $\mathbf{I}$ thus, we use

$$\hat{\mathbf{R}}(n)=\frac{1}{K+1}\sum_{p=n-K/2}^{n+K/2}\mathbf{x}(p)\mathbf{x}^H(p)+\rho\mathbf{I}. \tag{10.76}$$

instead of $\hat{\mathbf{R}}_{\mathbf{x}}(n)$ for the inverse calculation in (10.75) and (10.71).

10.8.2 MUSIC STFT

In the MUSIC formulation of the high resolution STFT the eigenvalue decomposition of the autocorrelation matrix (10.76) is used as

$$\hat{\mathbf{R}}(n) = \frac{1}{K+1}\sum_{p=n-K/2}^{n+K/2}\mathbf{x}(p)\mathbf{x}^H(p) + \rho\mathbf{I} = \mathbf{V}^H(n)\Lambda(n)\mathbf{V}(n),$$
$$\hat{\mathbf{R}}^{-1}(n) = \mathbf{V}^H(n)\Lambda^{-1}(n)\mathbf{V}(n).$$

Note that the Capon spectrogram, using eigenvalues and eigenvectors of the autocorrelation matrix, can be written as

$$S_{\text{Capon}}(n,\omega) = \frac{1}{\mathbf{a}^H(\omega)\mathbf{V}^H(n)\Lambda^{-1}(n)\mathbf{V}(n)\mathbf{a}(\omega)} = \frac{1}{\sum_{k=1}^{N}\frac{1}{\lambda_k}|STFT_k(n,\omega)|^2}$$

where

$$STFT_k(n,\omega) = \mathbf{a}^H(\omega)\mathbf{v}_k(n)$$

is the STFT of the kth eigenvector (column) of the autocorrelation matrix $\hat{\mathbf{R}}(n)$, corresponding to the eigenvalue λ_k. If the signal has $N-M$ components then the first $N-M$ largest eigenvalues λ_k (corresponding to the smallest values $1/\lambda_k$) will represent the signal space (components), and the remaining M eigenvalues will correspond to the noise space (represented by $\rho\mathbf{I}$ in the definition of autocorrelation matrix $\hat{\mathbf{R}}(n)$).

If a frequency ω corresponds to a signal component, then all eigenvectors corresponding to the noise space will be orthogonal to that harmonic, being represented by $\mathbf{a}^H(\omega)$. It means that the spectrograms of all noise space only components will be very small at the frequencies corresponding to the signal frequencies.

The MUSIC STFT is defined based on this fact. It is calculated using the eigenvectors corresponding to noise space, as

$$S_{\text{MUSIC}}(n,\omega) = \frac{1}{\mathbf{a}^H(\omega)\mathbf{V}_M^H\mathbf{V}_M\mathbf{a}(\omega)} = \frac{1}{\sum_{k=N-M+1}^{N}|STFT_k(n,\omega)|^2}, \tag{10.77}$$

where $\mathbf{V}_M$ is the eigenvector matrix containing only M eigenvectors corresponding to the M lowest eigenvalues in Λ, representing the space of noise. In this case the signal has $N-M$ components corresponding to the largest eigenvalues. A special case with $M=1$ is the Pisarenko method.

Example 10.20. Calculate high resolution forms of the spectrogram for two-component signal whose frequencies $\omega_0+\Delta\omega$ and $\omega_0-\Delta\omega$ may be considered as constants around the instant of interest $n=128$,

$$x(n) = \exp(jn(\omega_0+\Delta\omega)) + \exp(jn(\omega_0-\Delta\omega)),$$
$$\omega_0 = 1 \text{ and } \Delta\omega = 0.05.$$

In the STFT calculation use a rectangular window of the width $N = 16$. Use 15 samples for averaging (estimation) of the autocorrelation matrix, as well as its regularization by a $0.0001 \cdot \mathbf{I}$ (corresponding to noise signal $x(n) + \varepsilon(n)$, where $\varepsilon(n)$ is complex white noise with variance $\sigma_\varepsilon^2 = 0.0001$). Assume that signal samples needed for autocorrelation function estimation are also available.

★ Signal values around $n = 128$ are considered. The STFT is calculated using $N = 16$ signal samples

$$\mathbf{x}(128) = [x(128)\ x(129)\ x(130)\dots\ x(143)]^T$$

and a rectangular window. The mainlobe with of this window is $D = 4\pi/N = \pi/4 = 0.7854$. Its will not be able to resolve two components closer than $2\Delta\omega \sim D/2 = 0.3927$. Considered $\Delta\omega = 0.05$ is well below this limit. The STFT is interpolated in frequency up to 2048 samples. The result is shown in Fig. 10.28(a). Next the autocorrelation matrix

$$\hat{\mathbf{R}}(128) = \frac{1}{15}\sum_{p=128-7}^{128+7} \mathbf{x}(p)\mathbf{x}^H(p) + 0.00001 \cdot \mathbf{I}$$

is estimated using the signal vectors $\mathbf{x}(p) = [x(p)\ x(p+1)\ x(p+2)\dots\ x(p+15)]$. Note that values of signal from $x(128-7)$ for $p = 128-7$ up to $p = 128+7+15$ are needed for this calculation. Values of vector

$$\mathbf{a}(\omega) = [1\ e^{i\omega}\ e^{i\omega 2}\dot{s}e^{i\omega(N-1)}]^T$$

are calculated at the frequencies of interest $\omega = 2\pi k/2048$, for $k = 0,1,2,\dots,1023$. The Capon's STFT is then

$$S_{\text{Capon}}(128,\omega) = \frac{1}{\mathbf{a}^H(\omega)\hat{\mathbf{R}}^{-1}(128)\mathbf{a}(\omega)} = \frac{1}{\sum\limits_{k=1}^{16}\frac{1}{\lambda_k}|STFT_k(n,\omega)|^2}.$$

Its value is presented in Fig. 10.28(b),(d).

The MUSIC spectrogram is obtained by calculating the eigenvectors of $\hat{\mathbf{R}}(128)$ and using only $N-2$ eigenvectors corresponding to the noise space eigenvalues of this matrix (there are 2 signal components)

$$S_{\text{MUSIC}}(n,\omega) = \frac{1}{\mathbf{a}^H(\omega)\mathbf{V}_{14}^H\mathbf{V}_{14}\mathbf{a}(\omega)} = \frac{1}{\sum\limits_{k=3}^{16}|STFT_k(n,\omega)|^2}$$

where $\mathbf{V}_{14}$ is a 14×16 matrix containing 14 eigenvectors $\mathbf{v}_k(n)$, $k = 3,4,\dots 16$, corresponding to the noise space (2 eigenvectors corresponding to two largest eigenvalues, being the signal space, are omitted). The STFT of eigenvector $\mathbf{v}_k(n)$ is denoted by $STFT_k(n,\omega)$. The MUSIC spectrogram is presented in Fig. 10.28(c),(e).

The case corresponding to one eigenvector being used in the spectrogram $|STFT_{16}(n,\omega)|^2$ (a form of Pisarenko spectrogram, when only the lowest eigenvector is considered as the noise space) is presented in Fig. 10.28(f). Note that in the case of Pisarenko spectrogram it is sufficient (and required by its definition) to use only $N = 3$ window width (number of components plus one).

Normalized values of all spectrograms are presented in Fig. 10.28.

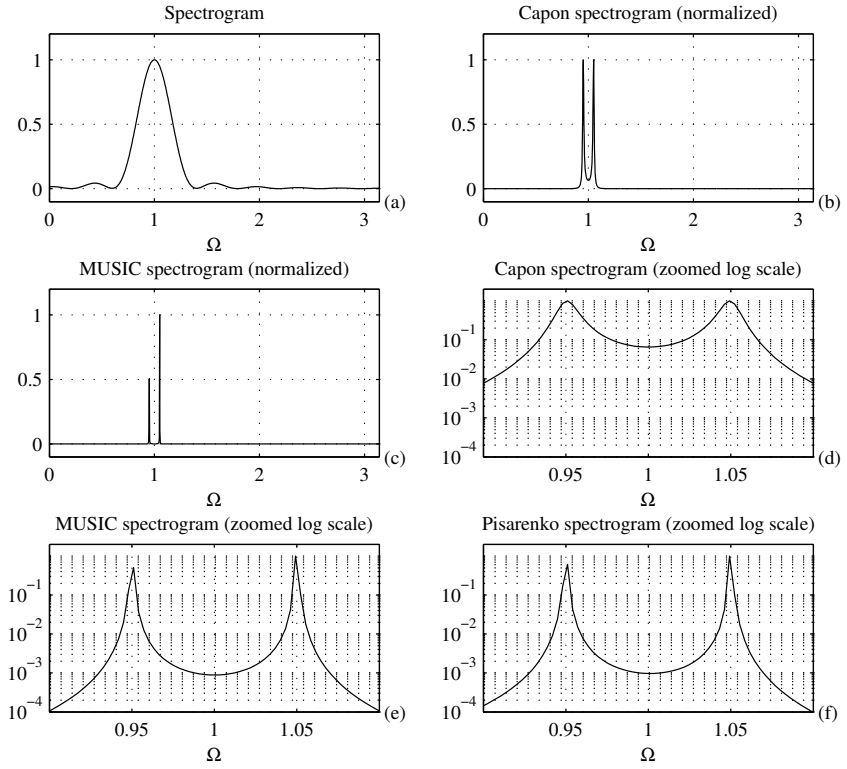


Figure 10.28 (a) The standard STFT using a rectangular window $N = 16$. The STFT is interpolated in frequency up to 2048 samples. (b) Capon's spectrogram calculated in 2048 frequency points. (c) MUSIC spectrogram calculated in 2048 frequency points. (d) Capon's spectrogram zoomed to the signal components. (e) MUSIC spectrogram zoomed to the signal components. (f) Pisarenko spectrogram zoomed to the signal components.

10.8.3 Capon's LPFT

With varying coefficients or appropriate signal multiplication, before the STFT calculation, a local polynomial version of Capon's transform could be defined. For example, for a linear frequency-modulated signal of the form

$$x(n) = Ae^{j(\alpha_0 n^2 + \omega_0 n + \varphi_0)}$$

we should use (10.75) or (10.71) with a signal of the form

$$\hat{\mathbf{R}}_{\mathbf{x}}(n,K,\alpha) = \frac{1}{K+1}\sum_{p=n-K/2}^{n+K/2} \mathbf{x}_{\alpha}(p)\mathbf{x}_a^H(p) \quad \text{with} \quad \mathbf{x}_{\alpha}(p) = \mathbf{x}(p)e^{-j\alpha p^2},$$

where α as a parameter. The high-resolution form of the LPFT can be used for efficient processing of close linear frequency-modulated signals, with the same rate within the considered interval.

Example 10.21. The Capon LPFT form is illustrated on an example with a signal with two close components

$$x(t) = \exp(j128\pi t(0.55 - t/2) + j5\pi t^3) + \exp(j128\pi t(0.45 - t/2) + j5\pi t^3),$$

that in addition to the linear frequency-modulated contained a small disturbing cubic phase term. The considered time interval was $-1 \leq t \leq 1 - \Delta t$ with $\Delta t = 2/512$, $\rho = 0.5$, $K = 30$, and the frequency domain is interpolated eight times. The standard STFT, LPFT, Capon's STFT, and Capon's LPFT-based representations are presented in Fig. 10.29.

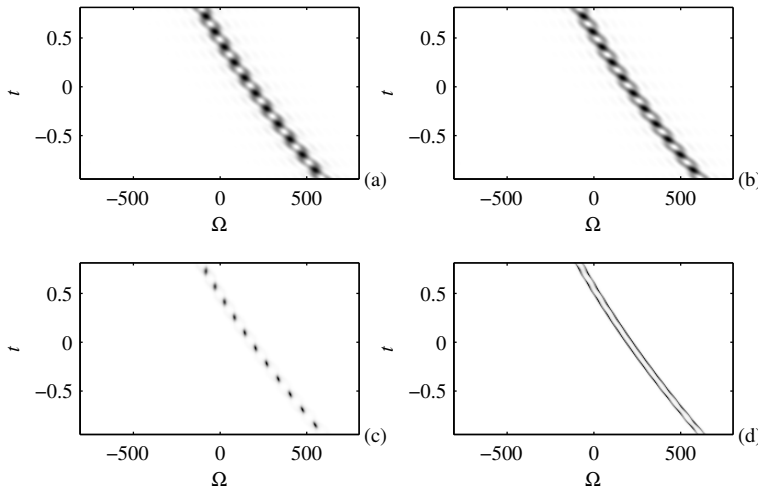


Figure 10.29 (a) The standard STFT, (b) the LPFT, (c) Capon's STFT, and (d) Capon's LPFT-based representations of two close almost linear frequency-modulated signals.

In general, higher-order polynomial or any other nonstationary signal, with appropriate parametrization, can be analyzed in the same way.

Chapter 11

Quadratic Time-Frequency Representations

The dimensions of the STFT blocks (resolutions) are determined by the window width. The best STFT for a signal would be the one whose window form fits the best to the signal's time-frequency content. Consider, for example, an important and simple signal such as a linear frequency modulated (LFM) chirp. For simplicity of analysis assume that its instantaneous frequency (IF) coincides with the time-frequency plane diagonal. It is obvious that, due to symmetry, both time and frequency resolution are equally important. Therefore, the best STFT would be the one calculated using a constant window whose (equivalent) widths are equal in time and frequency domain. With such a window both resolutions will be the same. However, these resolutions could be unacceptably low for many applications. It means that the STFT, including all of its possible time and/or frequency-varying window forms, would be unacceptable as a time-frequency representation of this signal. The overlapping STFT could be used for better signal tracking, without any effect on the resolution.

11.1 WIGNER DISTRIBUTION

A way to improve time-frequency representation of this signal is in transforming the signal into a sinusoid whose constant frequency is equal to the instantaneous frequency value of the linear frequency modulated signal at the considered instant. Then, a wide window can be used, with a high frequency resolution. The obtained result is valid for the considered instant only and the signal transformation procedure should be repeated for each instant of interest.

A simple way to introduce this kind of signal representation is presented. Consider an LFM signal,

$$x(t) = A\exp(j\phi(t)) = A\exp(j(at^2/2 + bt + c)).$$

Its instantaneous frequency changes in time as

$$\Omega_i(t) = d\phi(t)/dt = at + b.$$

One of the goals of time-frequency analysis is to obtain a function that will (in an ideal case) fully concentrate the signal power along its instantaneous frequency. The ideal representation would be

$$I(t,\Omega) = 2\pi A^2\delta(\Omega - \Omega_i(t)).$$

For a quadratic function $\phi(t)$, it is known that

$$\tau\frac{d\phi(t)}{dt} = \phi(t + \frac{\tau}{2}) - \phi(t - \frac{\tau}{2}) = \tau(at + b) = \tau\Omega_i(t).$$

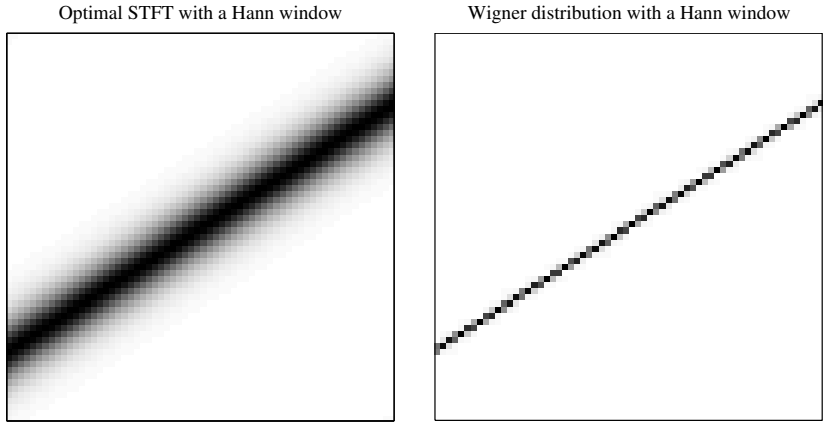


Figure 11.1 Optimal STFT (absolute value, calculated with optimal window width) and the Wigner distribution of a linear frequency modulated signal.

This property can easily be converted into an ideal time-frequency representation for the linear frequency modulated signal by using

$$\mathrm{FT}_\tau\{x(t+\tau/2)x^*(t-\tau/2)\} = \mathrm{FT}_\tau\{A^2 e^{j\Omega_i(t)\tau}\} = 2\pi A^2\delta(\Omega-\Omega_i(t)).$$

The Fourier transform of $x(t+\tau/2)x^*(t-\tau/2)$ over τ, for a given t, is called the Wigner distribution. It is defined as

$$WD(t,\Omega) = \int_{-\infty}^{\infty} x(t+\tau/2)x^*(t-\tau/2)e^{-j\Omega\tau}d\tau. \tag{11.1}$$

The Wigner distribution is originally introduced in quantum mechanics. The illustration of the Wigner distribution calculation is presented in Fig. 11.2.

Expressing $x(t)$ in terms of $X(\Omega)$ and substituting it into (11.1) we get

$$WD(t,\Omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\Omega+\theta/2)X^*(\Omega-\theta/2)e^{j\theta t}d\theta \tag{11.2}$$

what represents a definition of the Wigner distribution in the frequency domain.

It is easy to show that the Wigner distribution satisfies the marginal properties. From the Wigner distribution definition, it follows

$$x(t+\tau/2)x^*(t-\tau/2) = \mathrm{IFT}\{WD(t,\Omega)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} WD(t,\Omega)e^{j\Omega\tau}d\Omega \tag{11.3}$$

which, for $\tau = 0$, produces (11.29)

$$|x(t)|^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} WD(t,\Omega)d\Omega. \tag{11.4}$$

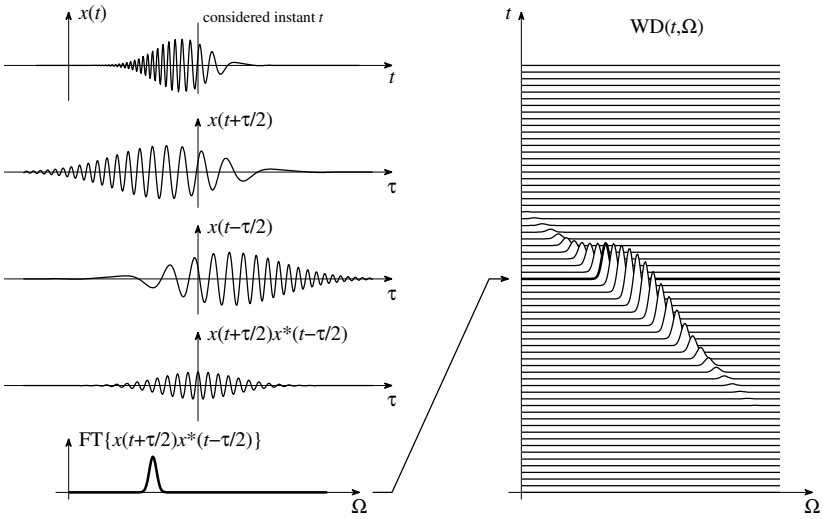


Figure 11.2 Illustration of the Wigner distribution calculation, for a considered time instant t. Real values of a linear frequency modulated signal (linear chirp) are presented.

Based on the definition of the Wigner distribution in the frequency domain, (11.2), one may easily prove the fulfillment of the frequency marginal.

Example 11.1. Find the Wigner distribution of signals: (a) $x(t) = \delta(t - t_1)$ and (b) $x(t) = \exp(j\Omega_1 t)$.

★ The Wigner distribution of signal $x(t) = \delta(t - t_1)$ is

$$WD(t,\Omega) = \int_{-\infty}^{\infty} \delta(t - t_1 + \tau/2)\delta(t - t_1 - \tau/2)e^{-j\Omega\tau}d\tau$$
$$= 2\delta(2(t - t_1))e^{-j2\Omega(t-t_1)} = \delta(t - t_1),$$

since $|a|\,\delta(at)x(t) = \delta(t)x(0)$. From the Wigner distribution definition in terms of the Fourier transform, for $x(t) = \exp(j\Omega_1 t)$ with $X(\Omega) = 2\pi\delta(\Omega - \Omega_1)$, follows

$$WD(t,\Omega) = 2\pi\delta(\Omega - \Omega_1).$$

A high concentration of time-frequency representation for both of these signals is achieved. Note that this fact does not mean that we will be able to achieve an arbitrary high concentration simultaneously, in a point, in the time-frequency domain. □

Example 11.2. Consider a linear frequency modulated signal, $x(t) = Ae^{jbt^2/2}$. Find its Wigner distribution.

★ In this case we have

$$x(t+\tau/2)x^*(t-\tau/2) = |A|^2 e^{jbt\tau}$$

with

$$WD(t,\Omega) = 2\pi|A|^2\delta(\Omega - bt).$$

Again, a high concentration along the instantaneous frequency in the time-frequency plane may be achieved for the linear frequency modulated signals. □

These two examples demonstrate that the Wigner distribution can provide superior time-frequency representation of one-component signal, in comparison to the STFT.

Example 11.3. Calculate the Wigner distribution for a linear frequency modulated signal, with Gaussian amplitude (Gaussian chirp signal)

$$x(t) = Ae^{-at^2/2}e^{j(bt^2/2+ct)}.$$

★ For the chirp signal, the local autocorrelation function reads as

$$R(t,\tau) = x(t+\tau/2)x^*(t-\tau/2) = |A|^2 e^{-at^2}e^{-a\tau^2/4}e^{jbt\tau+jc\tau}.$$

The Wigner distribution is obtained as the Fourier transform of $R(t,\tau)$,

$$WD(t,\Omega) = 2|A|^2 e^{-at^2}\sqrt{\frac{\pi}{a}}e^{-\frac{(\Omega-bt-c)^2}{a}}. \tag{11.5}$$

The Wigner distribution from the previous example is obtained with $c = 0$ and $a \to 0$, since $2\sqrt{\pi/a}e^{-\Omega^2/a} \to 2\pi\delta(\Omega)$ as $a \to 0$.

The Wigner distribution of the Gaussian chirp signal is always positive, as it could be expected from a distribution introduced with the aim to represent local density of signal energy. Unfortunately, this is the only signal when the Wigner distribution is always positive, for any point in the time-frequency plane (t,Ω). This drawback is not the only reason why the study of time-frequency distributions does not end with the Wigner distribution. □

11.1.1 Auto-Terms and Cross-Terms in the Wigner Distribution

For the multi-component signal

$$x(t) = \sum_{m=1}^{M} x_m(t)$$

the Wigner distribution has the form

$$WD(t,\Omega) = \sum_{m=1}^{M}\sum_{n=1}^{M}\int_{-\infty}^{\infty} x_m\left(t+\frac{\tau}{2}\right)x_n^*\left(t-\frac{\tau}{2}\right)e^{-j\Omega\tau}d\tau.$$

Besides the auto-terms

$$WD_{at}(t,\Omega) = \sum_{m=1}^{M}\int_{-\infty}^{\infty} x_m(t+\frac{\tau}{2})x_m^*(t-\frac{\tau}{2})e^{-j\Omega\tau}d\tau,$$

the Wigner distribution contains a significant number of cross-terms,

$$WD_{ct}(t,\Omega) = \sum_{m=1}^{M}\sum_{\substack{n=1\\ n\neq m}}^{M}\int_{-\infty}^{\infty} x_m(t+\frac{\tau}{2})x_n^*(t-\frac{\tau}{2})e^{-j\Omega\tau}d\tau.$$

Usually, they are not desirable in the time-frequency signal analysis. Cross-terms can mask the presence of auto-terms, which makes the Wigner distribution unsuitable for the time-frequency analysis of signals.

For a two-component signal with auto-terms located around (t_1,Ω_1) and (t_2,Ω_2) (see Fig.11.3) the oscillatory cross-terms are located around $((t_1+t_2)/2,(\Omega_1+\Omega_2)/2)$.

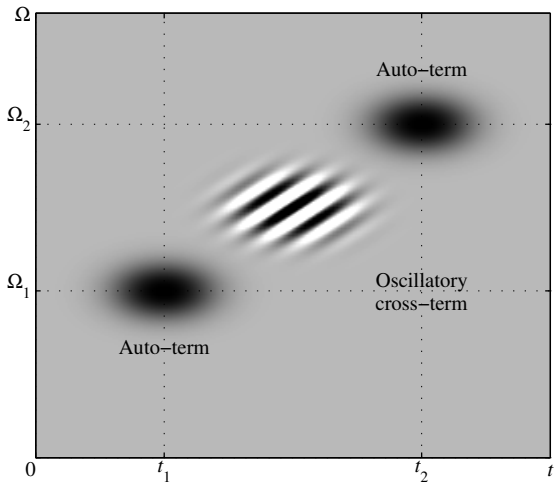


Figure 11.3 Wigner distribution of two component signal.

Example 11.4. Analyze auto-terms and cross-terms for two-component signal of the form

$$x(t) = e^{-\frac{1}{2}(t-t_1)^2 e^{j\Omega_1 t}} + e^{-\frac{1}{2}(t+t_1)^2 e^{-j\Omega_1 t}}$$

★ In this case we have

$$WD(t,\Omega) = 2\sqrt{\pi}e^{-(t-t_1)^2-(\Omega-\Omega_1)^2} + 2\sqrt{\pi}e^{-(t+t_1)^2-(\Omega+\Omega_1)^2} + 4\sqrt{\pi}e^{-t^2-\Omega^2}\cos(2t_1\Omega - 2\Omega_1 t)$$

where the first and second terms represent auto-terms while the third term is a cross-term. Note that the cross-term is oscillatory in both directions. The oscillation rate along the time axis is proportional to the frequency distance between components $2\Omega_1$, while the oscillation rate along frequency axis is proportional to the distance in time of components, $2t_1$. The oscillatory nature of cross-terms will be used for their suppression.

To analyze auto-terms and cross-terms, the well-known ambiguity function can be used as well. It is defined as:

$$AF(\theta,\tau) = \int_{-\infty}^{\infty} x\left(t+\frac{\tau}{2}\right)x^*\left(t-\frac{\tau}{2}\right)e^{-j\theta t}dt. \tag{11.6}$$

It is already a classical tool in optics as well as in radar and sonar signal analysis.

The ambiguity function and the Wigner distribution form a two-dimensional Fourier transform pair

$$AF(\theta,\tau) = \mathrm{FT}^{2D}_{t,\Omega}\{WD(t,\Omega)\},$$

$$WD(t,\Omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} x(u+\frac{\tau}{2})x^*(u-\frac{\tau}{2})e^{-j\theta u}du\right]e^{j\theta t-j\Omega\tau}d\tau d\theta,$$

where the integration over frequency related variable θ assumes factor $1/(2\pi)$ and the positive sign in the exponent $\exp(j\theta t)$.

Consider a signal whose components are limited in time to

$$x_m(t) \neq 0 \qquad \text{only for} \quad |t-t_m| < T_m.$$

In the ambiguity (θ,τ) domain we have $x_m(t+\tau/2)x_m^*(t-\tau/2) \neq 0$ only for

$$-T_m < t - t_m + \tau/2 < T_m$$
$$-T_m < t - t_m - \tau/2 < T_m.$$

It means that $x_m(t+\tau/2)x_m^*(t-\tau/2)$ is located within $|\tau| < 2T_m$, that is, around the θ-axis independently of the signal's position t_m. Cross-term between signal's m-th and n-th component is located within $|\tau + t_n - t_m| < T_m + T_n$. It is dislocated from $\tau = 0$ for two components that do not occur simultaneously, that is, when $t_m \neq t_n$.

From the frequency domain definition of the Wigner distribution a corresponding ambiguity function form follows

$$AF(\theta,\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X\left(\Omega+\frac{\theta}{2}\right)X^*\left(\Omega-\frac{\theta}{2}\right)e^{j\Omega\tau}d\Omega. \tag{11.7}$$

From this form we can conclude that the auto-terms of the components, limited in frequency to $X_m(\Omega) \neq 0$ only for $|\Omega - \Omega_m| < W_m$, are located in the ambiguity domain around τ-axis within the

region $|\theta/2| < W_m$. The cross-terms are within

$$|\theta + \Omega_n - \Omega_m| < W_m + W_n,$$

where Ω_m and Ω_n are the frequencies around which the Fourier transform of each component lies.

Therefore, all auto-terms are located along and around the ambiguity domain axis. The cross-terms, for the components which do not overlap in the time and frequency, simultaneously, are dislocated from the ambiguity axes, Fig. 11.4. This property will be used in the definition of the reduced interference time-frequency distributions.

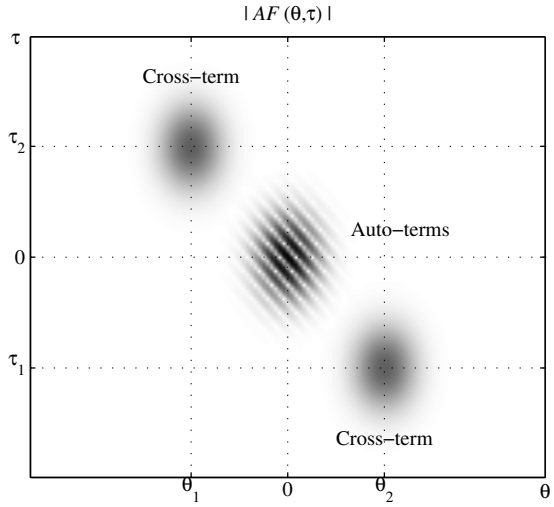


Figure 11.4 Auto and cross-terms for two-component signal in the ambiguity domain.

The ambiguity function of a four-component signal consisting of two Gaussian pulses, one sinusoidal and one linear frequency modulated component is presented in 11.5.

Example 11.5. Let us consider signals of the form

$$x_1(t) = e^{-\frac{1}{2}t^2} \quad \text{and} \quad x_2(t) = e^{-\frac{1}{2}(t-t_1)^2 e^{j\Omega_1 t}} + e^{-\frac{1}{2}(t+t_1)^2 e^{-j\Omega_1 t}}.$$

The ambiguity function of $x_1(t)$ is

$$AF_{x_1}(\theta,\tau) = \sqrt{\pi} e^{-\frac{1}{4}\tau^2 - \frac{1}{4}\theta^2}$$

while the ambiguity function of two-component signal $x_2(t)$ is

$$AF_{x_2}(\theta,\tau) = \sqrt{\pi} e^{-\frac{1}{4}\tau^2 - \frac{1}{4}\theta^2} e^{j\Omega_1\tau} e^{-jt_1\theta} + \sqrt{\pi} e^{-\frac{1}{4}\tau^2 - \frac{1}{4}\theta^2} e^{-j\Omega_1\tau} e^{jt_1\theta} + \sqrt{\pi} e^{-\frac{1}{4}(\tau-2t_1)^2 - \frac{1}{4}(\theta-2\Omega_1)^2} + \sqrt{\pi} e^{-\frac{1}{4}(\tau+2t_1)^2 - \frac{1}{4}(\theta+2\Omega_1)^2}$$

In the ambiguity domain (θ,τ) auto-terms are located around $(0,0)$ while cross-terms are located around $(2\Omega_1, 2t_1)$ and $(-2\Omega_1, -2t_1)$ as presented in Fig. 11.4.

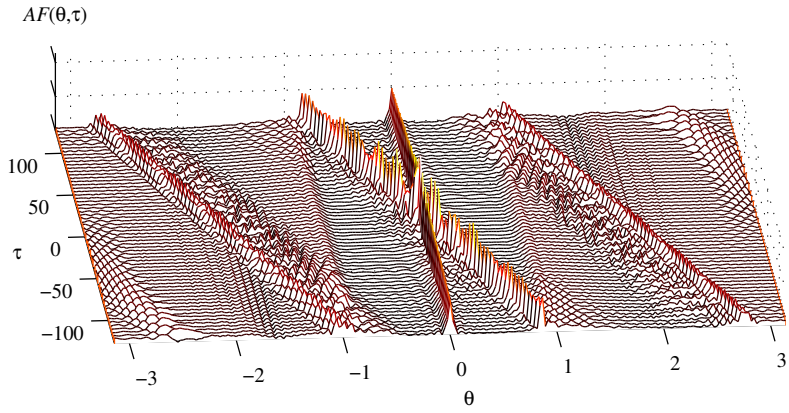


Figure 11.5 Ambiguity function of signal from Fig.10.4

11.1.2 Wigner Distribution Properties

A list of the properties satisfied by the Wigner distribution follows. The obvious ones will be just stated, while the proofs will be given for more complex ones. In the case when the Wigner distributions of more than one signal are considered, the signal will be added as an index in the Wigner distribution notation. Otherwise signal $x(t)$ is assumed, as a default signal in the notation.

P_1 – Realness

For any signal holds,

$$WD^*(t,\Omega) = WD(t,\Omega).$$

P_2 – Time-shift property

The Wigner distribution of a signal shifted in time

$$y(t) = x(t-t_0),$$

is

$$WD_y(t,\Omega) = WD_x(t-t_0,\Omega).$$

P_3 – Frequency shift property

For a modulated signal

$$y(t) = x(t)e^{j\Omega_0 t},$$

we have

$$WD_y(t,\Omega) = WD_x(t,\Omega-\Omega_0).$$

P_4 – Time marginal property

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} WD(t,\Omega)d\Omega = |x(t)|^2.$$

P_5 – Frequency marginal property

$$\int_{-\infty}^{\infty} WD(t,\Omega)dt = |X(\Omega)|^2.$$

P_6 – Time moments property

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} t^n WD(t,\Omega)dtd\Omega = \int_{-\infty}^{\infty} t^n|x(t)|^2dt.$$

This property follows from $\frac{1}{2\pi}\int_{-\infty}^{\infty} WD(t,\Omega)d\Omega = |x(t)|^2$. □

P_7 -Frequency moments property

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Omega^n WD(t,\Omega)d\Omega dt = \int_{-\infty}^{\infty} \Omega^n|X(\Omega)|^2d\Omega.$$

P_8 – Scaling

For a scaled version of the signal

$$y(t) = \sqrt{|a|}x(at),\ a \neq 0,$$

the Wigner distribution reads

$$WD_y(t,\Omega) = WD_x(at,\Omega/a).$$

P_9 – Instantaneous frequency property

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

$$\frac{\int_{-\infty}^{\infty}\Omega WD(t,\Omega)\,d\Omega}{\int_{-\infty}^{\infty} WD(t,\Omega)\,d\Omega} = \Omega_i(t) = \frac{d}{dt}\arg[x(t)] = \phi'(t). \tag{11.8}$$

In order to prove this property, we will use the derivative of the inverse Fourier transform of the Wigner distribution

$$\frac{d[x(t+\tau/2)x^*(t-\tau/2)]}{d\tau} = \frac{1}{2\pi}\int_{-\infty}^{\infty} j\Omega\, WD(t,\Omega)e^{j\Omega\tau}d\Omega$$

with $x(t) = A(t)e^{j\phi(t)}$, calculated at $\tau = 0$. It results in

$$\frac{j}{2\pi}\int_{-\infty}^{\infty}\Omega WD(t,\Omega)\,d\Omega = \frac{1}{2}[x'(t)x^*(t) - x(t)x^{*\prime}(t)] = j\phi'(t)A^2(t).$$

With the frequency marginal property $\int_{-\infty}^{\infty} WD(t,\Omega)\,d\Omega = 2\pi A^2(t)$, this property follows. □

P_{10} – Group delay

For signal whose Fourier transform is of the form $X(\Omega) = |X(\Omega)|\,e^{j\Phi(\Omega)}$, the group delay $t_g(\Omega) = -\Phi'(\Omega)$ is

$$\frac{\int_{-\infty}^{\infty} t\,WD(t,\Omega)\,dt}{\int_{-\infty}^{\infty} WD(t,\Omega)\,dt} = t_g(\Omega) = -\frac{d}{d\Omega}\arg[X(\Omega)] = -\Phi'(\Omega).$$

The proof is the same as in the instantaneous frequency case, using the frequency domain relations.

P_{11} – Time constraint

If $x(t) = 0$ for t outside $[t_1,t_2]$, then $WD(t,\Omega) = 0$ for t outside $[t_1,t_2]$.

The Wigner distribution is a function of $x(t+\tau/2)x^*(t-\tau/2)$. If $x(t)=0$ for t outside $[t_1,t_2]$ then $x(t+\tau/2)x^*(t-\tau/2)$ is different from zero within

$$t_1 \le t+\tau/2 \le t_2 \quad \text{and} \quad t_1 \le t-\tau/2 \le t_2.$$

The range of values of t defined by the previous inequalities is $t_1 \le t \le t_2$. □

P_{12} – Frequency constraint

If $X(\Omega)=0$ for Ω outside $[\Omega_1,\Omega_2]$, then, also $WD(t,\Omega)=0$ for Ω outside $[\Omega_1,\Omega_2]$.

P_{13} – Convolution

$$WD_y(t,\Omega)=\int_{-\infty}^{\infty} WD_h(t-\tau,\Omega)WD_x(\tau,\Omega)d\tau.$$

for

$$y(t)=\int_{-\infty}^{\infty} h(t-\tau)x(\tau)d\tau,$$

P_{14} – Product

$$WD_y(t,\Omega)=\frac{1}{2\pi}\int_{-\infty}^{\infty} WD_h(t,\Omega-v)WD_x(t,v)dv$$

for

$$y(t)=h(t)x(t).$$

The local autocorrelation of $y(t)$ is $h(t+\tau/2)h^*(t-\tau/2)x(t+\tau/2)x^*(t-\tau/2)$. Thus, the Wigner distribution of $y(t)$ is the Fourier transform of the product of local autocorrelations $h(t+\tau/2)h^*(t-\tau/2)$ and $x(t+\tau/2)x^*(t-\tau/2)$. It is a convolution in frequency of the corresponding Wigner distributions of $h(t)$ and $x(t)$. Property P_{13} could be proven in the same way using the Fourier transforms of signals $h(t)$ and $x(t)$. □

P_{15} – Fourier transform property

$$WD_y(t,\Omega)=WD_x(-\Omega/c,ct) \tag{11.9}$$

for

$$y(t)=\sqrt{|c|/(2\pi)}X(ct),\ c\neq 0.$$

Here the signal $y(t)$ is equal to the scaled version of the Fourier transform of signal $x(t)$,

$$\begin{aligned} WD_y(t,\Omega) &= \frac{|c|}{2\pi}\int_{-\infty}^{\infty} X\left(ct+\frac{c\tau}{2}\right)X^*\left(ct-\frac{c\tau}{2}\right)e^{-j\Omega\tau}d\tau \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} X\left(ct+\frac{\theta}{2}\right)X^*\left(ct-\frac{\theta}{2}\right)e^{j(-\Omega/c)\theta}d\theta. \end{aligned} \tag{11.10}$$

Comparing (11.2) to (11.1), with $ct\to\Omega$ and $(-\Omega/c)\to t$, we get

$$WD_y(t,\Omega)=\int_{-\infty}^{\infty} x\left(-\frac{\Omega}{c}+\frac{\tau}{2}\right)x^*\left(-\frac{\Omega}{c}-\frac{\tau}{2}\right)e^{-jct\tau}d\tau = WD_x\left(-\frac{\Omega}{c},ct\right).$$

□

P_{16} – Chirp convolution

$$WD_y(t,\Omega) = WD_x\left(t - \frac{\Omega}{c}, \Omega\right) \tag{11.11}$$

for

$$y(t) = x(t) * \sqrt{|c|}e^{jct^2/2}.$$

With $Y(\Omega) = \mathrm{FT}\{x(t) *_t \sqrt{|c|}e^{jct^2/2}\} = \sqrt{2\pi j}X(\Omega)e^{-j\Omega^2/(2c)}$ and the signal's Fourier transform-based definition of the Wigner distribution, proof of this property reduces to the next one. □

P_{17} – Chirp product

$$WD_y(t,\Omega) = WD_x(t,\Omega - ct)$$

for

$$y(t) = x(t)e^{jct^2/2}.$$

The Wigner distribution of $y(t)$ is

$$\begin{aligned}
&\int_{-\infty}^{\infty} x\left(t + \frac{\tau}{2}\right) e^{jc(t+\tau/2)^2/2} x^*\left(t - \frac{\tau}{2}\right) e^{-jc(t-\tau/2)^2/2} e^{-j\Omega\tau} d\tau \\
&= \int_{-\infty}^{\infty} x\left(t + \frac{\tau}{2}\right) x^*\left(t - \frac{\tau}{2}\right) e^{jct\tau} e^{-j\Omega\tau} d\tau = WD_x(t,\Omega - ct).
\end{aligned} \tag{11.12}$$

□

P_{18} – Moyal property

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} WD_x(t,\Omega)WD_y(t,\Omega)dtd\Omega = \left|\int_{-\infty}^{\infty} x(t)y(t)dt\right|^2. \tag{11.13}$$

This property follows from

$$\begin{aligned}
&\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\left(t + \frac{\tau_1}{2}\right) x^*\left(t - \frac{\tau_1}{2}\right) y\left(t + \frac{\tau_2}{2}\right) y^*\left(t - \frac{\tau_2}{2}\right) \int_{-\infty}^{\infty} e^{-j\Omega\tau_1} e^{-j\Omega\tau_2} d\Omega \; d\tau_1 d\tau_2 dt \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\left(t + \frac{\tau}{2}\right) x^*\left(t - \frac{\tau}{2}\right) y\left(t - \frac{\tau}{2}\right) y^*\left(t + \frac{\tau}{2}\right) d\tau dt.
\end{aligned}$$

With $t + \tau/2 = u$ and $t - \tau/2 = v$, we get

$$\int_{-\infty}^{\infty} x(u)\, y^*(u) du \int_{-\infty}^{\infty} x^*(v)\, y(v)\, dv = \left|\int_{-\infty}^{\infty} x(t)y(t)dt\right|^2.$$

11.1.3 Pseudo and Smoothed Wigner Distribution

In practical realizations of the Wigner distribution, we are constrained with a finite time lag τ. A pseudo form of the Wigner distribution is then used. It is defined as

$$PWD(t,\Omega) = \int_{-\infty}^{\infty} w(\tau/2)w^*(-\tau/2)x(t+\tau/2)x^*(t-\tau/2)e^{-j\Omega\tau}d\tau \tag{11.14}$$

where window $w(\tau)$ localizes the considered lag interval. If $w(0) = 1$, the pseudo Wigner distribution satisfies the time marginal property. Note that the pseudo Wigner distribution is smoothed in the frequency direction with respect to the Wigner distribution

$$PWD(t,\Omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} WD(t,\theta)W_e(\Omega-\theta)d\theta$$

where $W_e(\Omega)$ is a Fourier transform of $w(\tau/2)w^*(-\tau/2)$.

The pseudo Wigner distribution example for multi-component signals is presented in Fig.11.6. In the case of multi-component signal the cross-terms between components are emphatic.

Mono-component case with sinusoidally frequency modulated signal is presented in Fig.11.7. Note that significant inner interferences are present.

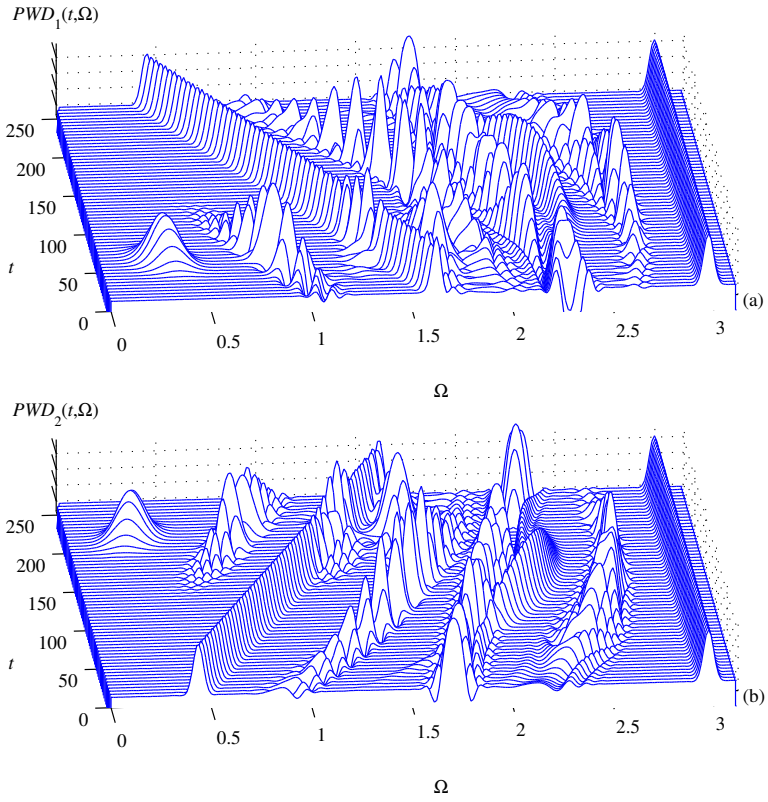


Figure 11.6 Pseudo Wigner distribution of the signals from Fig.10.4

Monocomponent case with sinusoidally frequency modulated signal is presented in Fig.11.7. Note that significant inner interferences are present.

Example 11.6. For a sinusoidally frequency modulated signal

$$x(t) = \exp(-j128\cos(\pi t/64))$$

calculate an approximate value of the pseudo Wigner distribution with a window $w(\tau)$ of the width defined by $T = 8$.

★ The pseudo Wigner distribution of this signal is

$$PWD(\Omega,t) = \int_{-8}^{8} e^{-j128\cos(\pi(t+\tau/2)/64)} e^{j128\cos(\pi(t-\tau/2)/64)} w(\tau) e^{-j\Omega\tau} d\tau.$$

By using the Taylor expansion with respect to a relatively small τ

$$\cos\left(\pi t/64 \pm \frac{\pi\tau}{128}\right) = \cos(\pi t/64) \mp \frac{\pi}{128}\sin(\pi t/64)\tau$$
$$-\left(\frac{\pi}{128}\right)^2 \cos(\pi t/64)\frac{\tau^2}{2} + \left(\frac{\pi}{128}\right)^3 \sin(\pi t/64)\frac{(\mp\tau_{1,2}^3)}{6},$$

with $|\tau_{1,2}| \leq 8$ in the Taylor series reminder, we get

$$WD(\Omega,t) = \int_{-8}^{8} e^{j2\pi\sin(\pi t/64)\tau} e^{j128\frac{\pi^3}{128^3}\sin(\pi t/64)\frac{\tau_1^3+\tau_2^3}{6}} w(\tau) e^{-j\Omega\tau} d\tau.$$

For $|\tau_{1,2}| \leq 8$ it holds $\left|128\frac{\pi^3}{128^3}\sin(\pi t/64)\frac{\tau_1^3+\tau_2^3}{6}\right| \leq 0.33$. By neglecting this term we may write

$$PWD(\Omega,t) \cong W(\Omega - 2\pi\sin(\pi t/64)),$$

where $W(\Omega)$ is the Fourier transform of window $w(\tau)$ Fig.11.7(a) (with a Hann(ing) window). For a wider window this approximation does not hold and the inner interferences in the Wigner distribution appear, Fig.11.7(b) (with a four times wider Hann(ing) window).

11.1.4 Discrete Pseudo Wigner Distribution

If the signal in (11.14) is discretized in τ with a sampling interval Δt, then a sum instead of an integral is formed. The pseudo Wigner distribution of a discrete-lag signal, for a given time instant t, is given by

$$PWD(t,\Omega) = \sum_{m=-\infty}^{\infty} w\left(m\tfrac{\Delta t}{2}\right) w^*\left(-m\tfrac{\Delta t}{2}\right) x\left(t+m\tfrac{\Delta t}{2}\right) x^*\left(t-m\tfrac{\Delta t}{2}\right) e^{-jm\Omega\Delta t}\Delta t. \qquad (11.15)$$

Sampling in τ with $\Delta t = \pi/\Omega_0$, $\Omega_0 > \Omega_m$ corresponds to the sampling of signal $x(t+\tau/2)$ in $\tau/2$ with $\Delta t/2 = \pi/(2\Omega_0)$.

The discrete-lag pseudo Wigner distribution is the Fourier transform of signal

$$R(t,m) = w\left(m\frac{\Delta t}{2}\right) w^*\left(-m\frac{\Delta t}{2}\right) x\left(t+m\frac{\Delta t}{2}\right) x^*\left(t-m\frac{\Delta t}{2}\right)\Delta t.$$

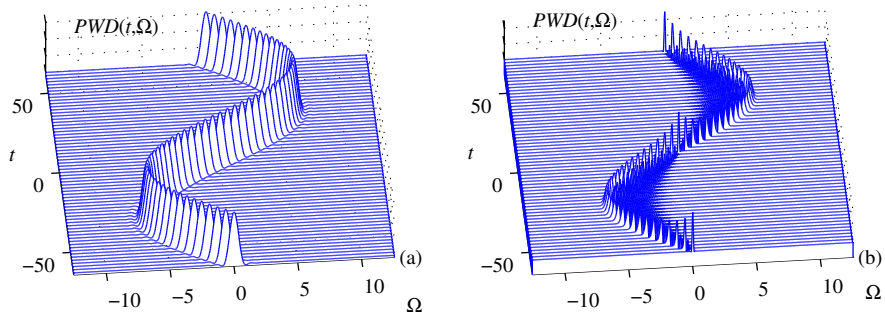


Figure 11.7 Pseudo Wigner distribution of sinusoidally frequency modulated signal. Narrow Hann(ing) window (left) and a four times wider window (right).

For a given instant t, it can be written as

$$PWD(t,\omega) = \sum_{m=-\infty}^{\infty} R(t,m)e^{-jm\omega}$$

with $\omega = \Omega\Delta t$. If the sampling interval satisfies the sampling theorem, then the sum in (11.15) is equal to the integral form (11.14). A discrete form of the pseudo Wigner distribution, with $N+1$ samples and $\omega = 2\pi k/(N+1)$, for a given time instant t, is

$$PWD(t,k) = \sum_{m=-N/2}^{N/2} R(t,m)e^{-j2\pi mk/(N+1)}.$$

Here, $N/2$ is an integer. This distribution could be calculated using the standard DFT routines.

For discrete-time instants $t = n\Delta t$, introducing the notation

$$R(n\Delta t, m\Delta t) = w\left(m\frac{\Delta t}{2}\right)w^*\left(-m\frac{\Delta t}{2}\right)x\left(n\Delta t + m\frac{\Delta t}{2}\right)x^*\left(n\Delta t - m\frac{\Delta t}{2}\right)\Delta t$$
$$R(n,m) = w\left(\frac{m}{2}\right)w^*\left(-\frac{m}{2}\right)x\left(n+\frac{m}{2}\right)x^*\left(n-\frac{m}{2}\right),$$

the discrete-time and discrete-lag pseudo Wigner distribution can be written as

$$PWD(n,\omega) = \sum_{m=-\infty}^{\infty} w\left(\frac{m}{2}\right)w^*\left(-\frac{m}{2}\right)x\left(n+\frac{m}{2}\right)x^*\left(n-\frac{m}{2}\right)e^{-jm\omega}. \tag{11.16}$$

Notation $x(n+m/2)$, for given n and m, should be understood as the signal value at the instant $x((n+m/2)\Delta t)$. In this notation, the discrete-time pseudo Wigner distribution is periodic in ω with period 2π.

Since various discretization steps are used (here and in open literature), we will provide a relation of discrete indexes to the continuous time and frequency, for each definition, as

$$PWD(t,\Omega)|_{t=n\Delta t,\ \Omega=\frac{2\pi k}{(N+1)\Delta t}} = PWD\left(n\Delta t, \frac{2\pi k}{(N+1)\Delta t}\right) \rightarrow PWD(n,k).$$

The sign $\rightarrow$ could be understood as the equality sign in the sense of sampling theorem (Example 2.13). Otherwise, it should be considered as a correspondence sign. The discrete form of (11.14), with $N+1$

samples, is

$$PWD\left(n\Delta t, \frac{2\pi k}{(N+1)\Delta t}\right) \to PWD(n,k)$$

$$PWD(n,k) = \sum_{m=-N/2}^{N/2} w\left(\frac{m}{2}\right) w^*\left(-\frac{m}{2}\right) x\left(n+\frac{m}{2}\right) x^*\left(n-\frac{m}{2}\right) e^{-j2\pi km/(N+1)},$$

where $N/2$ is an integer, $-N/2 \le k \le N/2$ and $\omega = \Omega\Delta t = 2\pi k/(N+1)$ or $\Omega = 2\pi k/((N+1)\Delta t)$.

In order to avoid different sampling intervals in time and lag in the discrete Wigner distribution definition, the discrete Wigner distribution can be oversampled in time, as it has been done in lag. It means that the same sampling interval $\Delta t/2$, for both time and lag axes, can be used. Then, we can write

$$R\left(n\frac{\Delta t}{2}, m\Delta t\right) \to R(n,m)$$

$$R\left(n\frac{\Delta t}{2}, m\Delta t\right) = w\left(m\frac{\Delta t}{2}\right) w^*\left(-m\frac{\Delta t}{2}\right) x\left(n\frac{\Delta t}{2}+m\frac{\Delta t}{2}\right) x^*\left(n\frac{\Delta t}{2}-m\frac{\Delta t}{2}\right)\Delta t$$

$$R(n,m) = w(m)w^*(-m)x(n+m)x^*(n-m)$$

The discrete-time and discrete-lag pseudo Wigner distribution, in this case, is of the form

$$PWD(n,\omega) = 2\sum_{m=-\infty}^{\infty} w(m)w^*(-m)x(n+m)x^*(n-m)e^{-j2m\omega}. \tag{11.17}$$

It corresponds to the continuous-time pseudo Wigner distribution (11.14) with substitution $\tau/2 \to \tau$

$$PWD(t,\Omega) = 2\int_{-\infty}^{\infty} w(\tau)w^*(-\tau)x(t+\tau)x^*(t-\tau)e^{-j2\Omega\tau}d\tau.$$

The discrete pseudo Wigner distribution is given here by

$$PWD\left(\frac{n\Delta t}{2}, \frac{4\pi k}{(N+1)\Delta t}\right) \to PWD(n,k)$$

$$PWD(n,k) = \sum_{m=-N/2}^{N/2} w(m)w^*(-m)x(n+m)x^*(n-m)e^{-j4\pi mk/(N+1)} \tag{11.18}$$

for $-N/2 \le 2k \le N/2$. Since, the standard DFT routines are commonly used for the pseudo Wigner distribution calculation, we may use every other ($2k$) sample in (11.18) or oversample the pseudo Wigner distribution in frequency (as it has been done in time). Then,

$$PWD\left(\frac{n\Delta t}{2}, \frac{2\pi k}{(N+1)\Delta t}\right) \to PWD(n,k)$$

$$PWD(n,k) = \sum_{m=-N/2}^{N/2} w(m)w^*(-m)x(n+m)x^*(n-m)e^{-j2\pi mk/(N+1)}. \tag{11.19}$$

This discrete pseudo Wigner distribution, oversampled in both time and in frequency by factor of 2, has finer time-frequency grid, producing smaller time-frequency estimation errors at the expense of the calculation complexity.

Example 11.7. Signal $x(t) = \exp(j31\pi t^2)$ is considered within $-1 \leq t \leq 1$. Find the sampling interval of signal for discrete pseudo Wigner distribution calculation. If the rectangular window of the width $N+1 = 31$ is used in analysis, find the pseudo Wigner distribution values and estimate the instantaneous frequency at $t = 0.5$ based on the discrete pseudo Wigner distribution.

★ For this signal the instantaneous frequency is $\Omega_i(t) = 62\pi t$. It is within the range $-62\pi \leq \Omega_i(t) \leq 62\pi$. Thus, we may approximately assume that the maximum frequency is $\Omega_m = 62\pi$.The sampling interval for the Fourier transform would be $\Delta t \leq 1/62$. For the direct pseudo Wigner distribution calculation, it should be twice smaller, $\Delta t/2 \leq 1/124$. Therefore, the discrete version of the pseudo Wigner distribution, normalized with $2\sqrt{\Delta t}$, at $t = 0.5$ or $n = 62$, is (11.18)

$$PWD(n,k) = \sum_{m=-15}^{15} e^{j31\pi((n+m)/124)^2} e^{-j31\pi((n-m)/124)^2} e^{-j4\pi mk/31}$$
$$= \sum_{m=-15}^{15} e^{j\pi mn/124} e^{-j4\pi mk/31} = \frac{\sin(\frac{\pi}{8}(n-16k))}{\sin(\frac{\pi}{248}(n-16k))}.$$

The argument k, when the pseudo Wigner distribution reaches maximum for $n = 62$, follows from $62 - 16k = 0$ as

$$\hat{k} = \arg\left\{\max_k PWD(n,k)\right\} = \left[\frac{62}{16}\right] = 4,$$

where $[\cdot]$ stands for the nearest integer. Obviously, the exact instantaneous frequency is not on the discrete frequency grid. The estimated value of the instantaneous frequency at $t = 1/2$ is

$$\hat{\Omega} = 4\pi\hat{k}/((N+1)\Delta t) = 16\pi/(31/62) = 32\pi.$$

The true value is $\Omega_i(1/2) = 31\pi$. When the true frequency is not on the grid, the estimation can be improved using the interpolation or displacement bin, as explained in Chapter 1. The frequency sampling interval is $\Delta\Omega = 4\pi/((N+1)\Delta t) = 8\pi$, with maximum estimation absolute error $\Delta\Omega/2 = 4\pi$.

If we used the standard DFT routine (11.19) with $N+1 = 31$ and all available frequency samples, we would get

$$PWD(n,k) = \text{DFT}_{31}\left\{e^{j31\pi((n+m)/124)^2} e^{-j31\pi((n-m)/124)^2}\right\}$$
$$= \sum_{m=-15}^{15} e^{j31\pi((n+m)/124)^2} e^{-j31\pi((n-m)/124)^2} e^{-j2\pi mk/31} = \frac{\sin(\frac{\pi}{8}(n-8k))}{\sin(\frac{\pi}{248}(n-8k))}.$$

The maximum would be at $\hat{k} = 8$, with the estimated frequency $\hat{\Omega} = 2\pi\hat{k}/((N+1)\Delta t)$. Thus, $\hat{\Omega} = 32\pi$, as expected. By this calculation, the frequency sampling interval is $\Delta\Omega = 2\pi/((N+1)\Delta t) = 4\pi$, with the maximum estimation absolute error $\Delta\Omega/2 = 2\pi$. □

Using an odd number of samples $N+1$ in the previous definitions, the symmetry of the product $x(n+m)x^*(n-m)$ is preserved in the summation. However, when an even number of samples is used, that is not the case. To illustrate this effect, consider a simple example of signal, for $n = 0$, with $N = 4$ samples. Then, four values of the signal $x(m)$, used in calculation, are

$x(m)$	$x(-2)$	$x(-1)$	$x(0)$	$x(1)$	
$x(-m)$		$x(1)$	$x(0)$	$x(-1)$	$x(-2)$

.

So, in forming the local autocorrelation function, there are several possibilities. One is to omit sample $x(-2)$ and to use an odd number of samples, in this case as well. Also, it is possible to periodically extend the signal and to form the product based on

$x(m)$	$\cdots$	$x(1)$	$x(-2)$	$x(-1)$	$x(0)$	$x(1)$	$x(-2)$	$x(-1)$
$x(-m)$	$\cdots$	$x(-1)$	$x(-2)$	$x(1)$	$x(0)$	$x(-1)$	$x(-2)$	$x(1)$
$w_e(m)$	$\cdots$	0	0	$w_e(1)$	$w_0(0)$	$w_e(1)$	0	0

Here, we can use four product terms, but with the first one formed as $x(-2)x^*(-2)$, that is, as $x(-N/2)x^*(-N/2)$. When a lag window with zero ending value is used (for example, a Hann(ing) window), this term does not make any influence to the result. The used lag window must also follow the symmetry, for example $w_e(m) = \cos^2(\pi m/N)$, when,

$$PWD\left(\frac{n\Delta t}{2}, \frac{2\pi k}{N\Delta t}\right) \to PWD(n,k)$$

$$PWD(n,k) = \sum_{m=-N/2}^{N/2-1} w_e(m)x(n+m)x^*(n-m)e^{-j2\pi mk/N}$$

$$= \sum_{m=-N/2+1}^{N/2-1} w_e(m)x(n+m)x^*(n-m)e^{-j2\pi mk/N},$$

since $w_e(-N/2) = 0$. However, if the window is nonzero at the ending point $m = -N/2$, this term will result in a kind of aliased distribution.

In order to introduce another way of the discrete Wigner distribution calculation, with an even number of samples, consider again the continuous form of the Wigner distribution of a signal with a limited duration. Assume that the signal is sampled in such a way that the sampling theorem can be applied and the equality sign used (Example 2.13). Then, the integral may be replaced by a sum

$$WD(t,\Omega) = \sum_{m=-N}^{N} x(t+m\tfrac{\Delta t}{2})x^*(t-m\tfrac{\Delta t}{2})e^{-jm\Omega\Delta t}\Delta t$$

$$= \sum_{m=-N/2}^{N/2} x(t+2m\tfrac{\Delta t}{2})x^*(t-2m\tfrac{\Delta t}{2})e^{-j2m\Omega\Delta t}\Delta t$$

$$+ \sum_{m=-N/2}^{N/2-1} x(t+(2m+1)\tfrac{\Delta t}{2})x^*(t-(2m+1)\tfrac{\Delta t}{2})e^{-j(2m+1)\Omega\Delta t}\Delta t. \qquad (11.20)$$

The initial sum is split into its even and odd terms part. Now, let us assume that the signal is sampled in such a way that twice wider sampling interval Δt is also sufficient to obtain the Wigner distribution (by using every other signal sample). Then, for the first sum (with an odd number of samples) holds,

$$\sum_{m=-N/2}^{N/2} x(t+m\Delta t)x^*(t-m\Delta t)e^{-j2m\Omega\Delta t}\Delta t = \frac{1}{2}WD(t,\Omega).$$

The factor $1/2$ comes from the sampling interval. Now, from (11.20) follows

$$\sum_{m=-N/2}^{N/2-1} x(t+(2m+1)\tfrac{\Delta t}{2})x^*(t-(2m+1)\tfrac{\Delta t}{2})e^{-j(2m+1)\Omega\Delta t}\Delta t = \frac{1}{2}WD(t,\Omega). \qquad (11.21)$$

This is just the discrete Wigner distribution with an even number of samples. If we denote

$$x(t+(2m+1)\tfrac{\Delta t}{2}) = x(t+m\Delta t+\tfrac{\Delta t}{2}) = x_e(t+m\Delta t)$$
$$x(n\Delta t+m\Delta t+\tfrac{\Delta t}{2})\sqrt{2\Delta t} = x_e(n+m)$$

then

$$x(t-m\Delta t-\tfrac{\Delta t}{2}) = x(t-m\Delta t+\tfrac{\Delta t}{2}-\Delta t)$$
$$x(n\Delta t-m\Delta t+\tfrac{\Delta t}{2}-\Delta t)\sqrt{2\Delta t} = x_e(n-m-1).$$

The summation terms, for example for $n=0$, are of the form

$x_e(m)$	$\dots$	$x_e(-2)$	$x_e(-1)$	$x_e(0)$	$x_e(1)$	$\dots$
$x_e(-m-1)$	$\dots$	$x_e(1)$	$x_e(0)$	$x_e(-1)$	$x_e(-2)$	$\dots$

.

They would produce a modulated version of the pseudo Wigner distribution, due to the shift of a half of the sampling interval. However, this shift can be corrected as (11.21)

$$WD(t,\Omega) = e^{-j\Omega\Delta t}\sum_{m=-N/2}^{N/2-1} x_e(t+m\Delta t)x_e^*(t-m\Delta t-\Delta t)e^{-j2m\Omega\Delta t}(2\Delta t)$$

for any t and Ω (having in mind the sampling theorem). Thus, we may also write

$$WD\left(n\Delta t, \frac{\pi k}{N\Delta t}\right) \to WD(n,k)$$
$$WD(n,k) = e^{-j\pi k/N}\sum_{m=-N/2}^{N/2-1} x_e(n+m)x_e^*(n-m-1)e^{-j2\pi mk/N}. \qquad (11.22)$$

In MATLAB notation, relation (11.1.4) can be implemented, as follows. The signal values are

$$\mathbf{x}_n^+ = [x_e(n-N/2),\ x_e(n-N/2+1),\dots,x_e(n+N/2-1)],$$
$$\mathbf{x}_n^- = [x_e^*(n+N/2-1),\ x_e^*(n+N/2-2),\dots,x_e^*(n-N/2)].$$

The vector of Wigner distribution values, for a given n and k, is

$$WD(n,k)=e^{-j\pi k/N}\left\{\mathbf{x}_n^+ * \left(\mathbf{x}_n^- .* e^{-j\pi k\mathbf{m}/N}\right)^T\right\},$$

where $e^{-j\pi k\mathbf{m}/N}$ is the vector with elements $e^{-j\pi km/N}$, for $-N/2 \le m \le N/2-1$, $*$ is the matrix multiplication and $.*$ denotes the vector multiplication term by term.

Thus, in the case of an even number of samples, the discrete Wigner distribution of a signal $x_e(n)$, calculated according to (11.1.4), corresponds to the original signal $x(t)$ related to $x_e(n)$ as

$$x_e(n) \leftrightarrow x(n\Delta t+\Delta t/2)\sqrt{2\Delta t}.$$

To check this statement, consider the time marginal property of this distribution. It is

$$\frac{1}{N}\sum_{k=-N/2}^{N/2-1} WD(n,k) = \sum_{m=-N/2}^{N/2-1}\left(x_e(n+m)x_e^*(n-m-1)\frac{1}{N}\sum_{k=-N/2}^{N/2-1} e^{-j(2m+1)\pi k/N}\right)$$
$$= \sum_{m=-N/2}^{N/2-1}\left(x_e(n+m)x_e^*(n-m-1)\frac{1}{N}e^{j(2m+1)\pi/2}\frac{1-e^{-j(2m+1)\pi}}{1-e^{-j(2m+1)\pi/N}}\right)$$
$$= \sum_{m=-N/2}^{N/2-1}(x_e(n+m)x_e^*(n-m-1)\delta(2m+1)) = \left|x_e(n-\frac{1}{2})\right|^2 = |x(n\Delta t)|^2(2\Delta t),$$

for $|2m+1| < N$.

Since for any signal $y(n)$ and its DFT holds

$$\mathrm{DFT}_{N/2}\{y(n)+y(n+N/2)\} = Y(2k),$$

where

$$Y(k) = \mathrm{DFT}_N\{y(n)\},$$

the pseudo Wigner distribution (11.1.4), without frequency ovesampling, in the case of an even N, can be calculated as

$$WD\left(n\Delta t, \frac{2\pi k}{N\Delta t}\right) \to WD(n,k)$$
$$WD(n,k) = e^{-j\pi k/(N/2)}\sum_{m=-N/4}^{N/4-1}(R(n,m)+R(n,m+N/2))\,e^{-j2\pi mk/(N/2)}$$

where

$$R(n,m) = x_e(n+m)x_e^*(n-m-1).$$

Periodicity in m, for a given n, with period N is assumed in $R(n,m)$, that is, $R(n,m+N) = R(n,m) = R(n,m-N)$. It is needed to calculate $R(n,m+N/2)$ for $-N/4 \le m \le N/4-1$ using $R(n,m)$ for $-N/2 \le m \le N/2-1$ only.

In the case of real-valued signals, in order to avoid the need for oversampling, as well as to eliminate cross-terms (that will be discussed later) between positive and negative frequency components, their analytic part is used in calculations.

11.2 FROM THE STFT TO THE WIGNER DISTRIBUTION VIA S-METHOD

The pseudo Wigner distribution can be calculated as

$$PWD(t,\Omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} STFT(t,\Omega+\theta)STFT^*(t,\Omega-\theta)d\theta. \tag{11.23}$$

Where STFT is defined as

$$STFT(t,\Omega) = \int_{-\infty}^{\infty} x(t+\tau)w(\tau)e^{-j\Omega\tau}d\tau. \tag{11.24}$$

This can be proven by substituting (11.24) into (11.23).

Relation (11.23) has led to the definition of a time-frequency distribution

$$SM(t,\Omega) = \frac{1}{\pi}\int_{-L_P}^{L_P} P(\theta)STFT(t,\Omega+\theta)STFT^*(t,\Omega-\theta)d\theta, \tag{11.25}$$

where $P(\theta)$ is a finite frequency domain window (we also assume rectangular form), $P(\theta) = 0$ for $|\theta| > L_P$. Distribution obtained in this way is referred to as the S-method. Two special cases are: the spectrogram $P(\theta) = \pi\delta(\theta)$ and the pseudo Wigner distribution $P(\theta) = 1$.

The S-method can produce a representation of a multi-component signal such that the distribution of each component is its Wigner distribution, avoiding cross-terms, if the STFTs of the components do not overlap in time-frequency plane.

Consider a signal

$$x(t) = \sum_{m=1}^{M} x_m(t)$$

where $x_m(t)$ are monocomponent signals. Assume that the STFT of each component lies inside the region $D_m(t,\Omega)$, $m = 1,2,\ldots,M$ and assume that regions $D_m(t,\Omega)$ do not overlap. Denote the length of the m-th region along Ω, for a given t, by $2B_m(t)$, and its central frequency by $\Omega_{0m}(t)$. Under this assumptions the S-method of $x(t)$ produces the sum of the pseudo Wigner distributions of each signal component

$$SM_x(t,\Omega) = \sum_{m=1}^{M} PWD_{x_m}(t,\Omega), \tag{11.26}$$

if the width of the rectangular window $P(\theta)$, for a point (t,Ω), is defined by

$$L_P(t,\Omega) = \begin{cases} B_m(t) - |\Omega - \Omega_{0m}(t)| \text{ for } (t,\Omega) \in D_m(t,\Omega) \\ 0 \text{ elsewhere.} \end{cases}$$

To prove this consider a point (t,Ω) inside a region $D_m(t,\Omega)$. The integration interval in (11.25), for the m-th signal component is symmetrical with respect to $\theta = 0$. It is defined by the smallest absolute value of θ for which $\Omega+\theta$ or $\Omega-\theta$ falls outside $D_m(t,\Omega)$, that is,

$$|\Omega \pm \theta - \Omega_{0m}(t)| \geq B_m(t).$$

For $\Omega > \Omega_{0m}(t)$ and positive θ, the integration limit is reached for $\theta = B_m(t) - (\Omega - \Omega_{0m}(t))$. For $\Omega < \Omega_{0m}(t)$ and positive θ, the limit is reached for $\theta = B_m(t) + (\Omega - \Omega_{0m}(t))$. Thus, having in mind the interval symmetry, an integration limit which produces the same value of integral (11.25) as the value of (11.23), over the region $D_m(t,\Omega)$, is given by $L_P(t,\Omega)$. Therefore, for $(t,\Omega) \in D_m(t,\Omega)$ we have $SM_x(t,\Omega) = PWD_{x_m}(t,\Omega)$. Since regions $D_m(t,\Omega)$ do not overlap we have

$$SM_x(t,\Omega) = \sum_{m=1}^{M} PWD_{x_m}(t,\Omega).$$

Note that any window $P(\theta)$ with constant width

$$L_P \geq \max_{(t,\Omega)}\{L_P(t,\Omega)\}$$

produces $SM_x(t,f) = \sum_{m=1}^{M} PWD_{x_m}(t,\Omega)$, if the regions $D_m(t,\Omega)$ for $m = 1,2,..,M$, are at least $2L_P$ apart along the frequency axis, $\left|\Omega_{0p}(t) - \Omega_{0q}(t)\right| > B_p(t) + B_q(t) + 2L_P$, for each p, q and t. This is the S-method with constant window width. The best choice of L_P is the value when $P(\theta)$ is wide enough to enable complete integration over the auto-terms, but narrower than the distance

between the auto-terms, in order to avoid the cross-terms. If two components overlap for some time instants t, then the cross-term will appear, but only between these two components and for that time instants.

A discrete form of the S-method (11.25) reads

$$SM_L(n,k) = \sum_{i=-L}^{L} S_N(n,k+i)S_N^*(n,k-i)$$

for $P(i) = 1$, $-L \leq i \leq L$ (a weighted form $P(i) = 1/(2L+1)$ could be used). A recursive relation for the S-method calculation is

$$SM_L(n,k) = SM_{L-1}(n,k) + 2\mathrm{Re}[S_N(n,k+L)S_N^*(n,k-L)], \tag{11.27}$$

The spectrogram is the initial distribution $SM_0(n,k) = |S_N(n,k)|^2$ and $2\mathrm{Re}[S_N(n,k+i)S_N^*(n,k-i)]$, $i = 1, 2, \ldots, L$ are the correction terms. Changing parameter L we can start from the spectrogram ($L = 0$) and gradually make the transition toward the pseudo Wigner distribution by increasing L.

For the S-method realization we have to implement the STFT first, based either on the FFT routines or recursive approaches suitable for hardware realizations. After we get the STFT we have to "correct" the obtained values, according to (11.27), by adding few "correction" terms to the spectrogram values. Note that S-method is one of the rare quadratic time-frequency distributions allowing easy hardware realization, based on the hardware realization of the STFT, presented in the first part, and its "correction" according to (11.27). There is no need for analytic signal since the cross-terms between negative and positive frequency components are removed in the same way as are the other cross-terms. If we take that $STFT(n,k) = 0$ outside the basic period, that is, when $k < -N/2$ or $k > N/2 - 1$, then there is no aliasing when the STFT is alias-free (in this way we can calculate the alias-free Wigner distribution by taking $L = N/2$ in (11.27)). The calculation in (11.27) can be performed for the whole matrix of the S-method and the STFT. This can significantly save time in some matrix based calculation tools.

There are two ways to implement summation in the S-method. The first one is with a constant L. Theoretically, in order to get the Wigner distribution for each individual component, the number of correcting terms L should be such that $2L$ is equal to the width of the widest auto-term. This will guarantee cross-terms free distribution for all components which are at least $2L$ frequency samples apart.

The second way to implement the S-method is with a time-frequency dependent $L = L_{(n,k)}$. The summation, for each point (n,k), is performed as long as the absolute values of $S_N(n,k+i)$ and $S_N^*(n,k-i)$ for that (n,k) are above an assumed reference level (established, for example, as a few percents of the STFT maximum value). Here, we start with the spectrogram, $L = 0$. Consider the correction term $S_N(n,k+i)S_N^*(n,k-i)$ with $i = 1$. If the STFT values are above the reference level then it is included in summation. The next term, with $i = 2$ is considered in the same way, and so on. The summation is stopped when a STFT in a correcting term is below the reference level. This procedure will guarantee cross-terms free distribution for components that do not overlap in the STFT.

Example 11.8. A signal consisting of three LFM components,

$$x(n) = \sum_{i=1}^{3} A_i \exp(ja_i\pi n/32 + jb_i\pi n^2/1024),$$

with

$$(a_1, a_2, a_3) = (-21, -1, 20)$$

and

$$(b_1, b_2, b_3) = (2, -0.75, -2.8),$$

is considered at the instant $n = 0$. The IFs of the signal components are $k_i = a_i$, while the normalized squared amplitudes of the components are indicated by dotted lines in Fig.11.8. An ideal time-frequency representation of this signal, at $n = 0$, would be

$$I(0,k) = A_1^2\delta(k - k_1) + A_2^2\delta(k - k_2) + A_3^2\delta(k - k_3).$$

The starting STFT, with the corresponding spectrogram, obtained using the cosine window of the width $N = 64$ is shown in Fig.11.8(a),(b). The first correction term is presented in Fig.11.8(c). The result of summing the spectrogram with the first correction term is the S-method with $L = 1$, Fig.11.8(d). The second correction term (Fig.11.8(e)) when added to $SM_1(0,k)$, produces the S-method with $L = 2$, Fig.11.8(f). The S-methods for $L = 3$, 5, and 8, ending with the Wigner distribution ($L = 31$) are presented in Fig.11.8(g)-(j). Just a few correction terms are sufficient in this case to achieve a high concentration. The cross-terms start appearing at $L = 8$ and increase as L increases toward the Wigner distribution. They make the Wigner distribution almost useless, since they cover a great part of the frequency range, including some signal components (Fig.11.8(j)).

The optimal number of correction terms L is the one that produces the best S-method concentration (sparsity), using the $\ell_{1/2}$-norm of the spectrogram and the S-method (corresponding to the ℓ_1-norm of the STFT). In this case the best concentrated S-method is detected for $L = 5$. The spectrogram is the initial distribution $SM_0(n,k) = |S_N(n,k)|^2$ and $2\mathrm{Re}[S_N(n,k+i)S_N^*(n,k-i)]$, $i = 1, 2, \ldots, L$ are the correction terms. Considering the parameter L as a frame index, we can make a video of the transition from the spectrogram to the Wigner distribution.

Example 11.9. The adaptive S-method realization will be illustrated on a five-component signal $x(t)$ defined for $0 \leq t < 1$ and sampled with $\Delta t = 1/256$. The Hamming window of the width $T_w = 1/2$ (128 samples) is used for STFT calculation. The spectrogram is presented in Fig.11.9(a), while the S-method with the constant $L_d = 3$ is shown in Fig.11.9(b). The concentration improvement with respect to the case $L_d = 0$, Fig.11.9(a), is evident. Further increasing of L_d would improve the concentration, but the cross-terms would also appear. Small changes are noticeable between the components with constant instantaneous frequency and between quadratic and constant instantaneous frequency component. An improved concentration, without cross-terms, can be achieved using the variable window width L_d. The regions $D_i(n,k)$, determining the summation limit $L_d(n,k)$ for each point (n,k), are obtained by imposing the reference level corresponding to 0.14% of its maximum value at that time instant n. They are defined as:

$$D_i(n,k) = \begin{cases} 1 \text{ when } |STFT_{x_i}(n,k)|^2 \geq R_n \\ 0 \text{ elsewhere} \end{cases}$$

and presented in Fig.11.9(c). White regions mean that the value of spectrogram is below 0.14% of its maximum value at that time instant n, meaning that the concentration improvement is not performed at these points. The signal dependent S-method is given in Fig.11.9(d). The method sensitivity, with respect to the reference level is low.

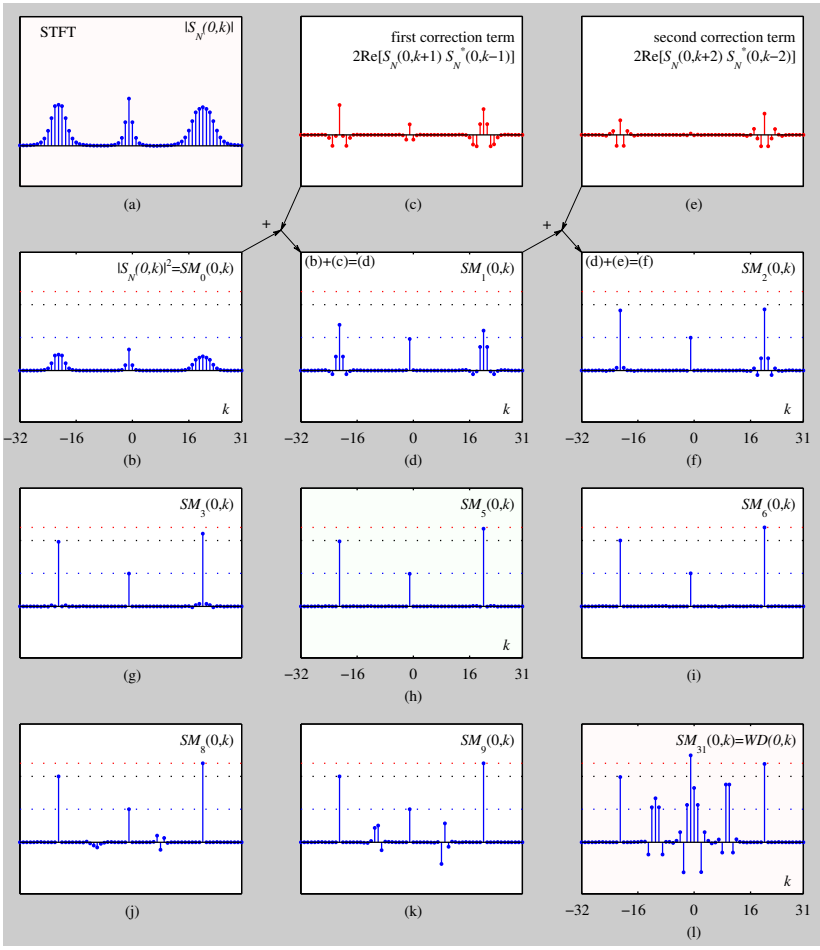


Figure 11.8 Analysis of a signal consisting of three LFM components (at the instant $n = 0$). (a) The STFT with a cosine window of the width $N = 64$. (b) The spectrogram. (c) The first correction term. (d) The S-method (SM) with one correction term. (e) The second correction term. (f) The S-method with two correction terms. (g) The S-method with three correction terms. (h) The S-method with five correction terms. (i) The S-method with six correction terms. (j) The S-method with eight correction terms.(k) The S-method with nine correction terms. (l) The Wigner distribution (the S-method with $L = 31$ correction term).

11.3 GENERAL QUADRATIC TIME-FREQUENCY DISTRIBUTIONS

In order to provide additional insight into the field of joint time-frequency analysis, as well as to improve concentration of time-frequency representation, energy distributions of signals were introduced. We have already mentioned the spectrogram which belongs to this class of representations and is a straightforward extension of the STFT. Here, we will discuss other distributions and their generalizations.

The basics condition for the definition of time-frequency energy distributions is that a two-dimensional function of time and frequency $P(t,\Omega)$ represents the energy density of a signal in the

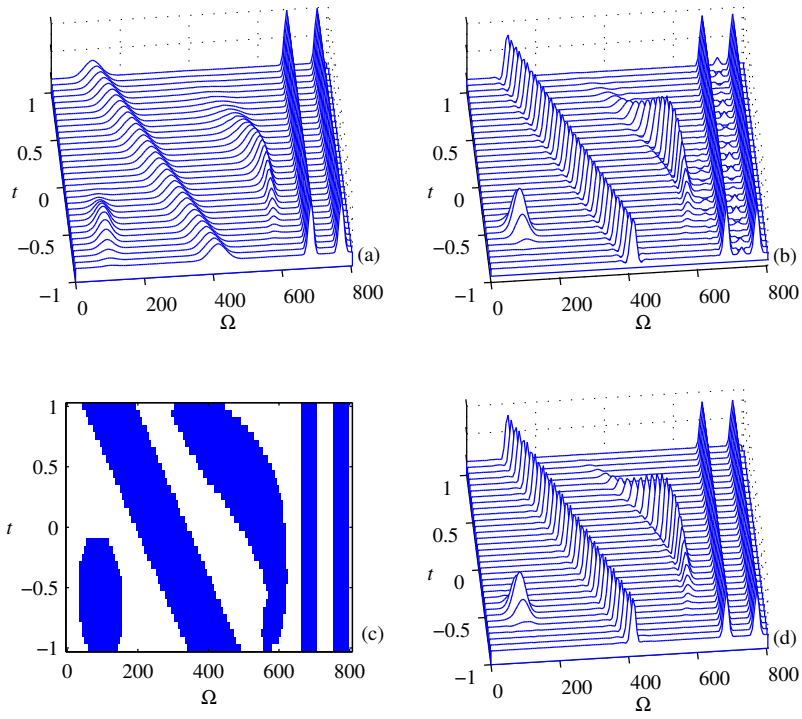


Figure 11.9 Time-frequency analysis of a multi-component signal: a) Spectrogram, b) The S-method with a constant window, with $L_P = 3$, c) Regions of support for the S-method with a variable window width calculation, corresponding to $Q^2 = 725$, d) The S-method with the variable window width calculated using regions in c).

time-frequency plane. Thus, the signal energy associated with the small time and frequency intervals Δt and $\Delta\Omega$, respectively, would be

$$\text{Signal energy within } [\Omega + \Delta\Omega, t + \Delta t] = P(t,\Omega)\Delta\Omega\Delta t.$$

However, point by point definition of time-frequency energy densities in the time-frequency plane is not possible, since the uncertainty principle prevents us from defining concept of energy at a specific instant and frequency. This is the reason why some more general conditions are being considered to derive time-frequency distributions of a signal. Namely, one requires that the integral of $P(t,\Omega)$ over Ω, for a particular instant of time should be equal to the instantaneous power of the signal $|x(t)|^2$, while the integral over time for a particular frequency should be equal to the spectral energy density $|X(\Omega)|^2$. These conditions are known as marginal conditions or marginal properties of time-frequency distributions.

Therefore, it is desirable that an energetic time-frequency distribution of a signal $x(t)$ satisfies:

– Energy property

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(t,\Omega)\,d\Omega\,dt = E_x, \tag{11.28}$$

– Time marginal properties

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} P(t,\Omega)\,d\Omega = |x(t)|^2, \text{ and} \tag{11.29}$$

– Frequency marginal property

$$\int_{-\infty}^{\infty} P(t,\Omega)\,dt = |X(\Omega)|^2, \tag{11.30}$$

where E_x denotes the energy of $x(t)$. It is obvious that if either one of marginal properties (11.29), (11.30) is fulfilled, so is the energy property. Note that relations (11.28), (11.29) and (11.30), do not reveal any information about the local distribution of energy at a point (t,Ω). The marginal properties are illustrated in Fig. 11.10.

Next we will introduce some distributions satisfying these properties.

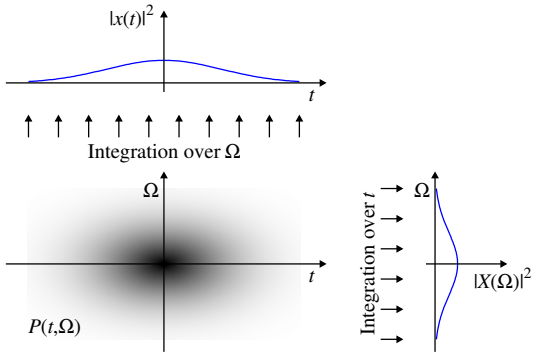


Figure 11.10 Illustration of the marginal properties

Time and frequency marginal properties (11.29) and (11.30) may be considered as the projections of the distribution $P(t,\Omega)$ along the time and frequency axes, that is, as the Radon transform of $P(t,\Omega)$ along these two directions. It is known that the Fourier transform of the projection of a two-dimensional function on a given line is equal to the value of the two-dimensional Fourier transform of $P(t,\Omega)$, denoted by $AF(\theta,\tau)$, along the same direction (inverse Radon transform property). Therefore, if $P(t,\Omega)$ satisfies marginal properties then any other function having two-dimensional Fourier transform equals to $AF(\theta,\tau)$ along the axes lines $\theta = 0$ and $\tau = 0$, and arbitrary values elsewhere, will satisfy marginal properties, Fig. 11.11.

Assuming that the Wigner distribution is a basic distribution which satisfies the marginal properties (any other distribution satisfying marginal properties can be used as the basic one), then any other distribution with two-dimensional Fourier transform

$$AF_g(\theta,\tau) = c(\theta,\tau)\mathrm{FT}^{2D}_{t,\Omega}\{WD(t,\Omega)\} = c(\theta,\tau)AF(\theta,\tau) \tag{11.31}$$

where $c(0,\tau) = 1$ and $c(\theta,0) = 1$, satisfies marginal properties as well.

The inverse two-dimensional Fourier transform of $AF_g(\theta,\tau)$ produces the Cohen class of distributions, introduced from quantum mechanics into the time-frequency analysis by Claasen and

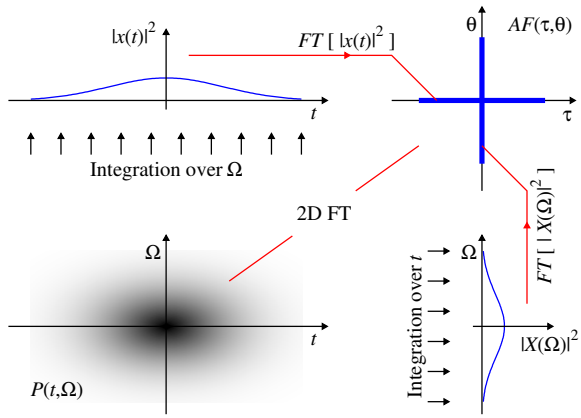


Figure 11.11 Marginal properties and their relation to the ambiguity function.

Mecklenbäuker, in the form

$$CD(t,\Omega)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}c(\theta,\tau)x(u+\tau/2)x^*(u-\tau/2)e^{j\theta t-j\Omega\tau-j\theta u}dud\tau d\theta \tag{11.32}$$

where $c(\theta,\tau)$ is called the kernel in the ambiguity domain.

Alternatively, the frequency domain definition of the Cohen class of distributions is

$$CD(t,\Omega)=\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}X(u-\theta/2)X^*(u+\theta/2)c(\theta,\tau)e^{j\theta t-j\tau\Omega+j\tau u}dud\tau d\theta. \tag{11.33}$$

Various distributions can be obtained by altering the kernel function $c(\theta,\tau)$. For example, $c(\theta,\tau)=1$ produces the Wigner distribution, while for $c(\theta,\tau)=e^{j\theta\tau/2}$ the Rihaczek distribution follows.

The Cohen class of distributions, defined in the ambiguity domain:

$$CD(t,\Omega)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}c(\theta,\tau)AF(\theta,\tau)e^{j\theta t-j\Omega\tau}d\tau d\theta \tag{11.34}$$

can be written in other domains, as well. The time-lag domain form is obtained from (11.32), after integration on θ, as:

$$CD(t,\Omega)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}c_T(t-u,\tau)x(u+\tau/2)x^*(u-\tau/2)e^{-j\Omega\tau}d\tau du. \tag{11.35}$$

The frequency-Doppler frequency domain form follows from (11.33), after integration on τ, as:

$$CD(t,\Omega)=\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}C_\Omega(\theta,\Omega-u)X(u+\theta/2)X^*(u-\theta/2)e^{j\theta t}d\theta du. \tag{11.36}$$

Finally, the time-frequency domain form is obtained as a two-dimensional convolution of the two-dimensional Fourier transforms, from (11.34), as:

$$CD(t,\Omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Pi(t-u,\Omega-\xi)WD(u,\xi)\,du\,d\xi. \tag{11.37}$$

Kernel functions in the respective time-lag, Doppler frequency-frequency and time-frequency domains are related to the ambiguity domain kernel $c(\theta,\tau)$ as:

$$c_T(t,\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} c(\theta,\tau)e^{j\theta t}d\theta \tag{11.38}$$

$$C_\Omega(\theta,\Omega) = \int_{-\infty}^{\infty} c(\theta,\tau)e^{-j\Omega\tau}d\tau \tag{11.39}$$

$$\Pi(t,\Omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} c(\theta,\tau)e^{j\theta t-j\Omega\tau}\,d\tau\,d\theta. \tag{11.40}$$

According to (11.37) all distributions from the Cohen class may be considered as 2D filtered versions of the Wigner distribution. Although any distribution could be taken as a basis for the Cohen class derivation, the form with the Wigner distribution is used because it is the best concentrated distribution from the Cohen class with the signal independent kernels.

11.3.1 Reduced Interference Distributions

The analysis performed on ambiguity function and Cohen class of time-frequency distributions leads to the conclusion that the cross-terms may be suppressed or eliminated, if a kernel $c(\theta,\tau)$ is a function of a two-dimensional low-pass type. In order to preserve the marginal properties $c(\theta,\tau)$ values along the axis should be $c(\theta,0)=1$ and $c(0,\tau)=1$.

Choi and Williams exploited one of the possibilities defining the distribution with the kernel of the form

$$c(\theta,\tau) = e^{-\theta^2\tau^2/\sigma^2}.$$

The parameter σ controls the slope of the kernel function which affects the influence of cross-terms. Small σ causes the elimination of cross-terms but it should not be too small because, for the finite width of the auto-terms around θ and τ coordinates, the kernel will cause their distortion, as well. Thus, there should be a trade-off in the selection of σ.

Here we will mention some other interesting kernel functions, producing corresponding distributions, Fig. 11.12:
Born-Jordan distribution

$$c(\theta,\tau) = \frac{\sin(\frac{\theta\tau}{2})}{\frac{\theta\tau}{2}},$$

Zhao-Atlas-Marks distribution

$$c(\theta,\tau) = w(\tau)\,|\tau|\,\frac{\sin(\frac{\theta\tau}{2})}{\frac{\theta\tau}{2}},$$

Sinc distribution

$$c(\theta,\tau) = \text{rect}(\frac{\theta\tau}{\alpha}) = \begin{cases} 1 & \text{for } |\theta\tau/\alpha| < 1/2 \\ 0 & \text{otherwise} \end{cases}$$

Butterworth distribution

$$c(\theta,\tau)=\frac{1}{1+(\frac{\theta\tau}{\theta_c\tau_c})^{2N}},$$

where $w(\tau)$ is a function corresponding to a lag window and α, N, θ_c and τ_c are constants in the above kernel definitions.

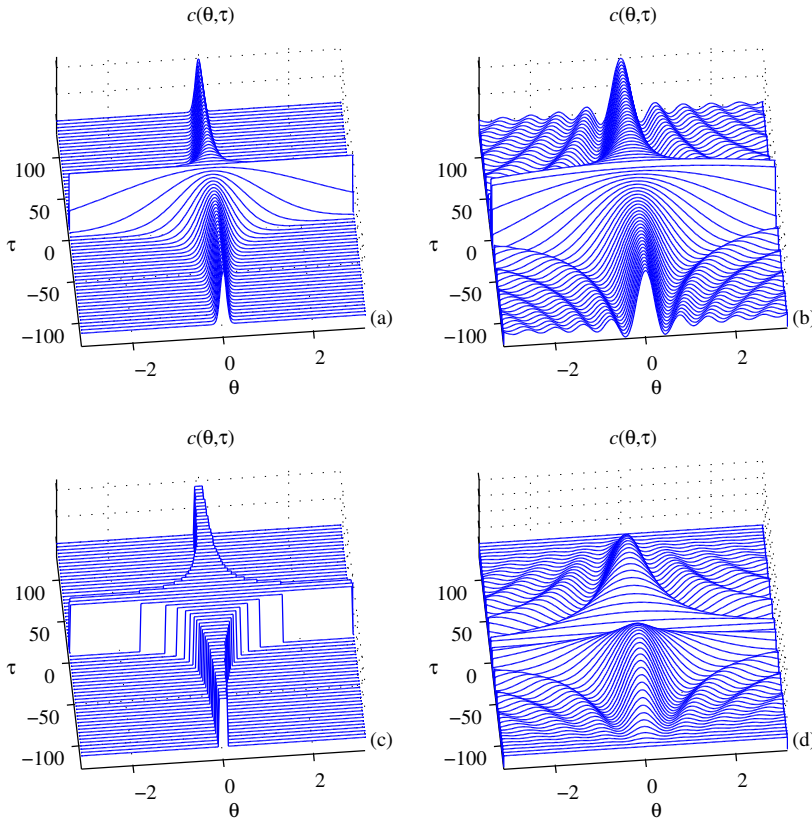


Figure 11.12 Kernel functions for: Choi-Williams distribution, Born-Jordan distribution, Sinc distribution and Zhao-Atlas-Marks distribution.

The spectrogram belongs to this class of distributions. Its kernel in (θ,τ) domain is the ambiguity function of the window

$$c(\theta,\tau)=\int_{-\infty}^{\infty} w\left(t-\frac{\tau}{2}\right)w\left(t+\frac{\tau}{2}\right)e^{-j\theta t}dt=AF_w(\theta,\tau).$$

Since the Cohen class is linear with respect to the kernel, it is easy to conclude that a distribution from the Cohen class is positive if its kernel can be written as

$$c(\theta,\tau)=\sum_{i=1}^{M}a_iAF_{w_i}(\theta,\tau),$$

where $a_i \geq 0$, $i = 1,2,\ldots,M$.

There are several ways for calculation of the reduced interference distributions from the Cohen class. The first method is based on the ambiguity function (11.34):

1. Calculation of the ambiguity function,
2. Multiplication with the kernel,
3. Calculation of the inverse two-dimensional Fourier transform of this product.

The reduced interference distribution can be calculated by using (11.35) or (11.37) with appropriate kernel transformations defined by (11.38) and (11.40). All these methods assume signal oversampling in order to avoid aliasing effects. Figure 11.13 shows the ambiguity function along with kernel (Choi-Williams). Figure 11.14(a) presents Choi-Williams distribution calculated according to the presented procedure. In order to reduce high side lobes of the rectangular window, the Choi-Williams distribution is also calculated with the Hann(ing) window in the kernel definition $c(\theta,\tau)w(\tau)$ and shown in Fig. 11.14(b). The pseudo Wigner distribution with Hann(ing) window is given in Fig. 11.6.

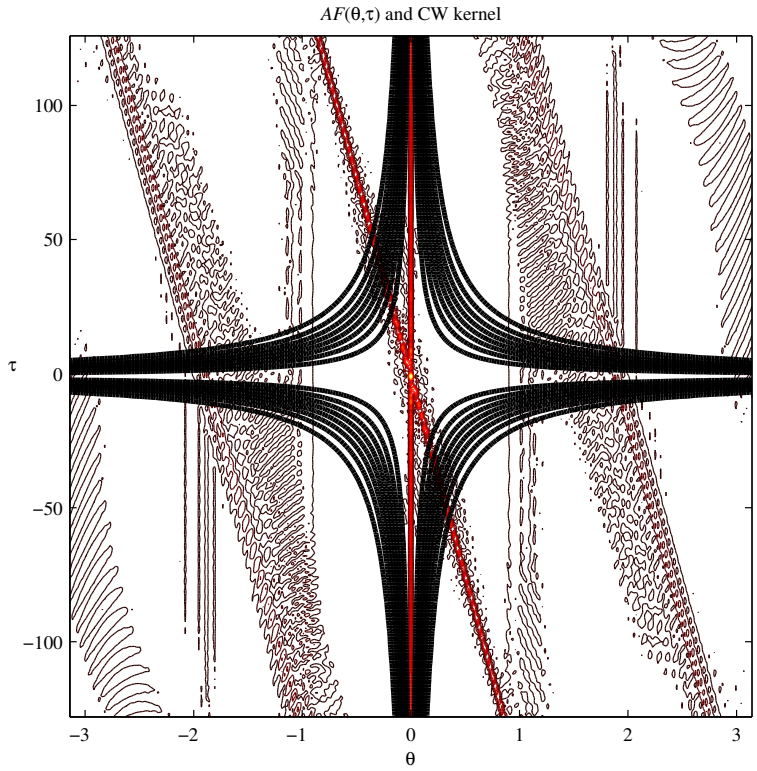


Figure 11.13 Ambiguity function for signal from Fig.10.4 with the Choi-Williams kernel

For the discrete-time signals. there are several ways to calculate a reduced interference distributions from the Cohen class, based on (11.34), (11.35), (11.36), or (11.37).

The kernel functions are usually defined in the Doppler-lag domain (θ,τ). Thus, here we should use (11.34) with the ambiguity function of a discrete-time signal

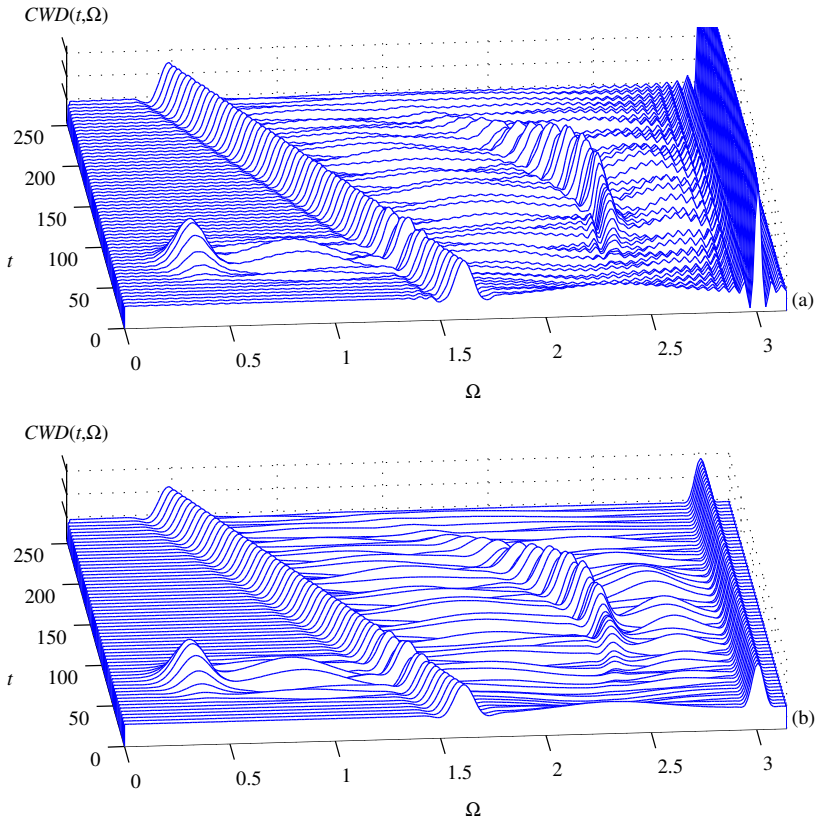


Figure 11.14 Choi-Williams distribution: (a) direct calculation, (b) calculation with the kernel multiplied by a Hann(ing) lag window.

$$AF(\theta, m\Delta t) = \sum_{p=-\infty}^{\infty} x\left(p\Delta t + m\frac{\Delta t}{2}\right) x^*\left(p\Delta t - m\frac{\Delta t}{2}\right) e^{-jp\theta\Delta t}\Delta t.$$

The signal should be sampled as in the Wigner distribution case. For a given lag instant m, the ambiguity function can be calculated using the standard DFT routines. Another way to calculate the ambiguity function is to take the inverse two-dimensional transform of the Wigner distribution. Note that the corresponding transformation pairs are $time \leftrightarrow Doppler$ and $lag \leftrightarrow frequency$, that is, $t \leftrightarrow \theta$ and $\tau \leftrightarrow \Omega$. The relation between discretization values in the Fourier transform pairs (considered interval, sampling interval in time Δt, number of samples N, sampling interval in frequency $\Delta\Omega = 2\pi/(N\Delta t)$) is discussed in Chapter 1.

The generalized ambiguity function is obtained as

$$\begin{aligned} AF_g(l\Delta\theta, m\Delta t) &= c(l\Delta\theta, m\Delta t) AF(l\Delta\theta, m\Delta t) \\ &= c(l\Delta\theta, m\Delta t) \sum_{p=-\infty}^{\infty} x\left(p\Delta t + m\frac{\Delta t}{2}\right) x^*\left(p\Delta t - m\frac{\Delta t}{2}\right) e^{-jl\Delta\theta p\Delta t}\Delta t, \end{aligned} \tag{11.41}$$

while a distribution, with kernel $c(\theta,\tau)$ is the two-dimensional inverse Fourier transform in the form

$$CD(n\Delta t,k\Delta\Omega)=\frac{1}{2\pi}\sum_{l=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}AF_g(l\Delta\theta,m\Delta t)e^{-jkm\Delta t\Delta\Omega}e^{jnl\Delta\theta\Delta t}\Delta t\Delta\theta.$$

In this notation we can calculate $CD(n,k)=\text{IDFT}^{2D}_{l,m}\left\{AF_g(l,m)\right\}$, where the values of $AF_g(l,m)$ are calculated according to (11.41).

In the time-lag domain, the discrete-time form reads

$$\begin{aligned}CD(n\Delta t,k\Delta\Omega)&=\sum_{p=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}c_T(n\Delta t-p\Delta t,m\Delta t)\\&\times x\left(p\Delta t+m\frac{\Delta t}{2}\right)x^*\left(p\Delta t-m\frac{\Delta t}{2}\right)e^{-jkm\Delta t\Delta\Omega}(\Delta t)^2\end{aligned}\tag{11.42}$$

with

$$c_T(n\Delta t-p\Delta t,m\Delta t)=\frac{1}{2\pi}\sum_{l=-\infty}^{\infty}c(l\Delta\theta,m\Delta t)e^{jnl\Delta\theta\Delta t}e^{-jlp\Delta\theta\Delta t}\Delta\theta.$$

For the discrete-time signals, it is common to write and use the Cohen class of distributions in the form

$$CD(n,\omega)=\sum_{p=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}c_T(n-p,m)x(p+m)x^*(p-m)e^{-j2m\omega},\tag{11.43}$$

where

$$\begin{aligned}x(p+m)x^*(p-m)&=x\left((p+m)\frac{\Delta t}{2}\right)x^*\left((p-m)\frac{\Delta t}{2}\right)\Delta t\\c_T(n-p,m)&=c_T\left((n-p)\frac{\Delta t}{2},m\Delta t\right)\frac{\Delta t}{2}\\CD(n,\omega)&\to CD\left(n\frac{\Delta t}{2},\Omega\Delta t\right).\end{aligned}$$

Here we should mention that the presented kernel functions are of infinite duration along the coordinate axis in (θ,τ) thus, they should be limited in calculations. Their transforms exist in a generalized sense only.

11.3.2 Kernel Decomposition Method

Distributions from the Cohen class can be calculated using decomposition of the kernel function in the time-lag domain. Starting from

$$CD(t,\Omega)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}c_T(t-u,\tau)x(u+\tau/2)x^*(u-\tau/2)e^{-j\Omega\tau}d\tau du$$

with substitutions $u+\tau/2=t+v_1$ and $u-\tau/2=t+v_2$ we get $t-u=-(v_1+v_2)/2$ and $\tau=v_1-v_2$, resulting in

$$CD(t,\Omega)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}c_T(-\frac{v_1+v_2}{2},v_1-v_2)x(t+v_1)x^*(t+v_2)e^{-j\Omega(v_1-v_2)}dv_1dv_2$$

The discrete-time version of the Cohen class of distribution can be written, as

$$CD(n,\omega)=\sum_{n_1}\sum_{n_2}c_T\left(-\frac{n_1+n_2}{2},n_1-n_2\right)[x(n+n_1)e^{-j\omega n_1}][x(n+n_2)e^{-j\omega n_2}]^*.$$

Assuming that $\mathbf{C}$ is a square matrix of finite dimension, with elements:

$$C(n_1,n_2)=c_T\left(-\frac{n_1+n_2}{2},n_1-n_2\right)$$

we can write

$$CD(n,\omega)=\mathbf{x}_n\mathbf{C}\mathbf{x}_n^H$$

where $\mathbf{x}_n$ is a vector with elements $x(n+n_1)e^{-j\omega n_1}$. We can now perform the eigenvalue decomposition, finding solutions of $\det(\mathbf{C}-\lambda\mathbf{I})=0$ and determining eigenvectors matrix $\mathbf{Q}$ that satisfies $\mathbf{Q}\mathbf{Q}^H=\mathbf{I}$ and

$$\mathbf{C}=\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^H,$$

where $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues. It results in

$$CD(n,\omega)=(\mathbf{x}_n\mathbf{Q})\mathbf{\Lambda}(\mathbf{x}_n\mathbf{Q})^H$$

Then, it is easy to conclude that the Cohen class of distribution can be written as a sum of spectrograms:

$$CD(n,\omega)=\sum_i\lambda_i|STFT_{\mathbf{q}_i}(n,\omega)|^2$$

where λ_i represents eigenvalues, while $\mathbf{q}_i$ are corresponding eigenvectors of $\mathbf{C}$, that is, columns of $\mathbf{Q}$, used as windows in the STFT calculations.

Example 11.10. A four-component real-valued signal with $M=384$ samples is considered. Its STFT is calculated with a Hann(ing) window of the width $N=128$ with a step of 4 samples. The spectrogram ($L=0$) is shown in Fig.11.15(a). The alias-free Wigner distribution ($L=N/2$) is presented in Fig. 11.15(b). The Choi-Williams distribution of analytic signal is shown in Fig. 11.15(c). Its cross-terms are smoothed by the kernel, that also spreads the auto-term of the LFM signal and chirps. The S-method with $L=10$ is shown in Fig. 11.15(d). For graphical presentation, the distributions are interpolated by a factor of 2. In all cases the pure sinusoidal signal is well concentrated. In the Wigner distribution and the SM the same concentration is achieved for the LFM signal.

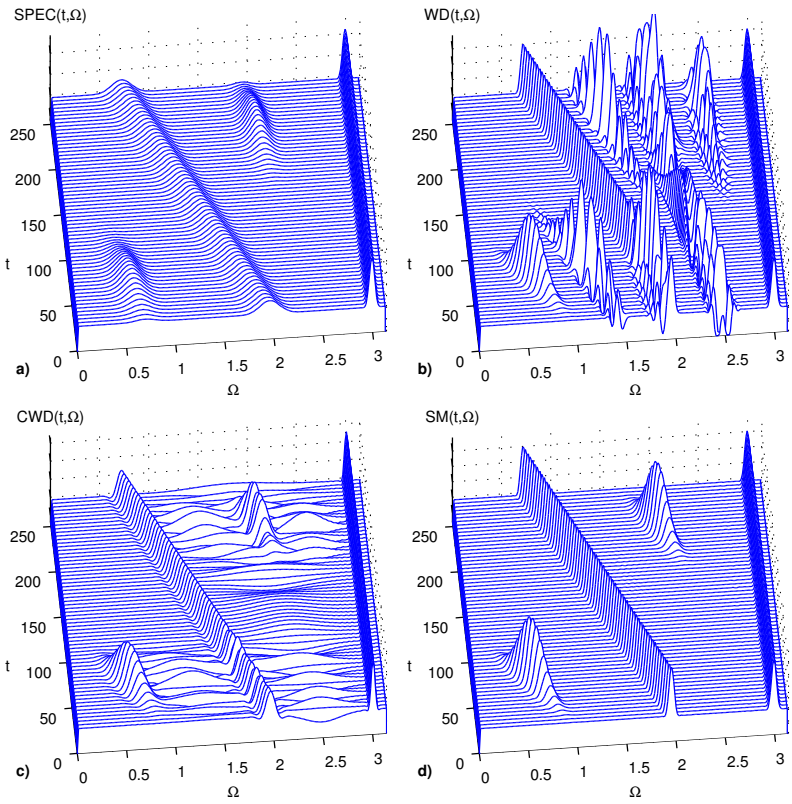


Figure 11.15 Time-frequency representation of a four component signal: (a) the spectrogram, (b) the Wigner distribution, (c) the Choi-Williams distribution, and (d) the S-method.

Chapter 12

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, 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.

12.1 CONTINUOUS MORLET WAVELET TRANSFORM

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. The window width is decreased as frequency increases.

The basis functions in the STFT are

$$STFT_{II}(t,\Omega_0) = \int_{-\infty}^{\infty} x(\tau)w^*(\tau-t)e^{-j\Omega_0\tau}d\tau$$

$$= \left\langle x(\tau), w(\tau-t)e^{j\Omega_0\tau} \right\rangle = \langle x(\tau), h(\tau-t)\rangle = \int_{-\infty}^{\infty} x(\tau)h^*(\tau-t)d\tau$$

where $h(\tau-t) = w(\tau-t)e^{j\Omega_0\tau}$ is a a band-pass signal, obtained when a real-valued window $w(\tau-t)$ is modulated by $e^{j\Omega_0\tau}$. Notice that $h(\tau-t) = w(\tau-t)e^{j\Omega_0(\tau-t)}$ is also used. This form follows from

$$STFT(t,\Omega_0) = \int_{-\infty}^{\infty} x(t+\tau)w^*(\tau)e^{-j\Omega_0\tau}d\tau = \int_{-\infty}^{\infty} x(\tau)w^*(\tau-t)e^{-j\Omega_0(\tau-t)}d\tau.$$

When the above idea about the wavelet transform is translated into the mathematical form and related to the STFT, one gets the definition of a continuous wavelet transform

$$WT(t,a) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(\tau) h^*(\frac{\tau - t}{a}) d\tau, \tag{12.1}$$

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

$$h(\tau - t) = w(\tau - t) e^{j\Omega_0(\tau - t)} \tag{12.2}$$

where $w(t)$ is a window function and Ω_0 is a constant frequency. For the Morlet wavelet we have a Gaussian function

$$w(\tau) = \sqrt{\frac{1}{2\pi}} e^{-\alpha\tau^2}$$

where the values of α and Ω_0 are chosen such that the ratio of $h(0)_{t=0} = 1/\sqrt{2\pi}$ and the first maximum (of the real part of the Morlet wavelet $w(\tau - t)\cos(\Omega_0\tau)$, which is also used in the analysis) at $\tau = 2\pi/\Omega_0$ is equal to $1/2 = \exp(-\alpha 4\pi^2/\Omega_0^2)$, that is, $\Omega_0 = 2\pi\sqrt{\alpha/\ln 2}$. Substitution of (12.2) into (12.1) leads to a continuous wavelet transform form suitable for a direct comparison with the STFT

$$WT(t,a) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(\tau) w^*(\frac{\tau - t}{a}) e^{-j\Omega_0(\tau - t)/a} d\tau = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t + \tau) w^*(\frac{\tau}{a}) e^{-j\Omega_0\tau/a} d\tau. \tag{12.3}$$

From the definition of $w(\tau/a)$ it is obvious that small Ω (that is, large a) corresponds to a wide wavelet, that is, a wide window, and vice versa. The basic idea of the wavelet transform and its comparison with the STFT is illustrated in Fig. 12.1.

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 $h^*(-t/a)\sqrt{|a|}$, that is,

$$WT(t,a) = x(t) *_t h^*(-t/a)\sqrt{|a|},$$

where $*_t$ denotes a convolution in time. Similarly the STFT, for a given Ω, may be considered as $STFT_{II}(t,\Omega) = x(t) *_t [w^*(-t)e^{j\Omega t}]$. If we consider these two bandpass 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

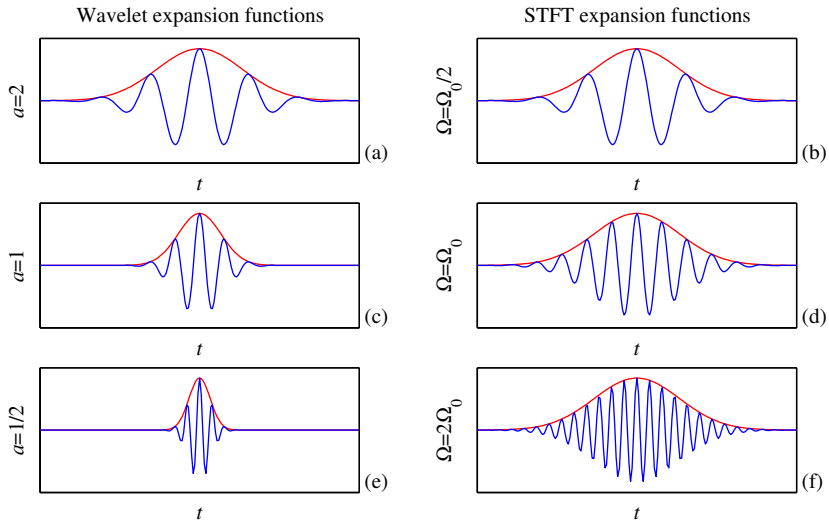


Figure 12.1 Expansion functions for the wavelet transform (left) and the short-time Fourier transform (right). Top row presents high scale (low frequency), middle row is for medium scale (medium frequency) and bottom row is for low scale (high frequency).

mother wavelet (wavelet in 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 Ω, 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 12.1. Find the wavelet transform of signal (10.3)

$$x(t) = \delta(t - t_1) + \delta(t - t_2) + e^{j\Omega_1 t} + e^{j\Omega_2 t}. \tag{12.4}$$

★ Its continuous wavelet transform is

$$WT(t,a) = \frac{1}{\sqrt{|a|}}\left[w((t_1 - t)/a)e^{-j\Omega_0(t_1-t)/a} + w((t_2 - t)/a)e^{-j\Omega_0(t_2-t)/a}\right]$$

$$+\sqrt{|a|}\left[e^{j\Omega_1 t}W[a(\Omega_0/a - \Omega_1)] + e^{j\Omega_2 t}W[a(\Omega_0/a - \Omega_2)]\right]. \tag{12.5}$$

where $w(t)$ is a real-valued function. The transform (12.5) has nonzero values in the region depicted in Fig. 12.2(a).

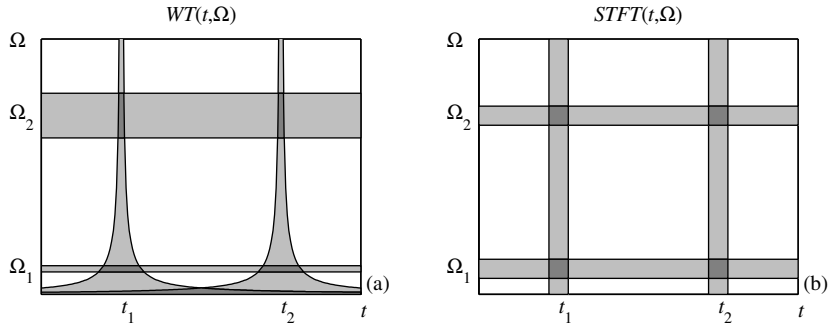


Figure 12.2 Illustration of the wavelet transform (a) of a sum of two delta pulses and two sinusoids compared to the STFT (b)

In analogy with spectrogram, the scalogram is defined as the squared magnitude of a wavelet transform:

$$SCAL(t,a) = \mid WT(t,a) \mid^2 . \tag{12.6}$$

The scalogram obviously loses the linearity property, and fits into the category of quadratic transforms.

12.1.1 S-Transform

The S-transform (the Stockwell transform) is conceptually a combination of the STFT analysis and wavelet analysis. It employs a common window, as in the STFT, with a frequency variable length as in the wavelet transform. The frequency-dependent window function produces a higher frequency resolution at lower frequencies, while at higher frequencies sharper time localization can be achieved, the same as in the continuous wavelet case.

For a signal $x(t)$ the S-transform is defined by

$$S_c(t,\Omega) = \frac{|\Omega|}{(2\pi)^{3/2}} \int_{-\infty}^{+\infty} x(\tau) e^{-\frac{(\tau-t)^2\Omega^2}{8\pi^2}} e^{-j\Omega\tau} d\tau, \tag{12.7}$$

with substitutions $\tau - t \to \tau$, the above equation can be rewritten as follows

$$S_c(t,\Omega) = \frac{|\Omega| e^{-j\Omega t}}{(2\pi)^{3/2}} \int_{-\infty}^{+\infty} x(t+\tau) e^{-\frac{\tau^2\Omega^2}{8\pi^2}} e^{-j\Omega\tau} d\tau. \tag{12.8}$$

For the window function of the form

$$w(\tau,\Omega) = \frac{|\Omega|}{(2\pi)^{3/2}} e^{-\frac{\tau^2\Omega^2}{8\pi^2}}, \tag{12.9}$$

the definition of the continuous S-transform can be rewritten as follows

$$S_c(t,\Omega) = e^{-j\Omega t} \int_{-\infty}^{+\infty} x(t+\tau) w(\tau,\Omega) e^{-j\Omega\tau} d\tau. \tag{12.10}$$

A discretization over τ of (12.10) results in the discrete form of S-transform

$$S_d(t,\Omega) = e^{-j\Omega t}\sum_n x(t+n\Delta t)w(n\Delta t,\Omega)e^{-j\Omega n\Delta t}\Delta t. \tag{12.11}$$

It may be considered as a STFT with frequency-varying window.

12.1.2 Spectral Meyer Wavelet Transform

The spectral domain STFT can be obtained from the corresponding time domain form

$$STFT(t,\Omega) = \int_{-\infty}^{\infty} x(t+\tau)w^*(\tau)e^{-j\Omega\tau}d\tau = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} X(\theta)e^{j(t+\tau)\theta}w^*(\tau)e^{-j\Omega\tau}d\theta d\tau$$
$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\theta)W^*(\theta-\Omega)e^{j\theta t}d\theta = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\theta)W^*_\Omega(\theta)e^{j\theta t}d\theta,$$

where $W(\theta)$ is the Fourier transform of the window function $w(\tau)$ and $W^*_\Omega(\theta)$ is its bandpass form centered at the frequency Ω (including the possibility that the form of $W^*_\Omega(\theta)$ is frequency-varying and that it changes with Ω as well, as in Section 10.6.1.1). The STFT can be considered as the projection (inner product) of the Fourier transform of the signal, $X(\theta)$, onto the kernel $W_\Omega(\theta)e^{-j\theta t}$.

The inversion relation can be derived in the same way as in Section 10.2. Assume that the STFT is calculated (available) for a set of discrete frequency values Ω_i. The Fourier transform of the signal is a projection of the STFT onto the kernel functions

$$\left\langle STFT(t,\Omega), W^*_\Omega(\theta)e^{j\theta t}\right\rangle_{\Omega,t} = \sum_{\Omega_i}\int_{-\infty}^{\infty} STFT(t,\Omega_i)W_{\Omega_i}(\theta)e^{-j\theta t}dt$$
$$= \sum_{\Omega_i}\left(X(\theta)W^*_{\Omega_i}(\theta)W_{\Omega_i}(\theta)\right) = X(\theta)\sum_{\Omega_i}\left(W^*_{\Omega_i}(\theta)W_{\Omega_i}(\theta)\right) = X(\theta)$$

if the condition

$$\sum_{\Omega_i}|W_{\Omega_i}(\theta)|^2 = 1, \tag{12.12}$$

holds. The inverse Fourier transform relation,

$$\int_{-\infty}^{\infty} STFT(t,\Omega)e^{-j\theta t}dt = X(\theta)W^*_\Omega(\theta),$$

is used. Notice that we have not used a factor of $1/(2\pi)$ within the scalar product definition and the summation over Ω_i, in order to simplify the notation. With this factor, the reconstruction condition would be $\sum_{\Omega_i}|W_{\Omega_i}(\theta)|^2/(2\pi) = 1$.

The spectral wavelet function is defined as a projection od the signal onto a set of the kernel functions $W(a_i\theta)e^{-j\theta t} = W_{\Omega_i}(\theta)e^{-j\theta t}$, where a_i is the scale which changes the position and the form of the basic bandpass function $W(\theta)$.

The Meyer wavelet transfer functions in the spectral domain, at a scale a_i, in the notation $W(a_i\theta)$, are defined as in (10.40), (10.42), and (10.43),

$$W_i(\theta) = W(a_i\theta) = \begin{cases} \sin\left(\frac{\pi}{2}\left(q(a_i\theta-1)\right)\right), & \text{for } 1 < a_i\theta \le M \\ \cos\left(\frac{\pi}{2}\left(q(\frac{a_i\theta}{M}-1)\right)\right), & \text{for } M < a_i\theta \le M^2 \end{cases} \tag{12.13}$$

and 0 elsewhere, for $2 \leq i \leq K-1$, where $q = 1/(M-1)$. The sine and cosine function arguments are such that they are either 0 or $\pi/2$ at the interval ending points. The scales a_i for each frequency interval are related through a geometric progression with a factor of $M > 1$, that is

$$a_i = a_{i-1}M = a_1 M^{i-1} = M^i/\theta_{\max},$$

where $\theta_{\max}$ is the maximum considered frequency.

The first function, corresponding to the highest and widest frequency band, is defined as

$$W_1(\theta) = W(a_1\theta) = \begin{cases} \sin\left(\frac{\pi}{2}\left(q(a_1\theta - 1)\right)\right), \text{ for } 1 < a_1\theta \leq M, \\ 0, \text{ elsewhere.} \end{cases} \tag{12.14}$$

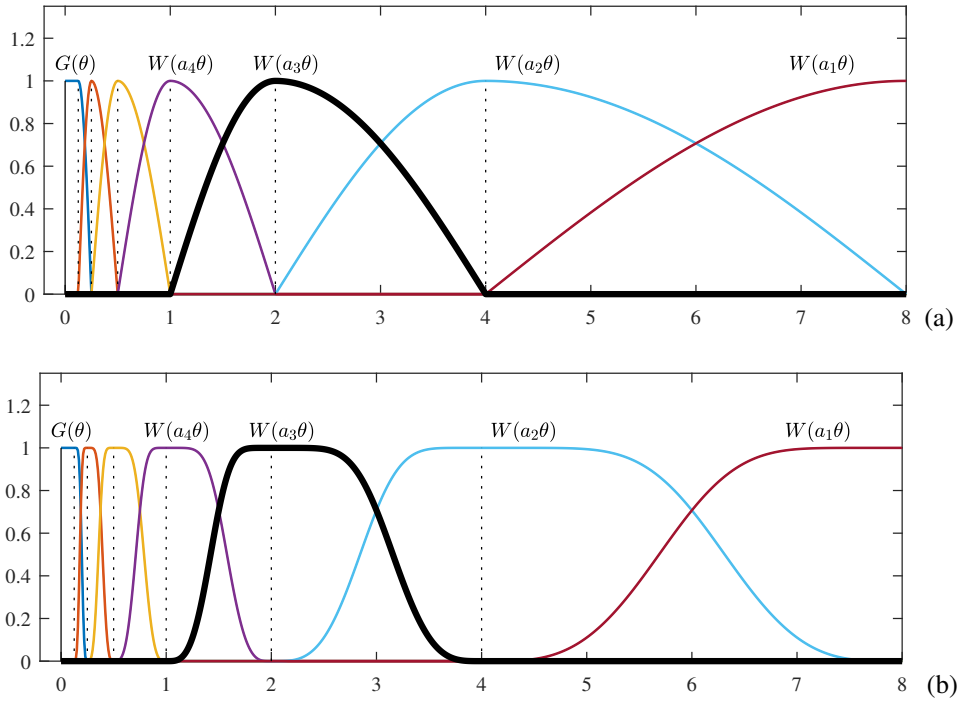


Figure 12.3 (a) Spectral domain windows (sine type) for the wavelet transform, $0 \leq \theta \leq 8$, that satisfy the reconstruction condition $\sum_i W^2(a_i\theta) = 1$, with , with $W(a_0\theta) = G(\theta)$, $M = 2$, $\theta_{\max} = 8$, and $K = 7$. (b) Spectral domain windows (Meyer spectral domain wavelet) for the wavelet transform, $0 \leq \theta \leq 8$, that satisfy the reconstruction condition $\sum_i W^2(a_i\theta) = 1$, with $W(a_0\theta) = G(\theta)$, $M = 2$, $\theta_{\max} = 8$, and $K = 7$.

Since $W(a_i\theta)$ are bandpass functions, to handle the lowpass spectral components (the interval for θ which includes $\theta = 0$), the lowpass type *scale function*, $G(\theta)$), is added in the form

$$G(\theta) = \begin{cases} 1, \text{ for } 0 \leq \theta \leq M/a_K = \theta_{\max}/M^{K-1} \\ \cos\left(\frac{\pi}{2}\left(q(\frac{a_K\theta}{M} - 1)\right)\right), \text{ for } M < a_K\theta \leq M^2 \\ 0, \text{ elsewhere.} \end{cases} \tag{12.15}$$

An example of the frequency domain windows (spectral transfer functions) of this wavelet is shown in Fig. 12.3(a).

The reconstruction condition in (12.12) can be written in the form of a sum of all normalized spectral transfer functions

$$\sum_{a_i}|W(a_i\theta)|^2 = |G(\theta)|^2 + \sum_{i=1}^{K-1}|W(a_i\theta)|^2 = 1. \tag{12.16}$$

The derivative discontinuity at the frequency band ending points can be avoided by introducing the polynomial argument into sine and cosine functions of the form

$$v_x(x) = x^4(35 - 84x + 70x^2 - 20x^3). \tag{12.17}$$

This polynomial will keep the property that the argument is such that $v_x(0) = 0$ and $v_x(1) = 1$, but it will make the derivatives smooth at the transition points. The Meyer wavelet functions are defined by

$$W(a_i\theta) = \begin{cases} \sin\left(\frac{\pi}{2}v_x\left(q(a_i\theta - 1)\right)\right), \text{ for } 1 < a_i\theta \le M \\ \cos\left(\frac{\pi}{2}v_x\left(q(\frac{a_i\theta}{M} - 1)\right)\right), \text{ for } M < a_i\theta \le M^2 \\ 0, \text{ elsewhere.} \end{cases} \tag{12.18}$$

The same form is used in the first transfer function $W(a_1\theta)$ and the scale function $G(\theta)$. The spectral Meyer wavelet functions with the same parameters as in the previous example, are shown in Fig. 12.3(b). They exhibit smooth transitions and they satisfy the reconstruction condition (12.16).

12.2 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

$$w(m) = \delta(m) + \delta(m-1),$$

with $N = 2$. A two-sample window STFT is

$$STFT(n,0) = \frac{1}{\sqrt{2}}\sum_{m=0}^{1}x(n+m)e^{-j0} = \frac{1}{\sqrt{2}}\left(x(n) + x(n+1)\right) = x_L(n), \tag{12.19}$$

for $k = 0$, corresponding to low frequency $\omega = 0$ and

$$x_H(n) = \frac{1}{\sqrt{2}}\left(x(n) - x(n+1)\right) \tag{12.20}$$

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 anymore. 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 used in further processing. After the signal samples $x(n)$ and $x(n+1)$ are processed using (12.19) and (12.20), then next two samples $x(n+2)$ and $x(n+3)$ are analyzed. The highpass part is again calculated $x_H(n+2) = (x(n+2) - x(n+3))/\sqrt{2}$ and kept as it is. Lowpass part $x_L(n+2) = (x(n+2) + x(n+3))/\sqrt{2}$ 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

$$\begin{bmatrix} \sqrt{2}W_1(0,H) \\ \sqrt{2}W_1(2,H) \\ \sqrt{2}W_1(4,H) \\ \sqrt{2}W_1(6,H) \\ 2W_2(0,H) \\ 2W_2(4,H) \\ 2\sqrt{2}W_4(0,H) \\ 2\sqrt{2}W_4(0,L) \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \\ x(4) \\ x(5) \\ x(6) \\ x(7) \end{bmatrix}. \tag{12.21}$$

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 bellow an assumed level of noise can be zero-valued and the signal-to-noise ratio in the reconstructed signal improved.

12.2.1 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 (12.19) and (12.20), based on the input signal $x(n)$. Transfer functions of the

discrete-time systems producing these two signals are

$$H_L(z) = \frac{1}{\sqrt{2}}(1+z) \qquad (12.22)$$
$$H_H(z) = \frac{1}{\sqrt{2}}(1-z).$$

Frequency responses of these systems assume the form

$$H_L(e^{j\omega}) = \frac{1}{\sqrt{2}}\left(1+e^{j\omega}\right)$$
$$H_H(e^{j\omega}) = \frac{1}{\sqrt{2}}\left(1-e^{j\omega}\right)$$

with amplitude characteristics $\left|H_L(e^{j\omega})\right| = \sqrt{2}|\cos(\omega/2)|$, and $\left|H_H(e^{j\omega})\right| = \sqrt{2}|\sin(\omega/2)|$, presented in Fig.12.4. 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.

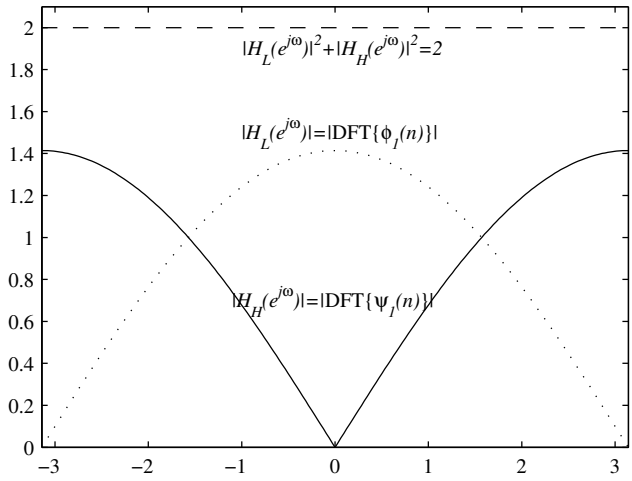


Figure 12.4 Amplitude of the Fourier transform of basic Haar wavelet and scale function divided by $\sqrt{2}$.

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

$$s_L(n) = x_L(2n)$$
$$s_H(n) = x_H(2n). \qquad (12.23)$$

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

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

$$X(z^{1/2}) + X(-z^{1/2}) = \sum_{n=-\infty}^{\infty} x(n)[(z^{-1/2})^n + (-z^{-1/2})^n] = \sum_{n=-\infty}^{\infty} 2x(2n)z^{-n}$$

$$\mathcal{Z}\{x(2n))\} = Y(z) = \frac{1}{2}X(z^{1/2}) + \frac{1}{2}X(-z^{1/2}). \tag{12.24}$$

For the signals $s_L(n) = x_L(2n)$ and $s_H(n) = x_H(2n)$ the system implementation is presented in Fig.12.5.

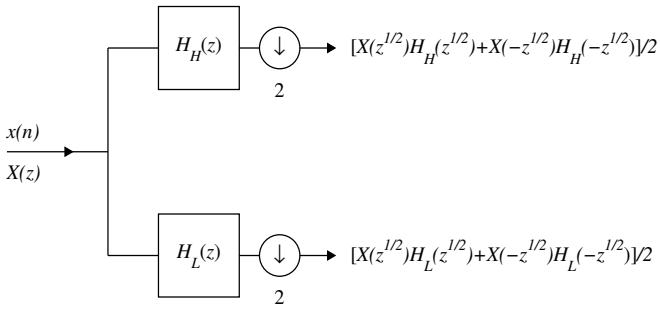


Figure 12.5 Signal filtering by a low pass and a high pass filter followed by downsaampling by 2.

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.

12.2.2 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

$$y(n) = [\ldots x(-2),\ 0,\ x(-1),\ 0,\ x(0),\ 0,\ x(1),0,\ x(2),\ 0,\ \ldots].$$

Its z-transform domain form is

$$Y(z) = X(z^2),$$

since

$$X(z^2)=\sum_{n=-\infty}^{\infty}x(n)z^{-2n}=\ldots x(-1)z^2+0\cdot z^1+x(0)+0\cdot z^{-1}+x(1)z^{-2}+\ldots. \qquad (12.25)$$

Upsampling of a signal $x(n)$ is defined by

$$y(n)=\left\{\begin{array}{ccc} x(n/2) & \text{for} & \text{even } n \\ 0 & \text{for} & \text{odd } n \end{array}\right. =\mathcal{Z}^{-1}\{X(z^2))\}.$$

If a signal $x(n)$ is downsampled first and then upsampled, the resulting signal transform is

$$Y(z)=\frac{1}{2}X(\left(z^{1/2}\right)^2)+\frac{1}{2}X(-\left(z^{1/2}\right)^2)$$
$$Y(z)=\frac{1}{2}X(z)+\frac{1}{2}X(-z). \qquad (12.26)$$

In the Fourier domain it means $Y(e^{j\omega})=(X\left(e^{j\omega}\right)+X\left(e^{j(\omega+\pi)}\right)$. This form indicates that an aliasing component $X\left(e^{j(\omega+\pi)}\right)$ appeared in this process.

12.2.3 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 (12.26). 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.

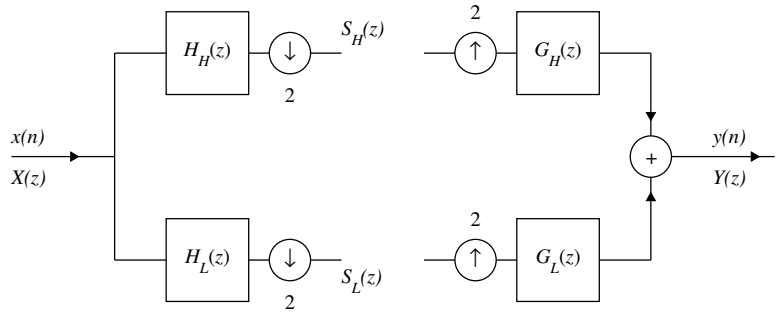


Figure 12.6 One stage of the filter bank with reconstruction, corresponding to the one stage of the wavelet transform realization.

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.12.6. The output signal transforms are

$$Y_L(z)=S_L(z^2)G_L(z)=[\frac{1}{2}H_L(z)X(z)+\frac{1}{2}H_L(-z)X(-z)]G_L(z)$$
$$Y_H(z)=S_H(z^2)G_H(z)=[\frac{1}{2}H_H(z)X(z)+\frac{1}{2}H_H(-z)X(-z)]G_H(z)$$

The total output is

$$\begin{aligned} Y(z) &= Y_L(z) + Y_H(z) \\ &= [\frac{1}{2}H_L(z)G_L(z) + \frac{1}{2}H_H(z)G_H(z)]X(z) \\ &+[\frac{1}{2}H_L(-z)G_L(z) + \frac{1}{2}H_H(-z)G_H(z)]X(-z). \end{aligned}$$

Condition for the alias-free reconstruction is

$$Y(z) = X(z).$$

This means that

$$H_L(z)G_L(z) + H_H(z)G_H(z) = 2 \tag{12.27}$$
$$H_L(-z)G_L(z) + H_H(-z)G_H(z) = 0. \tag{12.28}$$

These are general conditions for a correct (alias-free) signal reconstruction.

Based on the reconstruction conditions we can show that the lowpass filters satisfy

$$H_L(z)G_L(z) + H_L(-z)G_L(-z) = 2 \tag{12.29}$$
$$P(z) + P(-z) = 2, \tag{12.30}$$
$$\text{where } P(z) = H_L(z)G_L(z).$$

From (12.28) 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 (12.28) with z being replaced by $-z$, when $H_L(z)G_L(-z) + H_H(z)G_H(-z) = 0$. Substituting these values into (12.27) 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 (12.27) with z being replaced by $-z$) then

$$\frac{H_L(z)G_L(z)}{H_H(-z)G_H(-z)} = 1 \tag{12.31}$$

and (12.29) follows with

$$H_H(z)G_H(z) = H_L(-z)G_L(-z).$$

In the Fourier transform domain the reconstruction conditions are

$$H_L(e^{j\omega})G_L(e^{j\omega}) + H_H(e^{j\omega})G_H(e^{j\omega}) = 2 \tag{12.32}$$
$$H_L(-e^{j\omega})G_L(e^{j\omega}) + H_H(-e^{j\omega})G_H(e^{j\omega}) = 0.$$

12.2.4 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

$$\begin{aligned} &\langle h_L(m), h_L(m-2n)\rangle = \delta(n) \\ &\sum_m h_L(m)h_L(m-2n) = \delta(n). \end{aligned} \tag{12.33}$$

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 downsampled convolution of $h_L(n)$ and $h_L(-n)$

$$r(n) = h_L(n) * h_L(-n) = \sum_m h_L(m)h_L(m-n),$$

$$\mathcal{Z}\{r(n))\} = H_L(z)H_L(z^{-1}) \quad \text{or} \quad \text{FT}\{r(n))\} = \left|H_L(e^{j\omega})\right|^2.$$

The Fourier transform of the downsampled convolution, for real-valued $h_L(n)$ is, (12.24)

$$\text{FT}\{r(2n))\} = \frac{1}{2}\left|H_L(e^{j\omega/2})\right|^2 + \frac{1}{2}\left|H_L(-e^{j\omega/2})\right|^2.$$

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 (12.33), 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 (12.29) is

$$\begin{aligned} h_L(n) * g_L(n) + [(-1)^n h_L(n)] * [(-1)^n g_L(n)] &= 2\delta(n) \\ \sum_m h_L(m)g_L(n-m) + \sum_m (-1)^n h_L(m)g_L(n-m) &= 2\delta(n) \\ \sum_m h_L(m)g_L(2n-m) &= \delta(n). \end{aligned}$$

If the impulse response $h_L(n)$ is orthogonal, as in (12.33), 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 (12.29) in the form

$$G_L(z)G_L(z^{-1}) + G_L(-z)G_L(-z^{-1}) = 2 \tag{12.34}$$

or $P(z) + P(-z) = 2$ with $P(z) = G_L(z)G_L(z^{-1})$. Relation (12.34) may also written for $H_L(z)$

$$H_L(z)H_L(z^{-1}) + H_L(-z)H_L(-z^{-1}) = 2.$$

12.2.5 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

$$g_L(n) = h_L(-n) = \sum_{k=0}^{K-1} h_k \delta(n-k), \quad G_L(e^{j\omega}) = H_L(e^{-j\omega})$$

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 (12.27) and (12.31), 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 (12.34), with appropriate design conditions, then

$$\begin{aligned} H_L(e^{j\omega}) &= G_L(e^{-j\omega}) \\ G_H(e^{j\omega}) &= -e^{-j\omega K} G_L(-e^{-j\omega}) \\ H_H(e^{j\omega}) &= -e^{j\omega K} G_L(-e^{j\omega}). \end{aligned} \tag{12.35}$$

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

$$\begin{aligned}\sum_m h_L(m)h_H(m-2n) &= 0\\ \sum_m g_L(m)g_H(m-2n) &= 0\end{aligned} \tag{12.36}$$

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 (12.32)

$$H_L(-e^{j\omega})G_L(e^{j\omega}) + H_H(-e^{j\omega})G_H(e^{j\omega}) = 0$$

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

$$G_L(-e^{-j\omega})G_L(e^{j\omega}) - e^{j\omega K}G_L(e^{j\omega})e^{-j\omega K}G_L(-e^{-j\omega}) = 0.$$

12.2.6 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

$$g_L(n) = \frac{1}{\sqrt{2}}\left[\delta(n) + \delta(n-1)\right].$$

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

$$\begin{aligned}H_L(z) &= \frac{1}{\sqrt{2}}(1+z)\\ H_H(z) &= \frac{1}{\sqrt{2}}(1-z).\end{aligned}$$

The reconstruction filters are obtained from (12.27)-(12.28)

$$\begin{aligned}\frac{1}{\sqrt{2}}(1+z)G_L(z) + \frac{1}{\sqrt{2}}(1-z)G_H(z) &= 2\\ \frac{1}{\sqrt{2}}(1-z)G_L(z) + \frac{1}{\sqrt{2}}(1+z)G_H(z) &= 0\end{aligned}$$

as

$$G_L(z) = \frac{1}{\sqrt{2}}\left(1 + z^{-1}\right) \tag{12.37}$$
$$G_H(z) = \frac{1}{\sqrt{2}}\left(1 - z^{-1}\right)$$

with

$$g_L(n) = \frac{1}{\sqrt{2}}\delta(n) + \frac{1}{\sqrt{2}}\delta(n-1) \tag{12.38}$$
$$g_H(n) = \frac{1}{\sqrt{2}}\delta(n) - \frac{1}{\sqrt{2}}\delta(n-1).$$

The values impulse responses in the Haar wavelet transform (relations (12.22) and (12.38)) are:

n	$\sqrt{2}h_L(n)$	$\sqrt{2}h_H(n)$	n	$\sqrt{2}g_L(n)$	$\sqrt{2}g_H(n)$
0	1	1	0	1	1
-1	1	-1	1	1	-1

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 (12.23) are upsampled, according to (12.25), as

$$r_L(n) = [s_L(0)\ 0\ s_L(1)\ 0\ s_L(2)\ 0 \ldots s_L(N-1)\ 0]$$
$$r_H(n) = [s_H(0)\ 0\ s_H(1)\ 0\ s_H(2)\ 0 \ldots s_H(N-1)\ 0]$$

These signals are then passed trough the reconstruction filters. A sum of the outputs from these filters is

$$y(n) = r_L(n) * g_L(n) + r_H(n) * g_H(n)$$
$$= \frac{1}{\sqrt{2}}r_L(n) + \frac{1}{\sqrt{2}}r_L(n-1) + \frac{1}{\sqrt{2}}r_H(n) - \frac{1}{\sqrt{2}}r_H(n-1)$$
$$= \frac{1}{\sqrt{2}}[x_L(0)\ 0\ x_L(2)\ 0\ x_L(4)\ldots.0\ x_L(2N-2)\ 0] + \frac{1}{\sqrt{2}}[0\ x_L(0)\ 0\ x_L(2)\ldots.0\ x_L(2N-2)]$$
$$+\frac{1}{\sqrt{2}}[x_H(0)\ 0\ x_H(2)\ 0\ x_H(4)\ldots.0\ x_H(2N-2)\ 0] - \frac{1}{\sqrt{2}}[0\ x_H(0)\ 0\ x_H(2)\ldots.0\ x_H(2N-2)].$$

where $s_L(n) = x_L(2n)$ and $s_H(n) = x_H(2n)$. From the previous relation follows

$$y(0) = \frac{1}{\sqrt{2}}[x_L(0) + x_H(0)] = x(0)$$
$$y(1) = \frac{1}{\sqrt{2}}[x_L(0) - x_H(0)] = x(1)$$
$$\ldots$$
$$y(2n) = \frac{1}{\sqrt{2}}[x_L(2n) + x_H(2n)] = x(2n)$$
$$y(2n+1) = \frac{1}{\sqrt{2}}[x_L(2n) - x_H(2n)] = x(2n+1).$$

A system for the implementation of the Haar wavelet transform of a signal with eight samples is presented in Fig.12.7. It corresponds to the matrix form realization (12.21).

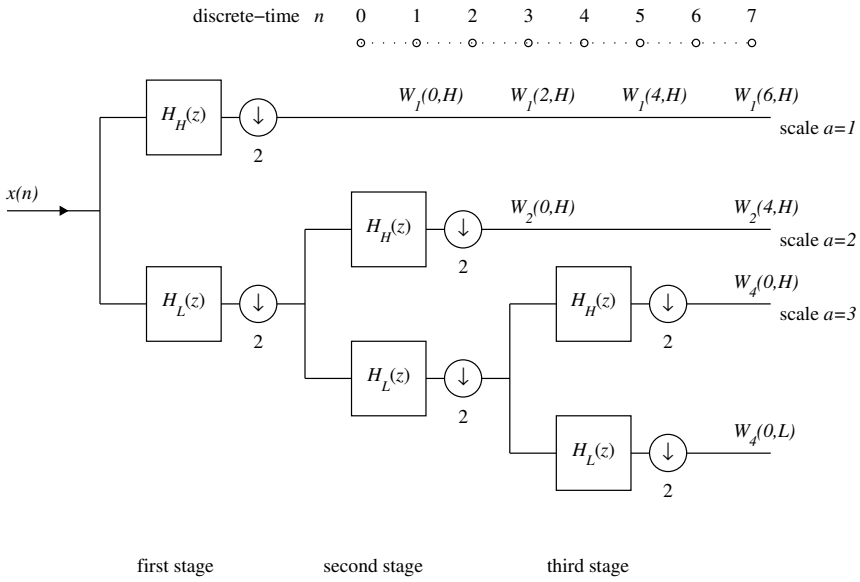


Figure 12.7 Filter bank for the wavelet transform realization

Example 12.2. For a signal $x(n) = [1, 1, 2, 0, 2, 2, 0, 0, 2, 2, 2, 2, 0, 0, 0, 0]$ 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.12.8(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 into its lowpass and highpass region $W_2(n,L) = [s_2(n) + s_2(n+1)]/\sqrt{2}$ and $W_2(n,H) = [s_2(n) - s_2(n+1)]/\sqrt{2}$, respectively, Fig.12.8(b). The same calculation is performed in the third and fourth stage, Fig.12.8(c) - (d).

12.2.7 Daubechies D4 Wavelet Transform

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.12.4 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 $h_L(n) = [h_L(0), h_L(-1), h_L(-2), h_L(-3)] = [h_0, h_1, h_2, h_3]$.

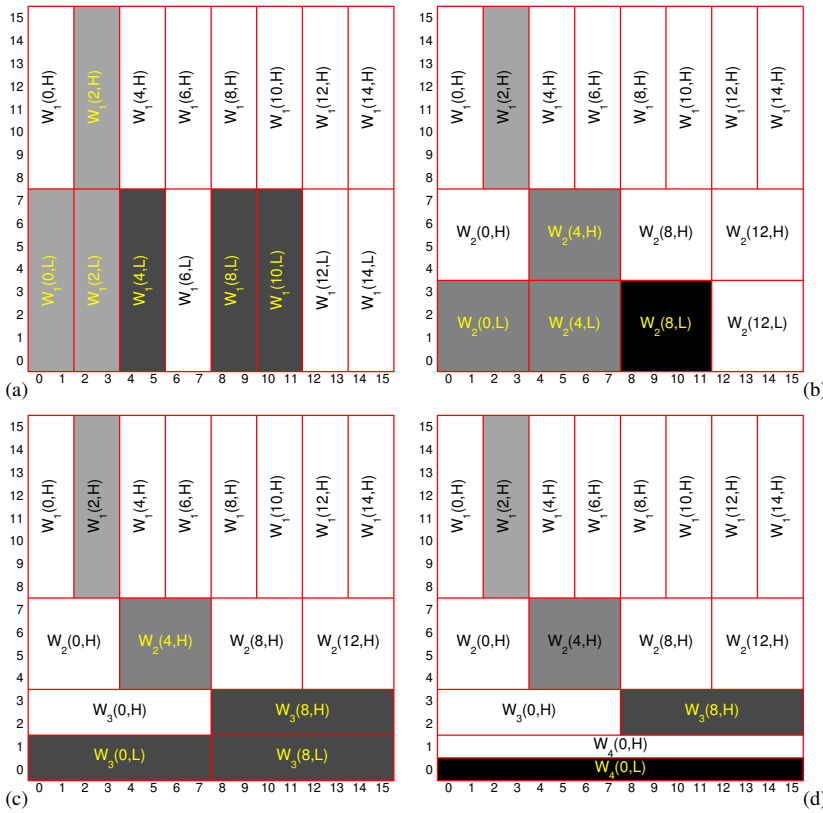


Figure 12.8 Wavelet transform of a signal with $M = 16$ samples at the output of stages 1, 2, 3 and 4, respectively. Notation $W_a(n,H)$ is used for the highpass value of coefficient after stage (scale of) a at an instant n. Notation $W_a(n,L)$ is used for the lowpass value of coefficient after stage (scale of) a at an instant n.

If the highpass and reconstruction filter coefficients are chosen such that

$$\begin{array}{cccccc} n & h_L(n) & h_H(n) & n & g_L(n) & g_H(n) \\ 0 & h_0 & h_3 & 0 & h_0 & h_3 \\ -1 & h_1 & -h_2 & 1 & h_1 & -h_2 \\ -2 & h_2 & h_1 & 2 & h_2 & h_1 \\ -3 & h_3 & -h_0 & 3 & h_3 & -h_0 \end{array} . \tag{12.39}$$

then relation (12.35) 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

$$\begin{aligned}
&H_L(z)G_L(z)+H_H(z)G_H(z)\\
&=\left(h_0+h_1z+h_2z^2+h_3z^3\right)\left(h_0+h_1z^{-1}+h_2z^{-2}+h_3z^{-3}\right)\\
&+\left(-h_0z^3+h_1z^2-h_2z+h_3\right)\left(-h_0z^{-3}+h_1z^{-2}-h_2z^{-1}+h_3\right)\\
&=2(h_0^2+h_1^2+h_2^2+h_3^2)=2
\end{aligned}$$

and

$$\begin{aligned}
&H_L(-z)G_L(z)+H_H(-z)G_H(z)\\
&=\left(h_0-h_1z+h_2z^2-h_3z^3\right)\left(h_0+h_1z^{-1}+h_2z^{-2}+h_3z^{-3}\right)\\
&+\left(h_0z^3+h_1z^2+h_2z+h_3\right)\left(-h_0z^{-3}+h_1z^{-2}-h_2z^{-1}+h_3\right)=0.
\end{aligned}$$

For the calculation of impulse response values h_0, h_1, h_2, h_3 of a fourth order system (12.39) 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

$$\left.\frac{dH_L(e^{j\omega})}{d\omega}\right|_{\omega=\pi}=0.$$

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,

$$\begin{aligned}
&h_0+h_1+h_2+h_3=\sqrt{2} \text{ from } H_L(e^{j0})=\sqrt{2}\\
&h_0^2+h_1^2+h_2^2+h_3^2=1 \text{ reconstruction condition}\\
&h_0-h_1+h_2-h_3=0 \text{ from } H_L(e^{j\pi})=0\\
&-h_1+2h_2-3h_3=0 \text{ from } \left.\frac{dH_L(e^{j\omega})}{d\omega}\right|_{\omega=\pi}=0.
\end{aligned}$$

Its solution produces the fourth order Daubechies wavelet coefficients (D4)

n	$h_L(n)$	$h_H(n)$	n	$g_L(n)$	$g_H(n)$
0	$\frac{1+\sqrt{3}}{4\sqrt{2}}$	$\frac{1-\sqrt{3}}{4\sqrt{2}}$	0	$\frac{1+\sqrt{3}}{4\sqrt{2}}$	$\frac{1-\sqrt{3}}{4\sqrt{2}}$
-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	$\frac{1-\sqrt{3}}{4\sqrt{2}}$	$-\frac{1+\sqrt{3}}{4\sqrt{2}}$

Note that this is just one of possible symmetric solutions of the previous system of equations, Fig.12.9.

The reconstruction conditions for the fourth order FIR filter

$$H_L(e^{j\omega})=h_0+h_1e^{j\omega}+h_2e^{j2\omega}+h_3e^{j3\omega}$$

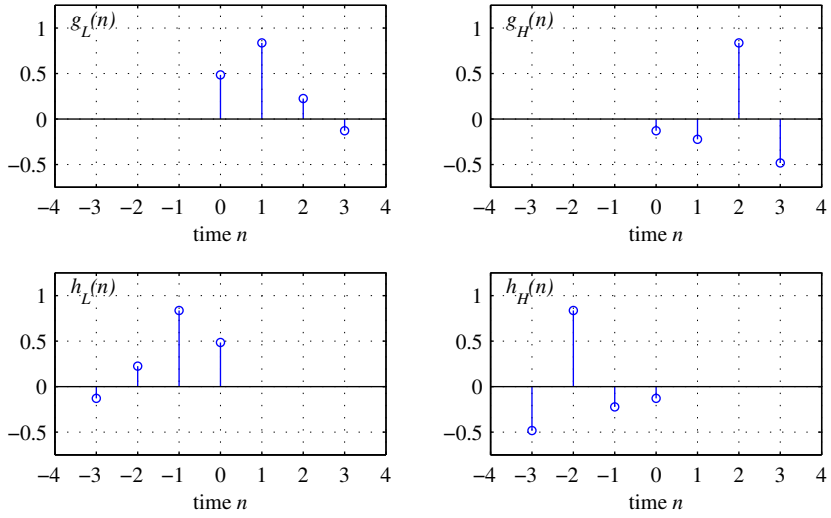


Figure 12.9 Impulse responses of the D4 filters.

with Daubechies wavelet coefficients (D4) can also be checked in a graphical way by calculating

$$\left|H_L(e^{j\omega})\right|^2 + \left|H_L(e^{j(\omega+\pi)})\right|^2 = 2$$
$$H_L(e^{j(\omega+\pi)})H_L^*(e^{j\omega}) + H_L(e^{j\omega})H_L^*(e^{j(\omega+\pi)}) = 0.$$

From Fig.12.10, we can see that it is much better approximation of low and high pass filters than in the Haar wavelet case, Fig.12.4.

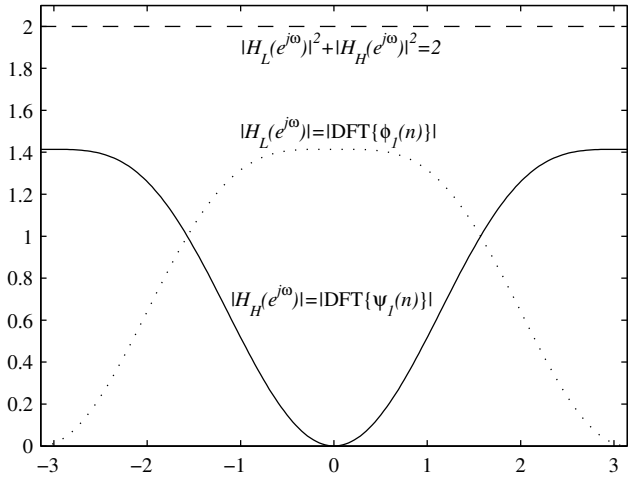


Figure 12.10 Amplitude of the Fourier transform of basic Daubechies D4 wavelet and scale function.

Another way to derive Daubechies wavelet coefficients (D4) is in using relation (12.34)

$$P(z) + P(-z) = 2$$

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 approach to the highpass zero value)

$$G_L(e^{j\omega})\Big|_{\omega=\pi} = 0$$
$$\frac{dG_L(e^{j\omega})}{d\omega}\Bigg|_{\omega=\pi} = 0.$$

Then $G_L(z)$ must contain a factor of the form $\left(1+z^{-1}\right)^2$. Since the filter order must be even (K must be odd), taking into account that $\left(1+z^{-1}\right)^2$ 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

$$R(z) = \left[a(1+z_1z^{-1})\right]\left[a(1+z_1z^{1})\right] = z_0z^{-1} + b + z_0z.$$

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

$$g_L(n) = \frac{1}{4\sqrt{2}}\left[\left(1+\sqrt{3}\right)\delta(n) + \left(3+\sqrt{3}\right)\delta(n-1) + \left(3-\sqrt{3}\right)\delta(n-2) + \left(1-\sqrt{3}\right)\delta(n-3)\right].$$

All other impulse responses follow from this one (as in the presented table).

Example 12.3. Consider a signal that is a linear function of time

$$x(n) = an + b.$$

Show that the condition

$$-h_L(-1) + 2h_L(-2) - 3h_L(-3) = 0 \text{ following from } \left.\frac{dH_L(e^{j\omega})}{d\omega}\right|_{\omega=\pi} = 0$$

is equivalent to the condition that highpass coefficients (output from $H_H(e^{j\omega})$) are zero-valued, Fig.12.10. 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

$$\begin{aligned} W_1(n,H) &= x(n) * h_H(n) \\ &= x(n)h_H(0) + x(n+1)h_H(-1) + x(n+2)h_H(-2) + x(n+3)h_H(-3) \\ &= x(n)h_3 - x(n+1)h_2 + x(n+2)h_1 - x(n+3)h_0 \\ &= (an+b)h_3 - ((n+1)a+b)h_2 + ((n+2)a+b)h_1 - ((n+3)a+b)h_0 \\ &= (a(n+3)+b)(-h_0 + h_1 - h_2 + h_3) - a(h_1 - 2h_2 + 3h_3) = 0 \end{aligned}$$

if

$$\begin{aligned} -h_0 + h_1 - h_2 + h_3 &= 0 \text{ and} \\ h_1 - 2h_2 + 3h_3 &= 0. \end{aligned}$$

The lowpass coefficients are obtained by downsampling

$$\begin{aligned} W_1(n,L) &= x(n) * h_L(n) \\ &= x(n)h_0 + x(n+1)h_1 + x(n+2)h_2 + x(n+3)h_3 \\ &= (an+b)h_0 + ((n+1)a+b)h_1 + ((n+2)a+b)h_2 + ((n+3)a+b)h_3 \\ &= (an+b)(h_0 + h_1 + h_2 + h_3) + a(h_1 + 2h_2 + 3h_3) \\ &= a_1 n + b_1 \end{aligned}$$

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 12.4. 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

$$\begin{matrix} h_0 & h_1 & h_2 & h_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & h_0 & h_1 & h_2 & h_3 & 0 & 0 \end{matrix}$$

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

$$\begin{bmatrix} W_1(0,L) \\ W_1(0,H) \\ W_1(2,L) \\ W_1(2,H) \\ W_1(4,L) \\ W_1(4,H) \\ W_1(6,L) \\ W_1(6,H) \end{bmatrix} = \begin{bmatrix} h_0 & h_1 & h_2 & h_3 & 0 & 0 & 0 & 0 \\ h_3 & -h_2 & h_1 & -h_0 & 0 & 0 & 0 & 0 \\ 0 & 0 & h_0 & h_1 & h_2 & h_3 & 0 & 0 \\ 0 & 0 & h_3 & -h_2 & h_1 & -h_0 & 0 & 0 \\ 0 & 0 & 0 & 0 & h_0 & h_1 & h_2 & h_3 \\ 0 & 0 & 0 & 0 & h_3 & -h_2 & h_1 & -h_0 \\ h_2 & h_3 & 0 & 0 & 0 & 0 & h_0 & h_1 \\ h_1 & -h_0 & 0 & 0 & 0 & 0 & h_3 & -h_2 \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \\ x(4) \\ x(5) \\ x(6) \\ x(7) \end{bmatrix}. \tag{12.40}$$

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 12.5. 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.

★ 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 12.6. 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.12.7) are

$$\begin{aligned} W_1(2n,H) &= x(2n)h_H(0) + x(2n+1)h_H(-1) \\ &+x(2n+2)h_H(-2) + x(2n+3)h_H(-3) \\ &= x(2n)h_3 - x(2n+1)h_2 + x(2n+2)h_1 - x(2n+3)h_0 \end{aligned}$$

with

$$[h_3,h_2,h_1,h_0] = [\frac{1-\sqrt{3}}{4\sqrt{2}}, \frac{3-\sqrt{3}}{4\sqrt{2}}, \frac{3+\sqrt{3}}{4\sqrt{2}}, \frac{1+\sqrt{3}}{4\sqrt{2}}].$$

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.12.11.

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 $W_1(2n, L) = 1.4142$ 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. 12.2(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.

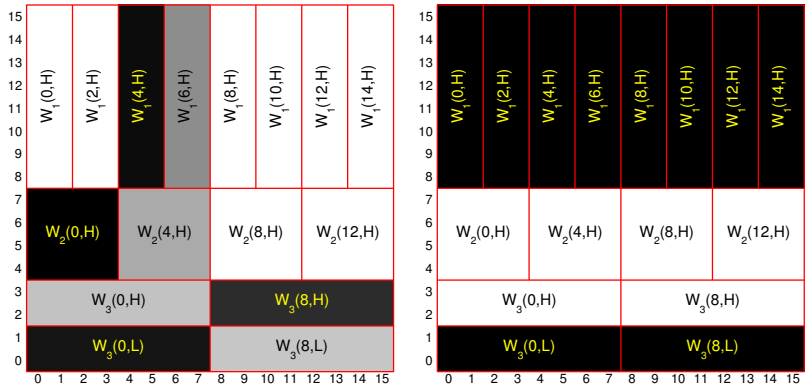


Figure 12.11 Daubechies D4 wavelet transform (absolute value) of the signal $x(n) = \delta(n-7)$ using $N = 16$ signal samples, $0 \leq n \leq N-1$ (left). The Daubechies D4 wavelet transform (absolute value) of the signal $x(n) = 2\cos(2\pi 8n/N) + 1$, $0 \leq n \leq N-1$, with $N = 16$ (right).

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

$$\begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \\ x(4) \\ x(5) \\ x(6) \\ x(7) \end{bmatrix} = \begin{bmatrix} h_0 & h_3 & 0 & 0 & 0 & 0 & h_2 & h_1 \\ h_1 & -h_2 & 0 & 0 & 0 & 0 & h_3 & -h_0 \\ h_2 & h_1 & h_0 & h_3 & 0 & 0 & 0 & 0 \\ h_3 & -h_0 & h_1 & -h_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & h_2 & h_1 & h_0 & h_3 & 0 & 0 \\ 0 & 0 & h_3 & -h_0 & h_1 & -h_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & h_2 & h_1 & h_0 & h_3 \\ 0 & 0 & 0 & 0 & h_3 & -h_0 & h_1 & -h_2 \end{bmatrix} \begin{bmatrix} W_1(0,L) \\ W_1(0,H) \\ W_1(2,L) \\ W_1(2,H) \\ W_1(4,L) \\ W_1(4,H) \\ W_1(6,L) \\ W_1(6,H) \end{bmatrix}. \tag{12.41}$$

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 12.7. For the Wavelet transform from the previous example find its inverse (reconstruct the signal).

★ The inversion is done backwards. From $W_3(0,H)$, $W_3(0,L)$,$W_3(8,H)$, $W_3(8,L)$ we get signal $s_3(n)$ or $W_2(2n,L)$ as

$$\begin{bmatrix} W_2(0,L) \\ W_2(4,L) \\ W_2(8,L) \\ W_2(12,L) \end{bmatrix} = \begin{bmatrix} h_0 & h_3 & h_2 & h_1 \\ h_1 & -h_2 & h_3 & -h_0 \\ h_2 & h_1 & h_0 & h_3 \\ h_3 & -h_0 & h_1 & -h_2 \end{bmatrix} \begin{bmatrix} W_3(0,L) \\ W_3(0,H) \\ W_3(8,L) \\ W_3(8,H) \end{bmatrix}$$

$$= \begin{bmatrix} h_0 & h_3 & h_2 & h_1 \\ h_1 & -h_2 & h_3 & -h_0 \\ h_2 & h_1 & h_0 & h_3 \\ h_3 & -h_0 & h_1 & -h_2 \end{bmatrix} \begin{bmatrix} 0.4668 \\ -0.1251 \\ -0.1132 \\ -0.4226 \end{bmatrix} = \begin{bmatrix} -0.1373 \\ 0.6373 \\ 0 \\ 0 \end{bmatrix}.$$

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

$$\begin{bmatrix} W_1(0,L) \\ W_1(2,L) \\ W_1(4,L) \\ W_1(6,L) \\ W_1(8,L) \\ W_1(10,L) \\ W_1(12,L) \\ W_1(14,L) \end{bmatrix} = \begin{bmatrix} h_0 & h_3 & 0 & 0 & 0 & 0 & h_2 & h_1 \\ h_1 & -h_2 & 0 & 0 & 0 & 0 & h_3 & -h_0 \\ h_2 & h_1 & h_0 & h_3 & 0 & 0 & 0 & 0 \\ h_3 & -h_0 & h_1 & -h_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & h_2 & h_1 & h_0 & h_3 & 0 & 0 \\ 0 & 0 & h_3 & -h_0 & h_1 & -h_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & h_2 & h_1 & h_0 & h_3 \\ 0 & 0 & 0 & 0 & h_3 & -h_0 & h_1 & -h_2 \end{bmatrix} \begin{bmatrix} W_2(0,L) \\ W_2(0,H) \\ W_2(4,L) \\ W_2(4,H) \\ W_2(8,L) \\ W_2(8,H) \\ W_2(12,L) \\ W_2(12,H) \end{bmatrix}.$$

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 16×16 order and it is formed using the same structure as the previous transformation matrix.

12.2.8 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.12.12. Note that this case and coefficient in the Haar transform would correspond to $W_4(0,L) = 1$ in (12.21) or in Fig.12.7.

The reconstruction process consists of signal upsampling and passing it trough 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

$$\begin{aligned}
\phi_0(n) &= \delta(n) * g_L(n) = g_L(n) \\
\phi_1(n) &= [\phi_0(0)\ 0\ \phi_0(1)\ 0\ \ \phi_0(2)\ 0\ \phi_0(3)] * g_L(n) \\
\phi_2(n) &= [\phi_1(0)\ 0\ \phi_1(1)\ 0 \ldots\ \phi_1(8)\ 0\ \phi_1(9)] * g_L(n) \\
&\ldots. \\
\phi_{a+1}(n) &= \sum_p \phi_a(p) g_L(n-2p),
\end{aligned}$$

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.12.14. 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

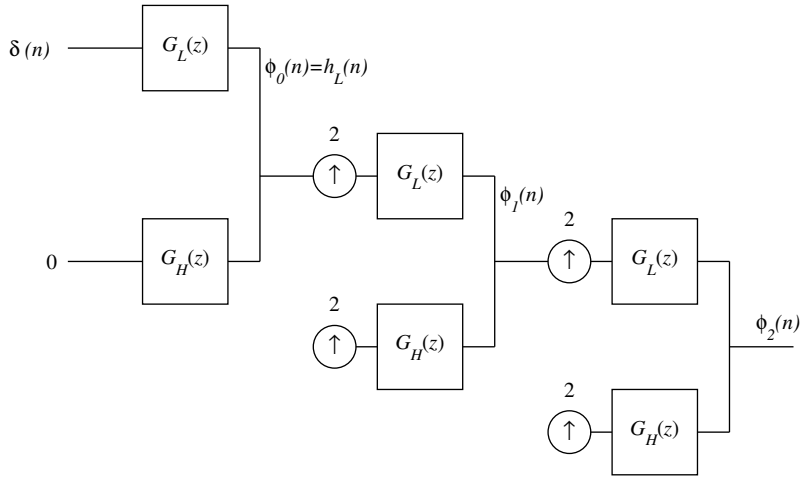


Figure 12.12 Calculation of the upsampled scale function.

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.12.13 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 (12.21) and Fig.12.7 this case would correspond to $W_4(0,H)=1$.

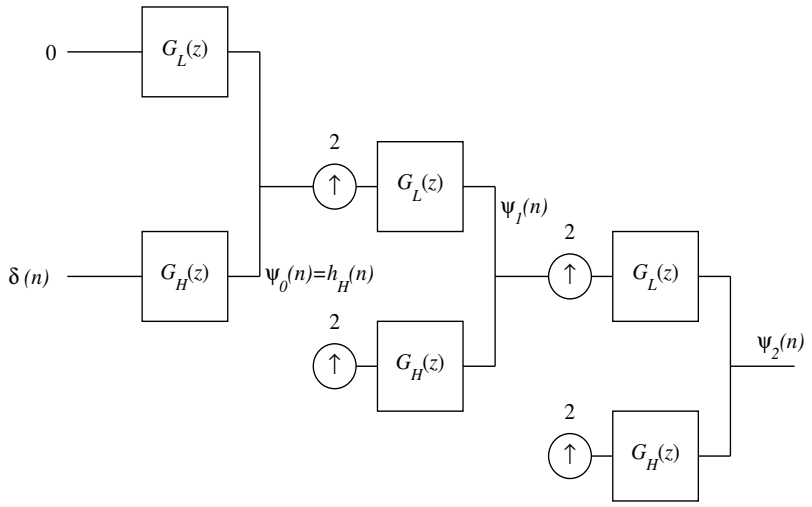


Figure 12.13 Calculation of the upsampled wavelet function

Calculation in the time of the wavelet function in different scales is done using

$$\begin{aligned}
\psi_0(n) &= \delta(n) * g_H(n) = g_H(n) \\
\psi_1(n) &= [\psi_1(0)\ 0\ \psi_1(1)\ 0\ \ \psi_1(2)\ 0\ \psi_1(3)] * g_L(n) \\
\psi_2(n) &= [\psi_2(0)\ 0\ \psi_2(1)\ 0 \ldots\ \psi_2(8)\ 0\ \psi_2(9)] * g_L(n) \\
&\ldots. \\
\psi_{a+1}(n) &= \sum_p \psi_a(p) g_L(n-2p)
\end{aligned}$$

Different scales of the wavelet function are presented in Fig.12.14, normalized using $\psi_a(n)2^{(a+1)/2}$.

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 (12.36).

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 12.8. 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.12.15.

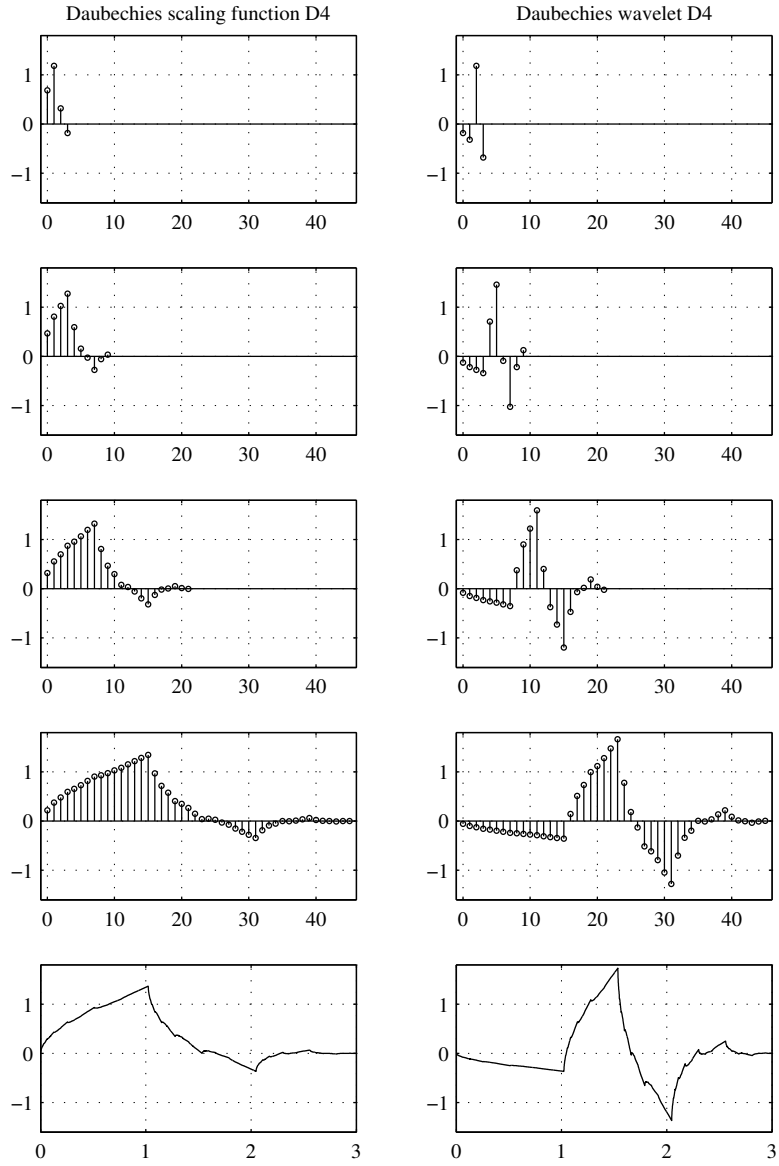


Figure 12.14 The Daubechies D4 wavelet scale function and wavelet calculated using the filter bank relation in different scales: $a = 0$ (first row), $a = 1$ (second row), $a = 2$ (third row), $a = 3$ (fourth row), $a = 10$ (fifth row-approximation of a continuous domain). The amplitudes are scaled by $2^{(a+1)/2}$ 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.

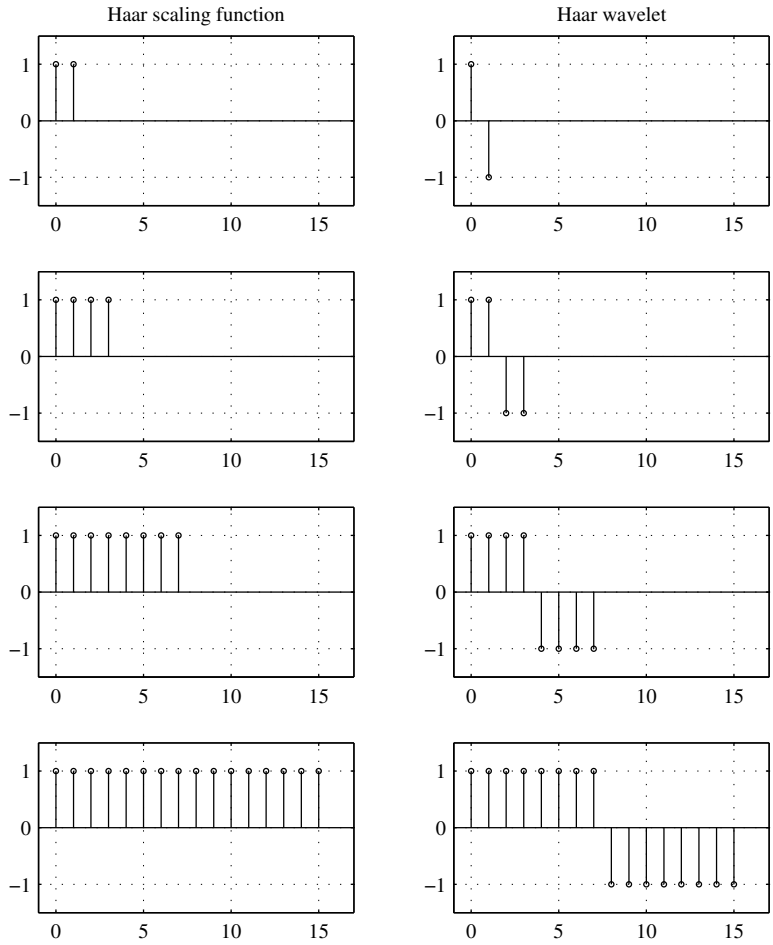


Figure 12.15 The Haar wavelet scale function and wavelet calculated using the filter bank relation in different scales. Values are normalized $2^{(a+1)/2}$.

12.2.9 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

$$\begin{aligned} h_L(n) &= [h_L(0), h_L(-1), h_L(-2), h_L(-3), h_L(-4), h_L(-5)] \\ &= [h_0, h_1, h_2, h_3, h_4, h_5]. \end{aligned}$$

In addition to the conditions $H_L(e^{j0}) = \sqrt{2}$ and $H_L(e^{j\pi}) = 0$, written as

$$\begin{aligned} h_0 + h_1 + h_2 + h_3 + h_4 + h_5 &= \sqrt{2} \\ h_0 - h_1 + h_2 - h_3 + h_4 - h_5 &= 0, \end{aligned}$$

the orthogonality conditions

$$h_0h_2 + h_1h_3 + h_2h_4 + h_3h_5 = 0$$
$$h_0h_4 + h_1h_5 = 0,$$

are added. Since the filter order is 6 then two orthogonality conditions must be used. One for shift 2 and the other for shift 4.

The linear signal cancellation condition is again used as

$$-h_1 + 2h_2 - 3h_3 + 4h_4 - 5h_5 = 0.$$

The final condition in the Daubechies D6 wavelet transform is that the quadratic signal cancellation is achieved for highpass filter, meaning

$$\left.\frac{d^2H_L(e^{j\omega})}{d\omega^2}\right|_{\omega=\pi} = \left.\frac{d^2\left(\sum_{n=0}^{5}h_ne^{j\omega n}\right)}{d\omega^2}\right|_{\omega=\pi} = -\left.\sum_{n=0}^{5}n^2h_ne^{j\omega n}\right|_{\omega=\pi} = 0.$$

This condition is of the form

$$-h_1 + 2^2h_2 - 3^2h_3 + 4^2h_4 - 5^2h_5 = 0$$

From the set of six equations the Daubechies D6 wavelet transform coefficients are obtained as

$$h_L(n) = [1.1411, 0.4705, 0.6504, 0.0498, -0.1208, -0.1909].$$

This is one of possible symmetric solutions of the previous system. From the definition it is obvious that the highpass coefficients will be zero as far as the signal is of quadratic nature within the considered interval. These coefficients can be used as a measure of the signal deviation from the quadratic form in each scale.

Implementation is the same as in the case of Haar or D4 wavelet transform. Only difference is in the filter coefficients form.

This form can be also derived from the reconstruction conditions and the fact that the transfer function $G_L(z)$ contains a factor of the form $(1+z^{-1})^3$ since $z=-1$ is its third order zero, according to the assumptions.

12.2.10 Coifflet Transform

In the Daubechies D6 wavelet transform the last condition is introduced so that the output of high-pass filter is zero when the input signal is quadratic. Another way to form filter coefficients for a six sample wavelet is to introduce the condition that the first moment of the scale function is zero, instead of the second order moment of the wavelet function. In this case symmetric form of coefficients should be used in the definition

$$h_L(-2) + h_L(-1) + h_L(0) + h_L(1) + h_L(2) + h_L(3) = \sqrt{2}$$
$$h_L^2(-2) + h_L^2(-1) + h_L^2(0) + h_L^2(1) + h_L^2(2) + h_L^2(3) = 1$$
$$-2h_L(-2) + h_L(-1) - h_L(1) + 2h_L(2) - 3h_L(3) = 0$$
$$h_L(-2)h_L(0) + h_L(-1)h_L(1) + h_L(0)h_L(2) + h_L(1)h_L(3) = 0$$
$$h_L(-2)h_L(2) + h_L(-1)h_L(3) = 0.$$

The first-order moment of $h_L(n)$ is

$$-2h_L(-2) - h_L(-1) + h_L(1) + 2h_L(2) + 3h_L(3) = 0$$

This is so called sixth order coifflet transform. Its coefficients are

$$h(-2) = (\sqrt{2} - \sqrt{14})/32,$$

$$h(-1) = (-11\sqrt{2} + \sqrt{14})/32,$$

$$h(0) = (7\sqrt{2} + \sqrt{14})/16,$$

$$h(1) = (-\sqrt{2} - \sqrt{14})/16,$$

$$h(2) = (\sqrt{2} - \sqrt{14})/32,$$

$$h(3) = (-3\sqrt{2} + \sqrt{14})/32.$$

12.2.11 Discrete Wavelet Transform - STFT

Originally the wavelet transform was introduced by Morlet as a frequency varying STFT. Its aim was to analyze spectrum of the signal with varying resolution in time and frequency. Higher resolution in frequency was required at low frequencies, while at high frequencies high resolution in time was the aim, for specific analyzed seismic signals.

The Daubechies D4 wavelet/scale function is derived from the condition that the highpass coefficients of a signal with linear change in time ($x(n) = an + b$) are zero-valued. Higher order Daubechies wavelet/scale functions are derived by increasing the order of the signal polynomial changes. Frequency of a signal does not play any direct role in the discrete-wavelet transform definition using Daubechies functions. In this sense it would be easier to relate the wavelet transform to the linear (D4) and higher order interpolations of functions (signals), within the intervals of various lengths (corresponding to various wavelet transform scales), than to the spectral analysis where the harmonic basis functions play the central role.

Example 12.9. Consider a signal $x(n)$ with $M = 16$ samples, $0 \leq n \leq M-1$. Write the Daubechies D4 wavelet transform based decomposition of this signal that will divide the frequency axis into four equal regions.

★ In the STFT a $4-$point ($N-$point) signal would be used to calculate 4 (or N) coefficients of the frequency plane. The wavelet transform divides the time-frequency plane into two regions (high and low) regardless of the number of the signal values (wavelet transform coefficients) being used. If the Haar wavelet is used in Fig.12.16 then by dividing both highpass bands and lowpass bands in the same way the short-time Walsh-Hadamard transform with 4-sample nonoverlapping calculation would be obtained. In the cases of Daubechies 4D wavelet transform, a kind of short time analysis with the Daubechies functions would be obtained. For the Daubechies D4 function the scale 2 functions:

$$\phi_1(n) = h_{LL}(n) = [h_L(0)\ 0\ h_L(1)\ 0\ h_L(2)\ 0\ h_L(3)] * h_L(n) \tag{12.42}$$

$$\varphi_1(n) = h_{LH}(n) = [h_H(0)\ 0\ h_H(1)\ 0\ h_H(2)\ 0\ h_H(3)] * h_L(n)$$

$$\psi_1(n) = h_{HL}(n) = [h_L(0)\ 0\ h_L(1)\ 0\ h_L(2)\ 0\ h_H(3)] * h_H(n)$$

$$\varkappa_1(n) = h_{HH}(n) = [h_H(0)\ 0\ h_H(1)\ 0\ h_H(2)\ 0\ h_H(3)] * h_H(n) \tag{12.43}$$

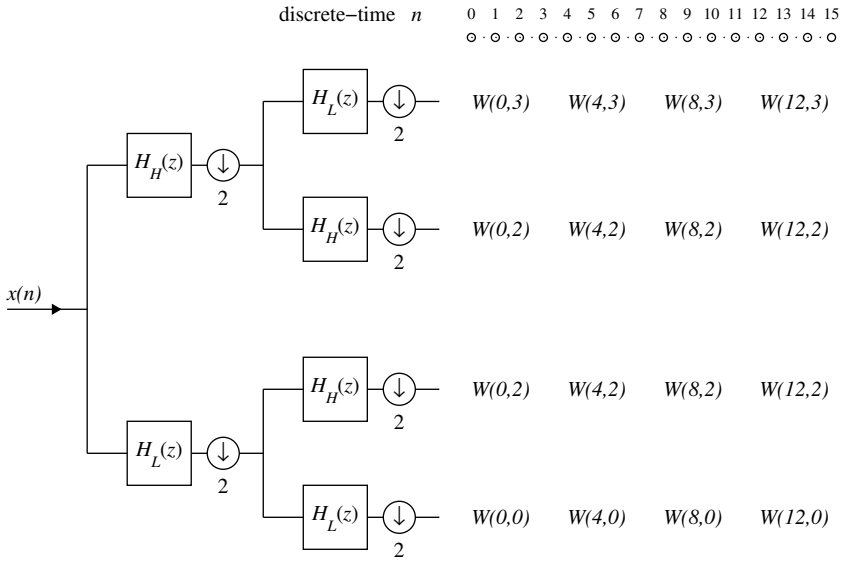


Figure 12.16 Full coverage of the time-frequency plane using the filter bank calculation and systems with impulse responses corresponding to the wavelet transformation.

would be used to calculate $W(4n,0)$, $W(4n,1)$, $W(4n,2)$, and $W(4n,3)$, Fig.12.17. The asymmetry of the frequency regions is visible.

Note that the STFT analysis of this case, with a Hann(ing) window of $N = 8$ and calculation step $R = 4$ will result in the same number of instants, however the frequency range will be divided in 8 regions, having a finer grid. This grid is redundant with respect to the signal and to the wavelet transform. Both, the signal and the wavelet transform have 16 values (coefficients).

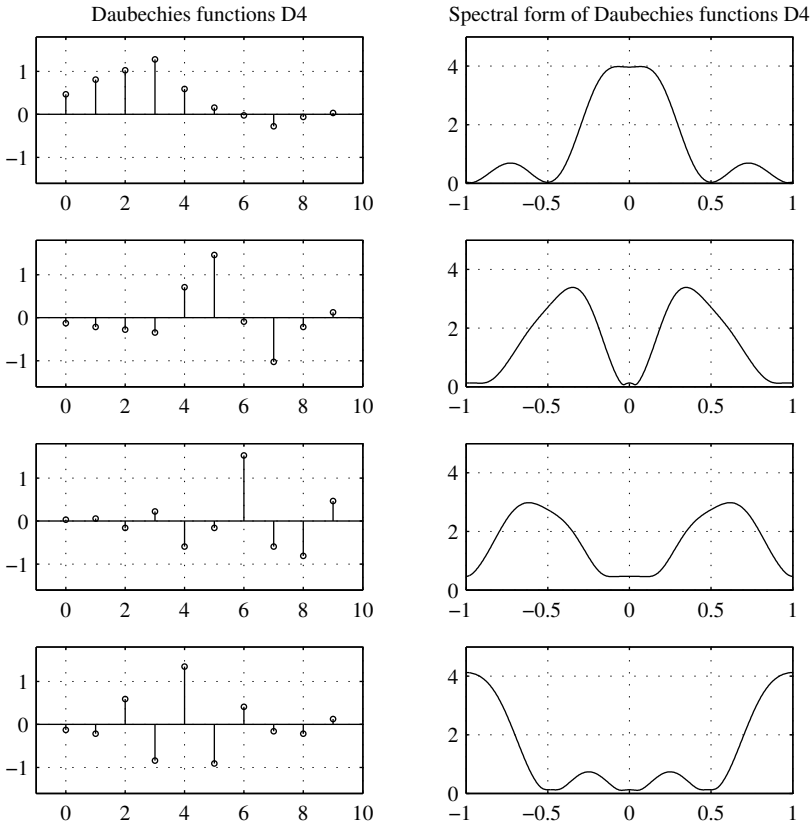


Figure 12.17 Daubechies functions: Scaling function (first row), Mother wavelet function (second row), Function producing the low-frequency part in the second stage of the high frequency part in the first stage (third), Function producing the high-frequency part in the second stage of the high frequency part in the first stage (fourth). Time domain forms of the functions are left while its spectral content is shown on the right.

Part VI

Sparse Signal Processing and Compressive Sensing

Chapter 13

Sensing of Sparse Signals

A discrete-time signal can be transformed into various domains using different signal transformations. Some signals that cover the whole considered interval in one domain could have only a few nonzero coefficients in a transformation domain. These signals are sparse in the considered transformation domain. An observation or measurement of a sparse signal is a linear combination of the sparsity domain coefficients. Since the signal samples are linear combinations of the signal transform coefficients they can be considered as the measurements of a sparse signal in the respective transformation domain.

Compressive sensing is a field dealing with a model for data acquisition including the problem of sparse signal recovery from a reduced set of measurements. A reduced set of measurements 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 the complete set of measurements. 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 from a reduced set of samples. Although the reduced set of measurements appears in the first case as a result of the user strategy to compress 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 obtained with a reduced set of measurements/samples, as in the case if a complete set of measurements/samples were available. A priori information about the sparse nature of the analyzed signal in a known transformation domain must be used in this analysis. Sparsity 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 its transformations domains. The DFT will be used as a case study. The compressive sensing results and algorithms are presented and used as a tool to solve engineering problems involving sparse signals.

13.1 ILLUSTRATIVE EXAMPLES

Before we start the analysis we will describe two simple examples that can be interpreted and solved within the context of sparse signal processing and compressive sensing.

Consider a large set of real numbers $X(0), X(1), \ldots, X(N-1)$. Assume that only one of them is nonzero (or different from a common and known expected value). We do not know either its position

Authors: *Ljubiša Stanković, Miloš Daković, Srdjan Stanković, Irena Orović*

or its value. The aim is to find the position and the value of this nonzero number. The nonzero valued sample will be denoted by $X(i)$. A direct way to find the position of nonzero sample would be to perform up to N measurements and to check which sample assumes a nonzero value. However, if N is very large and there is only one nonzero sample we can get the result using just a few measurements. A procedure to solve the problem with a reduced number of measurements is described next.

Take random numbers as weighting coefficients $a_0(k)$, $k = 0,1,2,\ldots,N-1$, for each sample. Measure the total value of all N weighted samples. Since only one of them is different from zero we will get the measurement

$$y(0) = \sum_{k=0}^{N-1} a_0(k)X(k) = a_0(i)X(i). \tag{13.1}$$

The same value will be obtained if there is only one sample different from the common and known expected value m of all other samples. Then, we will get the total measured value

$$M = a_1 m + a_2 m + \cdots + a_i(m + X(i)) + \cdots + a_N m.$$

After we subtract the expected value $M_T = (a_1 + a_2 + \cdots + a_N)m$ from M, the same measurement $y(0)$ as in (13.1) follows.

As an illustration consider a set of N bags. Assume that only one bag contains all false coins whose weight is $m + X(i)$. It is different from the known weight m of the true coins. The goal is to find the position, i, and the difference in weight, $X(i)$, of the false coins. From each of N bags we will take $a_k(0)$, $k = 0,1,\ldots N-1$, coins, respectively. The number of coins from the kth bag is denoted by $a_k(0)$. The total measured weight of all coins from N bags is M, Fig.13.1. After the expected value is subtracted, the measurement $y(0)$ value is obtained as

$$y(0) = \sum_{k=0}^{N-1} X(k)a_k(0). \tag{13.2}$$

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 an unknown position $k = i$. The cross-section of hyperplane (13.2) with any of the coordinate axes could be a solution to our problem Fig.13.2(a). Assuming that a single $X(k)$ is nonzero, a solution will exist for any k. Thus, one measurement would produce a set of N possible single nonzero values equal to

$$X(k) = y(0)/a_k(0), \quad a_k(0) \neq 0, k = 0,1,2,\ldots,N-1.$$

As expected, from one measurement we are not able to solve the problem (to find the position and the value of one nonzero sample).

If we perform one more measurement, with another set of weighting coefficients $a_k(1)$, $k = 0,1,\ldots,N-1$, and get measurement $y(1) = X(i)a_i(1)$, the result will be another hyperplane, Fig.13.2(b),

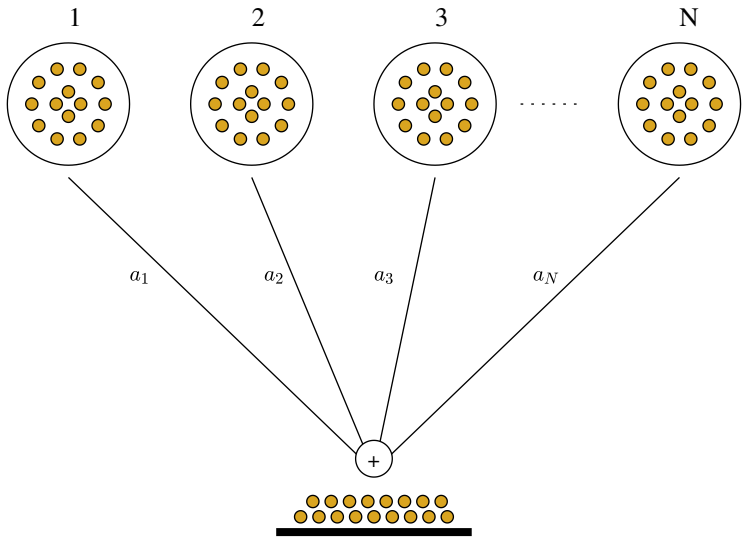
$$y(1) = \sum_{k=0}^{N-1} X(k)a_k(1). \tag{13.3}$$

This measurement will produce a new set of possible solutions for each $X(k)$ defined by

$$X(k) = y(1)/a_k(1), \quad a_k(1) \neq 0, k = 0,1,2,\ldots,N-1.$$

If these two hyperplanes (sets of possible solutions) produce only one common value for $k = i$,

$$X(i) = y(0)/a_i(0) = y(1)/a_i(1),$$



One bag with false coins
$y(0) = M - M_T = a_1 m + ... + a_i(m + X(i)) + ... + a_N m - (a_1 + ... + a_i + ... + a_N)m = a_i X(i)$
$i = ?, X(i) = ?$

Two bags with false coins
$y(0) = a_1 m + ... + a_i(m + X(i)) + ... + a_k(m + X(k)) + ... + a_N m - M_T = a_i X(i) + a_k X(k)$
$i = ?, k = ?, X(i) = ?, X(k) = ?$

Figure 13.1 There are N bags with coins. One of them, at an unknown position, contains false coins. False coins differ from the true ones in mass for an unknown $X(i) = \Delta m$. The mass of the true coins is m. A set of coins for the measurement is formed using: a_1 coins from the first bag, a_2 coins from the second bag, and so on. The total measured value is $M = a_1 m + \cdots + a_i(m + X(i)) + \cdots + a_N m$. The difference of this value from the total mass if all coins were true is $M - M_T$. The equations for the case with one and two bags with false coins are shown. The notation $a_k(0) = a_{k+1}$, for $k = 0, 1, \ldots, N-1$, is used in this illustration.

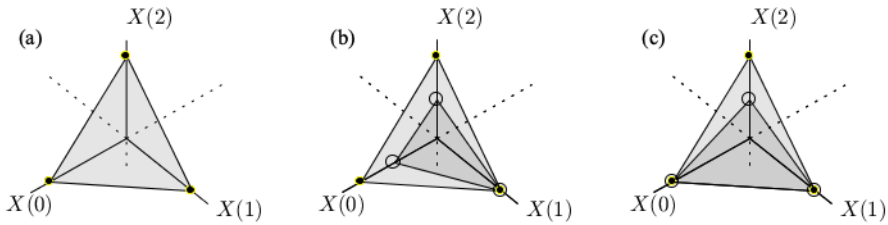


Figure 13.2 The solution illustration for $N = 3$, $K = 1$, and various possible cases: (a) Three possible solutions for one measurement plane. (b) Unique solution for two measurement planes. (c) Two possible solutions for two measurement planes.

then it is the solution to our problem, Fig.13.2(b).

Example 13.1. Consider a set of $N = 5$ bags of coins. In one of them all coins are false. The weight of true coins is $m = 2$.

In the first measurement we use $a_k(0) = k$ coins from the kth bag. The total weight of coins in this measurement is $M = 31$. This weight is equal to $(1+2+3+4+5)2 + iX(i) = M$, where $X(i)$ is the unknown weight difference of false coins. It means that $iX(i) = 1$, since all true coins would produce $M = (1+2+3+4+5)2 = 30$. If the false coins were in the first bag, their weight difference would be $X(1) = 1/1 = 1$, if they were in the second bag 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 th3 five bags.

Let us perform one more measurement with $a_k(1) = k^2$ coins from each bag. Assume that we get the total measured weight $M = 113$. It is equal to $M = 2(1^2+2^2+3^2+4^2+5^2) + i^2X(i) = 113$. Obviously $i^2X(i) = 3$. Again, if the false coins were in the first bag then $X(1) = 3/1$, the second false bag would produce $X(2) = 3/2^2 = 3/4$, and so on, $X(3) = 3/3^2 = 1/3$, $X(4) = 3/16$, $X(5) = 3/25$.

The common solution for both sets is $X(3) = 1/3$. Thus, the false coins are in the third bag. Their weight difference from the true coins is $1/3$.

Measurements (13.2) and (13.3) can be written in a matrix form as

$$\begin{bmatrix} y(0) \\ y(1) \end{bmatrix} = \begin{bmatrix} a_0(0) & a_1(0) & \dots & a_{N-1}(0) \\ a_0(1) & a_1(1) & \dots & a_{N-1}(1) \end{bmatrix} \begin{bmatrix} X(0) \\ X(1) \\ \dots \\ X(N-1) \end{bmatrix}$$

$$\mathbf{y} = \mathbf{A}\mathbf{X},$$

where $\mathbf{A}$ is the measurement matrix

$$\mathbf{A} = \begin{bmatrix} a_0(0) & a_1(0) & \dots & a_{N-1}(0) \\ a_0(1) & a_1(1) & \dots & a_{N-1}(1) \end{bmatrix} \tag{13.4}$$

and $\mathbf{y}$ are the measurements of the vector variable $\mathbf{X}$.

Common value for two measurements $X(i) = y(0)/a_i(0)$ and $X(i) = y(1)/a_i(1)$ is *unique* if

$$a_i(0)a_k(1) - a_i(1)a_k(0) = \det \begin{bmatrix} a_i(0) & a_k(0) \\ a_i(1) & a_k(1) \end{bmatrix} \neq 0$$

for all $i \neq k$. It also means that $\text{rank}\,(\mathbf{A}_2) = 2$ for all 2×2 submatrices, denoted by $\mathbf{A}_2$, of the measurement matrix $\mathbf{A}$ defined by (13.4).

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 (proof by contradiction)

$$a_i(0)X(i) = y(0), \quad a_i(1)X(i) = y(1)$$

and

$$a_k(0)X(k) = y(0), \quad a_k(1)X(k) = y(1).$$

Then

$$a_i(0)X(i) = a_k(0)X(k)$$
$$\text{and}$$
$$a_i(1)X(i) = a_k(1)X(k).$$

If we divide these two equations we get

$$\frac{a_i(0)}{a_i(1)} = \frac{a_k(0)}{a_k(1)}$$

or $a_i(0)a_k(1) - a_i(1)a_k(0) = 0$. This is contrary to the assumption that $a_i(0)a_k(1) - a_i(1)a_k(0) \neq 0$.

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

$$\begin{bmatrix} y(0) \\ y(1) \end{bmatrix} = \begin{bmatrix} a_i(0) & a_k(0) \\ a_i(1) & a_k(1) \end{bmatrix} \begin{bmatrix} X(i) \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} y(0) \\ y(1) \end{bmatrix} = \begin{bmatrix} a_i(0) & a_k(0) \\ a_i(1) & a_k(1) \end{bmatrix} \begin{bmatrix} 0 \\ X(k) \end{bmatrix}. \tag{13.5}$$

Subtraction of the previous matrix equations results in

$$\begin{bmatrix} a_i(0) & a_k(0) \\ a_i(1) & a_k(1) \end{bmatrix} \begin{bmatrix} X(i) \\ -X(k) \end{bmatrix} = 0.$$

If $a_i(0)a_k(1) - a_i(1)a_k(0) \neq 0$ is satisfied, then the trivial solution to the problem, $X(i) = X(k) = 0$, follows. Therefore, two different nonzero solutions $X(i)$ and $X(k)$ cannot exist in this case.

The previous experiment can be repeated assuming two nonzero values $X(i)$ and $X(k)$, Fig.13.1(second option). In the case of two nonzero elements in vector $\mathbf{X}$, two measurements,

$$y(0) = \sum_{l=0}^{N-1} X(l)a_l(0) = X(i)a_i(0) + X(k)a_k(0) \tag{13.6}$$
$$y(1) = \sum_{l=0}^{N-1} X(l)a_l(1) = X(i)a_i(1) + X(k)a_k(1),$$

will result in $X(i)$ and $X(k)$ for any i and k, $i \neq k$. They are the solution to a system with two equations of two unknowns. Therefore, with two measurements we cannot get a result of the problem and find the positions and the values of two nonzero coefficients. If two more measurements are performed then an additional system with two equations

$$y(2) = X(i)a_i(2) + X(k)a_k(2) \tag{13.7}$$
$$y(3) = X(i)a_i(3) + X(k)a_k(3)$$

is formed. Two systems of two equations (13.6) and (13.7) could be solved to find $X(i)$ 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 unique solution to our problem. As in the case of one nonzero coefficient, we may show that the sufficient condition for the unique solution is

$$\det \begin{bmatrix} a_{k_1}(0) & a_{k_2}(0) & a_{k_3}(0) & a_{k_4}(0) \\ a_{k_1}(1) & a_{k_2}(1) & a_{k_3}(1) & a_{k_4}(1) \\ a_{k_1}(2) & a_{k_2}(2) & a_{k_3}(2) & a_{k_4}(2) \\ a_{k_1}(3) & a_{k_2}(3) & a_{k_3}(3) & a_{k_4}(3) \end{bmatrix} \neq 0 \tag{13.8}$$

for all combinations of k_1, k_2, k_3 and k_4 or $\text{rank}\,(\mathbf{A}_4) = 4$ for all $\mathbf{A}_4$, where $\mathbf{A}_4$ is a 4×4 submatrix of the measurement matrix $\mathbf{A}$ defined, in this case, as

$$\mathbf{A} = \begin{bmatrix} a_0(0) & a_1(0) & \dots & a_{N-1}(0) \\ a_0(1) & a_1(1) & \dots & a_{N-1}(1) \\ a_0(2) & a_1(2) & \dots & a_{N-1}(2) \\ a_0(3) & a_1(3) & \dots & a_{N-1}(3) \end{bmatrix}. \tag{13.9}$$

Suppose that (13.8) 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,

$$\begin{bmatrix} y(0) \\ y(1) \\ y(2) \\ y(3) \end{bmatrix} = \begin{bmatrix} a_{k_1}(0) & a_{k_2}(0) & a_{k_3}(0) & a_{k_4}(0) \\ a_{k_1}(1) & a_{k_2}(1) & a_{k_3}(1) & a_{k_4}(1) \\ a_{k_1}(2) & a_{k_2}(2) & a_{k_3}(2) & a_{k_4}(2) \\ a_{k_1}(3) & a_{k_2}(3) & a_{k_3}(3) & a_{k_4}(3) \end{bmatrix} \begin{bmatrix} X(k_1) \\ X(k_2) \\ 0 \\ 0 \end{bmatrix}$$

and

$$\begin{bmatrix} y(0) \\ y(1) \\ y(2) \\ y(3) \end{bmatrix} = \begin{bmatrix} a_{k_1}(0) & a_{k_2}(0) & a_{k_3}(0) & a_{k_4}(0) \\ a_{k_1}(1) & a_{k_2}(1) & a_{k_3}(1) & a_{k_4}(1) \\ a_{k_1}(2) & a_{k_2}(2) & a_{k_3}(2) & a_{k_4}(2) \\ a_{k_1}(3) & a_{k_2}(3) & a_{k_3}(3) & a_{k_4}(3) \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ X(k_3) \\ X(k_4) \end{bmatrix}.$$

Subtracting these two systems we get

$$0 = \begin{bmatrix} a_{k_1}(0) & a_{k_2}(0) & a_{k_3}(0) & a_{k_4}(0) \\ a_{k_1}(1) & a_{k_2}(1) & a_{k_3}(1) & a_{k_4}(1) \\ a_{k_1}(2) & a_{k_2}(2) & a_{k_3}(2) & a_{k_4}(2) \\ a_{k_1}(3) & a_{k_2}(3) & a_{k_3}(3) & a_{k_4}(3) \end{bmatrix} \begin{bmatrix} X(k_1) \\ X(k_2) \\ -X(k_3) \\ -X(k_4) \end{bmatrix}.$$

Since (13.8) 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 if (13.8) holds.

The presented approach to solve the problem (and to check the solution uniqueness) is illustrative, however computationally not feasible. For example, in a simple case with $N = 1024$ and just two nonzero coefficients, we have to solve systems of equations (13.6) and (13.7) for each possible combination of i and k and to compare their solutions. The total number of combinations of two out of N indices is

$$\binom{N}{2} \sim 5 \times 10^5.$$

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, k_4 out of the set of N indices. The number of determinants is $\binom{N}{4} \sim 10^{10}$. If one determinant of the fourth 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 the second illustrative example consider a signal described by a weighted sum of K harmonics from a set of possible discrete oscillatory functions $e^{j2\pi kn/N}$, $k = 0, 1, 2, \dots, N-1$,

$$x(n) = A_1 e^{j2\pi k_1 n/N} + A_2 e^{j2\pi k_2 n/N} + \dots + A_K e^{j2\pi k_K n/N},$$

with $K \ll N$. This signal is sparse in the DFT domain. Its DFT $X(k)$ assumes only a few nonzero values at $k = k_i$, $i = 1, 2, \dots, K$.

In classical signal processing this signal is described by a full set of N signal samples/measurements $x(n)$ at $n = 0, 1, 2, \dots, N-1$.

However, if we know that the signal consists of only $K \ll N$ discrete oscillatory functions with unknown amplitudes and frequency indices k_i, then regardless of their frequencies, the signal can be

fully reconstructed from a reduced set of signal samples. As in the first illustrative example, a signal sample at an arbitrary instant n_1 can be considered as a weighted measurement of the sparse coefficients $X(k)$,

$$y(0) = x(n_1) = \sum_{k=0}^{N-1} X(k)\psi_k(n_1) = \sum_{k=0}^{N-1} X(k)a_k(0),$$

with the weighting factors $\psi_k(n_1) = \exp(j2\pi n_1 k/N)/N = a_k(0)$. The previous relation is the inverse DFT.

Now a similar analysis like in the first illustrative example can be performed, assuming, for example, $K = 1$ or $K = 2$. We can find the positions and the values of nonzero coefficients $X(k)$ using just a few signal samples/measurements $y(i)$. When the nonzero coefficient positions and their values are recovered then the whole DFT and the signal are recovered.

This model corresponds to many signals in real life. For example, in the Doppler-radar systems the speed of a radar target is transformed into a frequency of a sinusoidal signal. Since the returned signal contains only one or just a few targets, the signal representing target velocity is a sparse signal in the DFT domain. It can be reconstructed from fewer samples than the total number of radar return signal samples N.

13.2 BASIC DEFINITIONS

After the basic notions are introduced through illustrative examples in the previous section, here we will provide formal definitions of the key concepts in sparse signal processing and compressive sensing.

13.2.1 Sparsity

Consider a set of numbers $X(k)$, $k = 0,1,\ldots,N-1$. In signal processing this set of numbers corresponds to a signal in one of its representation domains.

A sequence $\{X(k)\}$, $k = 0,1,\ldots,N-1$ is referred to as *a sparse sequence* if the number, K, of its nonzero elements, $X(k) \neq 0$, is much smaller than its total length, N, that is,

$$X(k) = 0$$

for $k \notin \{k_1,k_2,\ldots,k_K\} = \mathbb{K}$, where the sparsity support set $\mathbb{K}$ is a subset of all possible indices,

$$\mathbb{K} = \{k_1,k_2,\ldots,k_K\} \subset \{0,1,\ldots,N-1\}.$$

The number of nonzero elements can be written as

$$\|\mathbf{X}\|_0 = \text{card}\{\mathbb{K}\} = K,$$

where $\text{card}\{\mathbb{K}\}$ is the notation for the number of elements in $\mathbb{K}$. Counting the nonzero elements in a signal representation $\mathbf{X}$ can be achieved using the so called ℓ_0-norm

$$\|\mathbf{X}\|_0 = \sum_{k=0}^{N-1} |X(k)|^0.$$

This function is referred to as the ℓ_0-norm (norm-zero) although it does not satisfy the norm properties ($\|c\mathbf{X}\|_0 = \|\mathbf{X}\|_0 \neq c\,\|\mathbf{X}\|_0$, for an arbitrary constant c). By definition $|X(k)|^0 = 0$ for $|X(k)| = 0$ and $|X(k)|^0 = 1$ for $|X(k)| \neq 0$.

The set of numbers $X(k)$, $k = 0,1,\ldots,N-1$, is sparse if

$$\text{card}\{\mathbb{K}\} = K \ll N.$$

A signal $x(n)$, $n = 0,1,\ldots,N$, is sparse in a representation domain with elements $X(k)$, $k = 0,1,\ldots,N$, if the set of these elements is sparse.

Example 13.2. Consider two sets of sparse numbers $X(k)$ and $H(k)$, $k = 0,1,\ldots,N-1$, in vector notations $\mathbf{X}$ and $\mathbf{H}$. Show that the sparsity of the sum of these numbers is not greater than the sum of their individual sparsities,

$$\|\mathbf{H}+\mathbf{X}\|_0 \leq \|\mathbf{H}\|_0 + \|\mathbf{X}\|_0, \tag{13.10}$$

meaning that the ℓ_0-norm satisfies the triangular property.

★ Assume that the sparsity support of $\mathbf{X}$ is $\mathbb{K}_X$ and the sparsity support of $\mathbf{H}$ is $\mathbb{K}_H$. We can differ the following cases:

- If $\mathbb{K}_X \cap \mathbb{K}_H = \emptyset$, then the number of nonzero numbers in $X(k)+H(k)$ is equal to the sum of the numbers of nonzero elements in $X(k)$ and $H(k)$, and $\|\mathbf{H}+\mathbf{X}\|_0 = \|\mathbf{H}\|_0 + \|\mathbf{X}\|_0$.

- If $\mathbb{K}_X \cap \mathbb{K}_H \neq \emptyset$, then the number of nonzero numbers in $X(k)+H(k)$ is always smaller than the sum of the numbers of nonzero elements in $X(k)$ and $H(k)$, for the number of overlapping indices and possible number of the elements that cancel out. Then $\|\mathbf{H}+\mathbf{X}\|_0 < \|\mathbf{H}\|_0 + \|\mathbf{X}\|_0$. Inequality (13.10) follows from these two cases. □

13.2.2 Measurements

A linear combination of the elements $X(k)$, $k = 0,1,\ldots,N-1$, of a vector $\mathbf{X}$, given by

$$y(n) = \sum_{k=0}^{N-1} a_k(n)X(k) = a_0(n)X(0) + \cdots + a_{N-1}(n)X(N-1), \tag{13.11}$$

is called *a measurement*, with the weighting coefficients (weights) denoted by $a_k(n)$. The measurements can written in a form of the system of M equations

$$\begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(M-1) \end{bmatrix} = \begin{bmatrix} a_0(0) & a_1(0) & a_{N-1}(0) \\ a_0(1) & a_1(1) & a_{N-1}(1) \\ \vdots & \vdots & \vdots \\ a_0(M-1) & a_1(M-1) & a_{N-1}(M-1) \end{bmatrix} \begin{bmatrix} X(0) \\ X(1) \\ \vdots \\ X(N-1) \end{bmatrix} \tag{13.12}$$

or in the matrix form as

$$\mathbf{y} = \mathbf{A}\mathbf{X},$$

where $\mathbf{A}$ is an $M \times N$ measurement matrix. An illustration of the measurements calculation is given in Fig. 13.3.

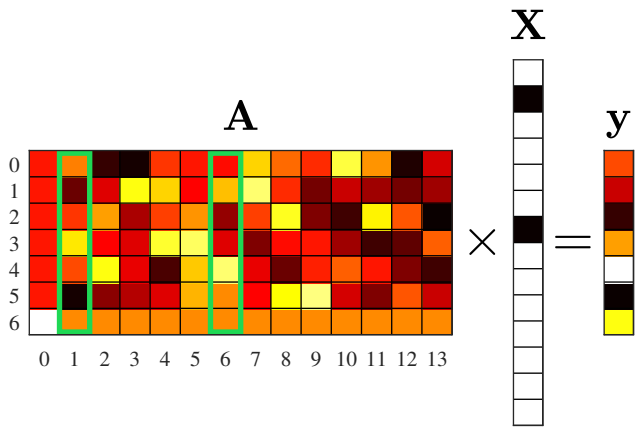


Figure 13.3 Principle of compressive sensing. The short and wide measurement matrix $\mathbf{A}$ maps the original N-dimensional K-sparse vector, $\mathbf{X}$, to an M-dimensional dense vector of measurements, $\mathbf{y}$, with $M < N$ and $K \ll N$. In our case $N = 14$, $M = 7$, and $K = 2$.

13.2.2.1 Sparsity Aware Form of the Measurements

The fact that the signal is sparse with $X(k) = 0$ for $k \notin \{k_1, k_2, \ldots, k_K\} = \mathbb{K}$ is not included in the measurement matrix $\mathbf{A}$ since the positions of the nonzero values are unknown. If the knowledge that $X(k) = 0$ for $k \notin \{k_1, k_2, \ldots, k_K\} = \mathbb{K}$ were included, then a reduced measurement matrix would be obtained as

$$\begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(M-1) \end{bmatrix} = \begin{bmatrix} a_{k_1}(0) & a_{k_2}(0) & a_{k_K}(0) \\ a_{k_1}(1) & a_{k_2}(1) & a_{k_K}(1) \\ \vdots & \vdots & \vdots \\ a_{k_1}(M-1) & a_{k_2}(M-1) & a_{k_K}(M-1) \end{bmatrix} \begin{bmatrix} X(k_1) \\ X(k_2) \\ \vdots \\ X(k_K) \end{bmatrix} \tag{13.13}$$

or

$$\mathbf{y} = \mathbf{A}_K \mathbf{X}_K.$$

The $M \times K$ matrix $\mathbf{A}_K$ would be formed if we knew the positions of nonzero samples $k \in \{k_1, k_2, \ldots, k_K\} = \mathbb{K}$. It would follow from the measurement matrix $\mathbf{A}$ by omitting the columns corresponding to the zero-valued elements $X(k)$. Vector $\mathbf{X}_K$ consists of the assumed nonzero elements $X(k)$.

Assuming that there are K nonzero elements $X(k)$, the total number of possible different matrices $\mathbf{A}_K$ is equal to the number of combinations with K out of N positions. It is equal to $\binom{N}{K}$. This matrix will play an important role in the analysis that follows.

13.2.2.2 Signal Samples as Measurements

In signal processing, the sparsity domain is commonly one of the signal transformation domains. For a linear signal transform $\mathbf{X} = \mathbf{\Phi x}$ and its inverse transform $\mathbf{x} = \mathbf{\Psi X}$ the signal samples are

$$x(n) = \sum_{k=0}^{N-1} X(k)\psi_k(n),$$

$\sigma = 0.1$ are assumed. The assumed threshold for considering hyperparameters extremely large is $T_h = 100$. Hyperparameters above this threshold are omitted from calculation (along with the corresponding values in $\mathbf{X}$, $\mathbf{A}$, $\mathbf{D}$ and $\mathbf{V}$). The results for estimated mean value $\mathbf{V}$ in the first iteration are shown in Fig.14.38(c), along with the values of hyperparameters $\mathbf{V}$ in Fig.14.38(d). The hyperparameters whose value is above T_h are omitted (pruned) along with the corresponding values at the same positions in all other matrices. In the second iteration the values of remaining hyperparameters $\mathbf{V}$ are shown in Fig.14.38(e). After the elimination of hyperparameters above the threshold, the third iteration is calculated with the remaining positions of the hyperparameters. In this iteration all hyperparameters, except those whose values are close to one, are eliminated Fig.14.38(f). The remaining positions, after this iteration, correspond to the nonzero elements $X(k_i), i = 1,2,\ldots,K$ positions, with corresponding pruned matrices $\Sigma_K, \mathbf{A}_K, \mathbf{D}_K$. The values of $X(k_i)$ are estimated using V_i from

$$\mathbf{V}_K = \Sigma_K \mathbf{A}_K^T \mathbf{y}/\sigma^2 = (\mathbf{A_K}^T \mathbf{A_K} + \sigma^2 \mathbf{D_K})^{-1} \mathbf{A}_K^T \mathbf{y}$$

in the final iteration. If the measurements were noise-free this would be exact recovery. The values of estimated $X(k_i), i = 1,2,\ldots,K$ are shown in Fig.14.38(g). The diagonal values of Σ_K are the variances of $X(k_i)$.

Index

χ-squared distribution, 361

Activation function, 529
Adaptive systems, 446
Allpass system, 224
Ambiguity function, 611
Analog signals, 14
Analytic part, 35
Antenna array, 488
Anticausal systems, 174
Attenuation, 55
Auto-regressive (AR), 173
Autocorrelation function, 318
Autocovariance function, 318

Back-propagation in CNN, 538
Backward difference, 207
Bandpass filter, 223
Basis pursuit, 754
Bayes' theorem, 309
Bayesian inference, 330
Bayesian reconstruction, 806
Bias-to-variance trade-off, 372, 427
Bilinear transform, 211
Binomial random variable, 329
Blackman window, 562
Block LMS algorithm, 495
Bootstrap method, 354
Born-Jordan distribution, 632
Butterworth filter, 52, 219
  discrete-time, 216

Capon's method, 599
  local polynomial Fourier transform (LPFT), 604
  short-time Fourier transform (STFT), 601
Cascade realization, 249
Causal system, 22, 65
Causal systems, 169
Central limit theorem, 334, 336
Characteristic function, 321
Characteristic polynomial, 174
Choi-Willimas distribution, 637
Cohen class of distributions, 631
  discrete form, 635
  kernel decomposition, 636
Coherence, 688, 691
comb filter, 183
Complex sinusoidal signal, 18
Compressive sensing, 674
Confidence Intervals, 347
Constant overlap-add (COLA), 557
Constant overlap-add (WOLA), 557
Continuous signals, 17
Convolution
  circular, 108
  continuous, 21
  discrete-time, 63
  in frequency domain, 34, 71
Convolution filter, 525
Convolution kernel, 525
Convolutional neural network – CNN, 525
Cramer-Rao bound, 341
Critically dumped response, 48
Cross-validation method, 427
Cumulants, 418

Daniell periodogram, 377, 378
Deep neural network, 514
Delta method, 353
Derivative
  complex function, 30
Difference equation, 171, 174
Differential equation, 47
Differentiator, 71
Digital signals, 14
Direct realization I, 242
Direct realization II, 243
Dirichlet conditions, 23
Discrete Cosine transform (DCT), 132
Discrete Fourier transform (DFT), 102, 117
Discrete Hartley transform (DCT), 155
Discrete pseudo Wigner distribution, 618
Discrete Sine transform (DST), 135
Discrete system, 62
Discrete-time signals (discrete signals), 58
Displacement, 130
Downsampling, 647

Duality property, 32

Eigenvalues, 458
Eigenvectors, 458
Energy, 19, 61
Equiangular Tight Frame, 690
ergodic process, 321
Ergodicity, 320
Error function, 336
Error signal
  adaptive system, 451

Fast Fourier transform
  decimaton-in-frequency, 121
Fast Fourier transform, 119
  decimaton-in-time, 122
Final value theorem, 46
Finite impulse response (FIR), 173
  frequency domain design, 262
  realization, 257
First-order statistics, 287
Fixed point arithmetic, 404
Flattening, 531
Floating point arithmetic, 410
  IEEE standard, 412
  mu-law and A-law, 412
Forced response, 50
Forward propagation, 526
Fourier series, 24, 37, 117
Fourier transform, 31, 37, 117
  matrix, 106
  of discrete-time signals, 66, 117
  optimal window, 372
  properties, 33
  signum function, 33
  variance, 371
Fractional Fourier transform, 598
  relation to the LPFT, 599
  windowed, 599
Frequency estimation, 128

Gershgorin theorem, 706
Gershgorin theorem, 707
Goertzel algorithm, 176
Gradient, 756
Gradient descent with sparsification (GDS), 739
Gram matrix, 701
Gramm matrix, 698
Group delay, 41, 228, 257, 614

Haar transform, 143, 646
Hadamard transform, 142
Hamming window, 561
Hann(ing) window, 69, 559
Hard therholding, 717
Hard thresholding, 715
Hartley series, 26
He initial values, 538
Highpass filter, 222
Hilbert transform, 36
Histogram, 289
Homogeneous equation, 174
Hypothesis testing, 356

Image reconstruction, 799
Impulse invariance method, 200
Impulse signal
  continuous (delta function), 17
  discrete-time, 59
Indirect measurements, 778
Infinite impulse response (IIR), 173
Initial condition
  continuous, 44
Initial value theorem, 46
Instantaneous frequency, 38, 614
Interpolation, 112
Intersecton of confidence intervals, 373
Inverse system, 226
Isometry, 686
ISTA algorithm, 756
Iterative hard thresholding (IHT), 739

Kalman filter, 505
Kronecker delta function, 59
Kurtosis, 435

L-statistics, 296
Lagrangian, 758
Laplace transform, 42
LASSO minimization, 754
Leakage effect, 128
Least absolute shrinkage and selection operator LASSO, 305
Linear adaptive adder, 449
Linear phase systems, 257
Linear system, 20
LMS algorithm, 474
  antenna systems, 488
  block, 495
  complex, 498
  convergence, 475
  echo cacellation, 492
  identification, 476
  noise cancelation, 480
  prediction, 485
  sign, 494
  sinusoidal disturbance, 483
  variable step, 497
Local polynomial Fourier transform, 594
  moments, 596
  relation to fractional Fourier transform, 598
Log-likelihood function, 332

Lowpass filter, 216

Magnitude, 19
Marginal probability, 310
Marginal properties, 629
Matched filter, 392
Matched z-transform method, 205
Maximum a posterior, 334
Maximum likelihood estimation, 333
Measurement Matrix
  Bernoulli Random, 685
  Indirect, 684
  Structured Random, 685, 779
Measurement matrix, 677, 681
Median, 293, 381
Memoryless property, 366
Minimum phase system, 226
Moment generating function, 366
Moment genrating function, 418
Moments, 418
  LPFT, 597
Morlet wavelet, 640
Moving average (MA), 173
Moyal property, 616
MUSIC, 602

Narrowband signals
  spectral estimation, 388
Natural response, 51
Neural networks, 510
  activation function, 512
  acyclic, 514
  continuous output, 519
  cyclic, 514
  error backpropagation, 522
  layer, 514
  multilayer, 522
  network function, 512
  perceptron, 516
  supervised, 515
  unsupervised, 524
  voting mashines, 524
Neuron, 511
Noise, 327
  Bernoulli, 328
  binary, 328
  compex Gaussian, 359
  Gaussian, 335
  Laplacian, 361
  missing samples, 727
  Poisson, 363
  reconstruction, 732
  shot, 363
  unform, 327
Noisy signal
  Fourier transform, 367

Norm zero, 709
Norm-one, 739
  ball, 744
Norm-zero, 680
Notch filter, 171, 198

Opinion poll estimation, 350
Optimal filter, 397, 455
Orthogonality principle, 464
Overdumped response, 49
Overflow, 406

Parallel realization, 253
Parseval's theorem, 72, 108
Perceptron, 516
Period of a discrete signal, 60
Periodogram, 373
Pisarenko method, 602
Polynomial fitting, 306
Pooling, 531
Power, 19, 61
Power spectral density, 322, 385
Probability
  density function, 306, 312
Probability density function, 312
Pseudoinverse matrix, 712

Quantization, 400

Random signals, 287
Rank of matrix, 701
Rayleigh distribution
  unform, 359
Reconstruction uniqueness, 771
Reconstruction uniquness, 775
Rectangular window, 68
Recursive systems
  adaptive, 503
Reduced interference distributions
  discrete form, 632
Region of convergence, 159
ReLU activation function, 530
Resolution, 560
Restricted isometry, 686, 691
  constant, 687, 693
  eigenvalues, 697
  uniqueness, 702
Ridge regression, 304, 754
RLS algorithm
  variable step, 500

S-method, 377, 624
S-transform (the Stockwell transform), 642
Sampling
  nonunform, 779
Sampling theorem, 99

for periodic signals, 123
in the frequency domain, 37
in the time domain, 74
Schwartz's inequality, 690
Second-order statistics, 318
Sensitivity of system, 246
Shallow neural network, 514
Short-time Fourier transform (STFT), 550
discrete, 565
discrete-time, 564
filter bank, 569
frequency-varying, 592
hybrid, 594
inversion, 557, 572
optimal window, 553
optimisation, 584
overlapping, 570
recursive, 568
time-varying, 581
Sign LMS algorithm, 494
Sinc distribution
discrete form, 632
Smoothed periodogram, 377
Soft-thresholding, 756
Spacification, 717
Spark of matrix, 701
coherence, 753
uniqueness, 705
Sparse signals, 674
Sparsity, 680
Sparsity support, 680
Stable system, 22, 65
Stable systems, 169
Standard deviation, 297
starred transform, 179
Stationary phase method, 37
Stationary signals, 320
Steepest descend method, 465
Stride, 530

Taylor series, 38
Trace of matrix, 467

Unbiased autocorrelation estimator, 377
Underdumped response, 49
Unit step signal
continuous (Heaviside function), 17
discrete-time, 59
Upsampling, 648

Variance, 297
Variance stabilization, 352
Voting mashines, 524

Walsh-Hadamard transform, 142
Wavelet transform, 639
Coifflet, 669
Daubechies D4, 655
Daubechies D6, 668
filter bank, 645
Haar, 653
orthogonality, 651
reconstruction condition, 649
scale function, 664
wavelet function, 664
Welch bound, 688
Welch periodogram, 378, 390
Wide sense stationary signals, 320
Wide-sense cyclostationary, 320
Wiener filter, 397
Wigner distribution, 378, 606, 631
auto-terms, 610
cross-terms, 610
discrete form, 619
properties, 613
pseudo, 616
S-method, 624
smoothed, 617
Window, 558
Bartlett (triangular), 260, 558
Blackman, 562
Hamming, 262, 561
Hann(ing), 262
Hann(ing) (Hann), 559
Kaiser, 562
rectangular, 558
Windows, 258

Xavier initial values, 538

Yule-Walk equation, 389

z-transform, 158, 179
inverse, 163
Zero-padding, 112
Zhao-Atlas-Marks distribution
discrete form, 632

Bibliography

[1] S. T. Alexander, *Adaptive Signal Processing: Theory and Applications,* Springer-Verlag, 1986.

[2] R. Allred, *Digital Filters for Everyone,* CreateSpace Independent Publishing Platform; 2 edition, 2013.

[3] M. Amin, *Compressive Sensing for Urban Radar*, CRC Press, 2014.

[4] A. Antoniou, *Digital Signal Processing: Signals, Systems, and Filters,* McGraw-Hill Education, 2005.

[5] F. Auger and F. Hlawatsch, *Time-Frequency Analysis: Concepts and Methods*, Wiley-ISTE, 2008.

[6] D. Blandford and J. Parr, *Introduction to Digital Signal Processing,* Prentice Hall, 2012.

[7] B. Boashash, *Time-Frequency Signal Analysis and Processing: A Comprehensive Reference*, Second Edition, Academic Press, 2015.

[8] P. Bremaud, *Mathematical Principles of Signal Processing: Fourier and Wavelet Analysis*, Springer, 2002.

[9] S. A. Broughton, Discrete Fourier Analysis and Wavelets: Applications to Signal and Image Processing, Wiley-Interscience, 2008.

[10] L. F. Chaparro, *Signals and Systems using MATLAB*, Academic Press, 2011.

[11] C. T. Chen, *Signals and Systems: A Fresh Look*, CreateSpace Independent Publishing Platform, 2011.

[12] V. Chen, D. Tahmoush, and W. J. Miceli, *Radar Micro-Doppler Signature-Processing and Applications,* The Institution of Engineering and Technology (IET), 2014.

[13] L. Cohen, *Time-frequency Analysis*, *Prentice-Hall*, 1995.

[14] A. G. Constantinides, *System Function of Discrete-Time Systems*, Academic Press, 2001.

[15] M. Davenport, *Digital Signal Processing,* Kindle edition, 2011.

[16] P. S. R. Diniz, E. A. B. da Silva, and S. L. Netto, *Digital Signal Processing: System Analysis and Design* , Cambridge University Press; 2 edition, 2010.

[17] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, 1992.

[18] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing,* Springer, 2010.

[19] Y. C. Eldar and G. Kutyniok *Compressed Sensing: Theory and Applications*, Cambridge University Press, 2012.

[20] M. E. Frerking, *Digital Signal Processing in Communication Systems,* Springer, 1994.

[21] P. Flandrin, *Time-Frequency/Time-Scale Analysis*, Academic Press, 1999.

[22] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Prentice Hall, 2007.

[23] N. Hamdy, *Applied Signal Processing: Concepts, Circuits, and Systems*, CRC Press, 2008.

[24] M. H. Hayes, *Schaums Outline of Digital Signal Processing*, 2nd Edition, Schaum's Outline Series, 2011.

[25] S. Haykin, *Least-Mean-Square Adaptive Filters,* Wiley-Interscience, 2003.

[26] S. Haykin and B. Van Veen, *Signals and Systems*, Wiley, 2002.

[27] E. Ifeachor and B. Jervis, *Digital Signal Processing: A Practical Approach*, Prentice Hall; 2 edition, 2001.

[28] V. K. Ingle and J. G. Proakis, *Digital Signal Processing Using MATLAB*, CL Engineering; 3 edition, 2011.

[29] S. Kay, *Fundamentals of Statistical Signal Processing*, *Volume I: Estimation Theory*, Prentice Hall, 1993.

[30] S. Kay, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*, Prentice Hall, 1998.

[31] E. Kudeki and D. C. Munson Jr., *Analog Signals and Systems,* Prentice Hall, 2008.

[32] S. M. Kuo, B. H. Lee, and W. Tian, *Real-Time Digital Signal Processing: Fundamentals, Implementations and Applications*, Wiley, third edition, 2013.

[33] B. P. Lathi and Z. Ding, *Modern Digital and Analog Communication Systems* , Oxford University Press, 2009.

[34] B. P. Lathi and R. A. Green, *Essentials of Digital Signal Processing,* Cambridge University Press, 2014.

[35] S. R. Laxpati and V. Goncharoff, *Practical Signal Processing and its Applications, Part I,* CreateSpace Independent Publishing Platform, 2013.

[36] R. G. Lyons, *Understanding Digital Signal Processing,* Pearson, 2010.

[37] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, second edition, 1999.

[38] D. P. Mandić and S. L. Goh, *Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear and Neural Models*, John Wiley & Sons, 2009.

[39] D. P. Mandić, M. Golz, A. Kuh, D. Obradovic, and T. Tanaka, *Signal Processing Techniques for Knowledge Extraction and Information Fusion*, Springer, 2008.

[40] D. G. Manolakis and V. K. Ingle, *Applied Digital Signal Processing: Theory and Practice*, Cambridge University Press, 2011.

[41] D. G. Manolakis, D. Manolakis, V. K. Ingle, S. M. and Kogon, *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering and Array Processing*, Artech House, 2005.

[42] J. H. McClellan, R. W. Schafer, and M. A. Yoder, *Signal Processing First*, Prentice Hall, 2003.

[43] S. K. Mitra, *Digital Signal Processing*, McGraw-Hill, 2010.

[44] B. Mulgrew, P. Grant, and J. Thompson, *Digital Signal Processing: Concepts and Applications,* Palgrave Macmillan; 2nd Edition edition, 2002.

[45] R. Newbold, *Practical Applications in Digital Signal Processing,* Prentice Hall, 2012.

[46] A. V. Oppenheim and A. S. Willsky, *Signals and Systems*, Prentice Hall, 2008.

[47] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing,* Prentice Hall, 2009.

[48] A. V. Oppenheim and G. C. Verghese, *Signals, Systems and Inference,* Prentice Hall, 2015.

[49] A. Papulis, *Signal Analysis*, McGraw-Hill, 1997

[50] C. L. Phillips, J. Parr, and E. Riskin, *Signals, Systems, and Transforms,* Prentice Hall, 2013.

[51] B. Porat, *A Course in Digital Signal Processing*, Wiley, 1996.

[52] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications,* Pearson; 4 edition, 2013.

[53] A. Quinquis, *Digital Signal Processing Using Matlab,* Wiley-ISTE, 2008.

[54] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing,* Prentice Hall, 1975.

[55] M. A. Richards, *Fundamentals of Radar Signal Processing,* McGraw-Hill Education, second edition, 2014.

[56] M. J. Roberts, *Signals and Systems: Analysis Using Transform Methods & MATLAB,* McGraw-Hill Education, 2011.

[57] A. H. Sayed, *Adaptive Filters*, Wiley-IEEE, 2008.

[58] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*, Prentice Hall, 1991.

[59] M. Soumekh, *Synthetic Aperture Radar Signal Processing with MATLAB Algorithms*, Wiley-Interscience, 1999.

[60] L. Stanković, *Digital Signal Processing*, Naucna knjiga, Beograd, Second edition 1989 (in Montenegrin/Serbian/Croat/Bosnian).

[61] L. Stanković, M. Daković, and T. Thayaparan, *Time-frequency Signal Analysis with Applications*, Artech House, Boston, 2013.

[62] L. Stanković and I. Djurović, *Solved Problems in Digital Signal Processing*, University of Montenegro, 1998 (in Montenegrin/Serbian/Croat/Bosnian).

[63] S. Stanković, I. Orović, E. Sejdić, *Multimedia Signals and Systems,* Springer, second edition, 2015.

[64] H. Stark and J. W. Woods, *Probability and Random Processes with Applications to Signal Processing*, Prentice Hall, third edition, 2001.

[65] L. Tan and J. Jiang, *Digital Signal Processing: Fundamentals and Applications,* Academic Press; 2 edition , 2013.

[66] S. Theodoridis and R. Chellappa, eds., *Academic Press Library in Signal Processing*, Vols.1-4, Academic Press, 2013.

[67] M. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine", *Journal of Machine Learning Research*, JMLR.org, 1, 2001, pp.211- 244.

[68] A. Uncini, *Fundamentals of Adaptive Signal Processing (Signals and Communication Technology)*, Springer, 2014.

[69] M.Vetterli and J.Kovačević, *Wavelets and Subband Coding*, CreateSpace Independent Publishing Platform, 2013.

[70] M. Vetterli, J. Kovačević, and V. K. Goyal, *Foundations of Signal Processing,* Cambridge University Press, 2014.

[71] B. Widrow and D. Stearns, *Adaptive Signal Processing*, Prentice Hall, 1985.

About the Author

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 degree 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, MASS. 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 Universiteit Eindhoven.

From 2003 to 2008, Stanković was the rector of the University of Montenegro. He was the ambassador of Montenegro to the United Kingdom, Iceland, and Ireland 2011-2015. During his stay in the United Kingdom he was a visiting academic at Imperial College London, 2013-2014.

His current interests are in signal processing. He published about 500 technical papers, more than 170 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, an associate editor of the IEEE Transactions on Signal Processing, and an associate editor of the IET Signal Processing.

Stanković is a member of the Editorial Board of Signal Processing (Elsevier), associate editor of the CN Computer Sciences (Springer Nature), deputy editor of the IET Signal Processing, and a senior area editor of the IEEE Transactions on Image Processing.

He is a member of the National Academy of Sciences and Arts of Montenegro (CANU) since 1996, vice-president of CANU since 2015, a member of the Academia Europaea, 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.

Stanković (with coauthors) won the Best paper award from the European Association for Signal Processing (EURASIP) for 2017 for a paper published in Signal Processing.

For bibliographic data and copies of the published papers see **www.tfsa.ac.me**

Library of Congress Cataloging-in-Publication Data
Library of Congress Control Number: 2015912465

ISBN-13: 978-1514179987
ISBN-10: 1514179989

© 2015 Ljubiša Stanković, All Rights Reserved

Printed by CreateSpace Independent Publishing Platform,
An Amazon.com Company
North Charleston, South Carolina, USA.

Available from Amazon.com and other on-line and bookstores

All right reserved. Printed and bounded in the United States of America.

© 2020 Ljubiša Stanković, All Rights Reserved

No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the copyright holder.