Contents

Preface *XIX*

Acknowledgments *XXI*

Glossary of Symbols *XXIII*

Acronyms and Abbreviations *XXVII*

Part I Background

1 Introduction **3**
- 1.1 Definition of the Quantizer 3
- 1.2 Sampling and Quantization (Analog-to-Digital Conversion) 9
- 1.3 Exercises 10

2 Sampling Theory **13**
- 2.1 Linvill's Frequency Domain Description of Sampling 14
- 2.2 The Sampling Theorem; Recovery of the Time Function from its Samples 18
- 2.3 Anti-Alias Filtering 22
- 2.4 A Statistical Description of Quantization, Based on Sampling Theory 25
- 2.5 Exercises 28

3 Probability Density Functions, Characteristic Functions, Moments **31**
- 3.1 Probability Density Function 31
- 3.2 Characteristic Function and Moments 33
- 3.3 Joint Probability Density Functions 35
- 3.4 Joint Characteristic Functions, Moments, and Correlation Functions 40
- 3.5 First-Order Statistical Description of the Effects of Memoryless Operations on Signals 43

3.6 Addition of Random Variables and Other Functions of Random Variables 46
3.7 The Binomial Probability Density Function 47
3.8 The Central Limit Theorem 49
3.9 Exercises 53

Part II Uniform Quantization

4 Statistical Analysis of the Quantizer Output 61
4.1 PDF and CF of the Quantizer Output 61
4.2 Comparison of Quantization with the Addition of Independent Uniformly Distributed Noise, the PQN Model 66
4.3 Quantizing Theorems I and II 69
4.4 Recovery of the PDF of the Input Variable x from the PDF of the Output Variable x' 70
4.5 Recovery of Moments of the Input Variable x from Moments of the Output Variable x' when QT II is Satisfied; Sheppard's Corrections and the PQN Model 80
4.6 General Expressions of the Moments of the Quantizer Output, and of the Errors of Sheppard's Corrections: Deviations from the PQN Model 84
4.7 Sheppard's Corrections with a Gaussian Input 84
4.8 Summary 85
4.9 Exercises 87

5 Statistical Analysis of the Quantization Noise 93
5.1 Analysis of the Quantization Noise and the PQN Model 93
5.2 Satisfaction of Quantizing Theorems I and II 99
5.3 Quantizing Theorem III/A 99
5.4 General Expressions of the First- and Higher-Order Moments of the Quantization Noise: Deviations from the PQN Model 102
5.5 Quantization Noise with Gaussian Inputs 106
5.6 Summary 107
5.7 Exercises 108

6 Crosscorrelations between Quantization Noise, Quantizer Input, and Quantizer Output 113
6.1 Crosscorrelations when Quantizing Theorem II is Satisfied 113
6.1.1 Crosscorrelation between Quantization Noise and the Quantizer Input 113
6.1.2 Crosscorrelation between Quantization Noise and the Quantizer Output 115

6.1.3 Crosscorrelation between the Quantizer Input and the Quantizer Output 116
6.2 General Expressions of Crosscorrelations 116
6.2.1 Crosscorrelation between Quantization Noise and the Quantizer Input 116
6.2.2 Crosscorrelation between Quantization Noise and the Quantizer Output Signal 119
6.2.3 Crosscorrelation between the Quantizer Input and Output Signals 122
6.3 Correlation and Covariance between Gaussian Quantizer Input and Its Quantization Noise 123
6.4 Conditions of Orthogonality of Input x and Noise ν: Quantizing Theorem III/B 126
6.5 Conditions of Uncorrelatedness between x and ν: Quantizing Theorem IV/B 127
6.6 Summary 128
6.7 Exercises 129

7 General Statistical Relations among the Quantization Noise, the Quantizer Input, and the Quantizer Output 131
7.1 Joint PDF and CF of the Quantizer Input and Output 131
7.2 Quantizing Theorems for the Joint CF of the Quantizer Input and Output 138
7.3 Joint PDF and CF of the Quantizer Input and the Quantization Noise: Application of the PQN Model 140
7.4 Quantizing Theorems for the Joint CF of the Quantizer Input and the Quantization Noise: Application of the PQN Model 146
7.5 Joint Moments of the Quantizer Input and the Quantization Noise: Quantizing Theorem III 149
7.5.1 General Expressions of Joint Moments when Quantizing Theorem III is not satisfied 151
7.6 Joint Moments of the Centralized Quantizer Input and the Quantization Noise: Quantizing Theorem IV 152
7.6.1 General Expressions of Joint Moments 153
7.7 Joint PDF and CF of the Quantization Noise and the Quantizer Output 154
7.8 Three-Dimensional Probability Density Function and Characteristic Function 158
7.8.1 Three-Dimensional Probability Density Function 158
7.8.2 Three-Dimensional Characteristic Function 159
7.9 General Relationship between Quantization and the PQN Model 160
7.10 Overview of the Quantizing Theorems 162

7.11 Examples of Probability Density Functions Satisfying Quantizing Theorems III/B or QT IV/B 165
7.12 Summary 170
7.13 Exercises 171

8 Quantization of Two or More Variables: Statistical Analysis of the Quantizer Output 173
8.1 Two-Dimensional Sampling Theory 174
8.2 Statistical Analysis of the Quantizer Output for Two-Variable Quantization 179
8.3 A Comparison of Multivariable Quantization with the Addition of Independent Quantization Noise (PQN) 184
8.4 Quantizing Theorem I for Two and More Variables 186
8.5 Quantizing Theorem II for Two and More Variables 187
8.6 Recovery of the Joint PDF of the Inputs x_1, x_2 from the Joint PDF of the Outputs x_1', x_2' 187
8.7 Recovery of the Joint Moments of the Inputs x_1, x_2 from the Joint Moments of the Outputs x_1', x_2': Sheppard's Corrections 190
8.8 Summary 192
8.9 Exercises 193

9 Quantization of Two or More Variables: Statistical Analysis of Quantization Noise 197
9.1 Analysis of Quantization Noise, Validity of the PQN Model 197
9.2 Joint Moments of the Quantization Noise 200
9.3 Satisfaction of Quantizing Theorems I and II 203
9.4 Quantizing Theorem III/A for N Variables 204
9.5 Quantization Noise with Multiple Gaussian Inputs 206
9.6 Summary 207
9.7 Exercises 207

10 Quantization of Two or More Variables: General Statistical Relations between the Quantization Noises, and the Quantizer Inputs and Outputs 209
10.1 Joint PDF and CF of the Quantizer Inputs and Outputs 209
10.2 Joint PDF and CF of the Quantizer Inputs and the Quantization Noises 210
10.3 Joint PDF, CF, and Moments of the Quantizer Inputs and Noises when Quantizing Theorem I or II is Satisfied 211
10.4 General Expressions for the Covariances between Quantizer Inputs and Noises 213
10.5 Joint PDF, CF, and Moments of the Quantizer Inputs and Noises when Quantizing Theorem IV/B is Satisfied 214

10.6 Joint Moments of Quantizer Inputs and Noises with Quantizing Theorem III Satisfied 216
10.7 Joint Moments of the Quantizer Inputs and Noises with Quantizing Theorem IV Satisfied 217
10.8 Some Thoughts about the Quantizing Theorems 218
10.9 Joint PDF and CF of Quantization Noises and Quantizer Outputs under General Conditions 218
10.10 Joint PDF and CF of Quantizer Inputs, Quantization Noises, and Quantizer Outputs 219
10.11 Summary 221
10.12 Exercises 222

11 Calculation of the Moments and Correlation Functions of Quantized Gaussian Variables 225
11.1 The Moments of the Quantizer Output 225
11.2 Moments of the Quantization Noise, Validity of the PQN Model 233
11.3 Covariance of the Input x and Noise ν 237
11.4 Joint Moments of Centralized Input $\tilde{x}$ and Noise ν 240
11.5 Quantization of Two Gaussian Variables 242
11.6 Quantization of Samples of a Gaussian Time Series 249
11.7 Summary 252
11.8 Exercises 253

Part III Floating-Point Quantization

12 Basics of Floating-Point Quantization 257
12.1 The Floating-Point Quantizer 257
12.2 Floating-Point Quantization Noise 260
12.3 An Exact Model of the Floating-Point Quantizer 261
12.4 How Good is the PQN Model for the Hidden Quantizer? 266
12.5 Analysis of Floating-Point Quantization Noise 272
12.6 How Good is the PQN Model for the Exponent Quantizer? 280
12.6.1 Gaussian Input 280
12.6.2 Input with Triangular Distribution 285
12.6.3 Input with Uniform Distribution 286
12.6.4 Sinusoidal Input 290
12.7 A Floating-Point PQN Model 302
12.8 Summary 303
12.9 Exercises 304

13 More on Floating-Point Quantization 307
13.1 Small Deviations from the Floating-Point PQN Model 307

13.2 Quantization of Small Input Signals with High Bias 311
13.3 Floating-Point Quantization of Two or More Variables 313
13.3.1 Relationship between Correlation Coefficients ρ_{v_1,v_2} and $\rho_{v_{FL_1},v_{FL_2}}$ for Floating-Point Quantization 324
13.4 A Simplified Model of the Floating-Point Quantizer 325
13.5 A Comparison of Exact and Simplified Models of the Floating-Point Quantizer 331
13.6 Digital Communication with Signal Compression and Expansion: "μ-law" and "A-law" 332
13.7 Testing for PQN 333
13.8 Practical Number Systems: The IEEE Standard 343
13.8.1 Representation of Very Small Numbers 343
13.8.2 Binary Point 344
13.8.3 Underflow, Overflow, Dynamic Range, and SNR 345
13.8.4 The IEEE Standard 346
13.9 Summary 348
13.10 Exercises 351

14 Cascades of Fixed-Point and Floating-Point Quantizers 355
14.1 A Floating-Point Compact Disc 355
14.2 A Cascade of Fixed-Point and Floating-Point Quantizers 356
14.3 More on the Cascade of Fixed-Point and Floating-Point Quantizers 360
14.4 Connecting an Analog-to-Digital Converter to a Floating-Point Computer: Another Cascade of Fixed- and Floating-Point Quantization 367
14.5 Connecting the Output of a Floating-Point Computer to a Digital-to-Analog Converter: a Cascade of Floating-Point and Fixed-Point Quantization 368
14.6 Summary 369
14.7 Exercises 369

Part IV Quantization in Signal Processing, Feedback Control, and Computations

15 Roundoff Noise in FIR Digital Filters and in FFT Calculations 373
15.1 The FIR Digital Filter 373
15.2 Calculation of the Output Signal of an FIR Filter 374
15.3 PQN Analysis of Roundoff Noise at the Output of an FIR Filter 376
15.4 Roundoff Noise with Fixed-Point Quantization 377
15.5 Roundoff Noise with Floating-Point Quantization 381
15.6 Roundoff Noise in DFT and FFT Calculations 383
15.6.1 Multiplication of Complex Numbers 385

15.6.2 Number Representations in Digital Signal Processing Algorithms, and Roundoff 386
15.6.3 Growing of the Maximum Value in a Sequence Resulting from the DFT 387
15.7 A Fixed-Point FFT Error Analysis 388
15.7.1 Quantization Noise with Direct Calculation of the DFT 388
15.7.2 Sources of Quantization Noise in the FFT 389
15.7.3 FFT with Fixed-Point Number Representation 392
15.8 Some Noise Analysis Results for Block Floating-Point and Floating-Point FFT 394
15.8.1 FFT with Block Floating-Point Number Representation 394
15.8.2 FFT with Floating-Point Number Representation 394
15.9 Summary 397
15.10 Exercises 397

16 Roundoff Noise in IIR Digital Filters 403
16.1 A One-Pole Digital Filter 403
16.2 Quantization in a One-Pole Digital Filter 404
16.3 PQN Modeling and Moments with FIR and IIR Systems 406
16.4 Roundoff in a One-Pole Digital Filter with Fixed-Point Computation 407
16.5 Roundoff in a One-Pole Digital Filter with Floating-Point Computation 414
16.6 Simulation of Floating-point IIR Digital Filters 416
16.7 Strange Cases: Exceptions to PQN Behavior in Digital Filters with Floating-Point Computation 418
16.8 Testing the PQN Model for Quantization Within Feedback Loops 419
16.9 Summary 425
16.10 Exercises 427

17 Roundoff Noise in Digital Feedback Control Systems 431
17.1 The Analog-to-Digital Converter 432
17.2 The Digital-to-Analog Converter 432
17.3 A Control System Example 434
17.4 Signal Scaling Within the Feedback Loop 442
17.5 Mean Square of the Total Quantization Noise at the Plant Output 447
17.6 Satisfaction of QT II at the Quantizer Inputs 449
17.7 The Bertram Bound 455
17.8 Summary 460
17.9 Exercises 461

18 Roundoff Errors in Nonlinear Dynamic Systems – A Chaotic Example 465
18.1 Roundoff Noise 465

18.2 Experiments with a Linear System 467
18.3 Experiments with a Chaotic System 470
18.3.1 Study of the Logistic Map 470
18.3.2 Logistic Map with External Driving Function 478
18.4 Summary 481
18.5 Exercises 481

Part V Applications of Quantization Noise Theory

19 Dither 485
19.1 Dither: Anti-alias Filtering of the Quantizer Input CF 485
19.2 Moment Relations when QT II is Satisfied 488
19.3 Conditions for Statistical Independence of x and ν, and d and ν 489
19.4 Moment Relations and Quantization Noise PDF when QT III or QT IV is Satisfied 492
19.5 Statistical Analysis of the Total Quantization Error $\xi = d + \nu$ 493
19.6 Important Dither Types 497
19.6.1 Uniform Dither 497
19.6.2 Triangular Dither 500
19.6.3 Triangular plus Uniform Dither 501
19.6.4 Triangular plus Triangular Dither 502
19.6.5 Gaussian Dither 502
19.6.6 Sinusoidal Dither 503
19.6.7 The Use of Dither in the Arithmetic Processor 503
19.7 The Use of Dither for Quantization of Two or More Variables 504
19.8 Subtractive Dither 506
19.8.1 Analog-to-Digital Conversion with Subtractive Dither 508
19.9 Dither with Floating-Point 512
19.9.1 Dither with Floating-Point Analog-to-Digital Conversion 512
19.9.2 Floating-Point Quantization with Subtractive Dither 515
19.9.3 Dithered Roundoff with Floating-Point Computation 516
19.10 The Use of Dither in Nonlinear Control Systems 520
19.11 Summary 520
19.12 Exercises 522

20 Spectrum of Quantization Noise and Conditions of Whiteness 529
20.1 Quantization of Gaussian and Sine-Wave Signals 530
20.2 Calculation of Continuous-Time Correlation Functions and Spectra 532
20.2.1 General Considerations 532
20.2.2 Direct Numerical Evaluation of the Expectations 535
20.2.3 Approximation Methods 536

20.2.4 Correlation Function and Spectrum of Quantized Gaussian Signals 538
20.2.5 Spectrum of the Quantization Noise of a Quantized Sine Wave 544
20.3 Conditions of Whiteness for the Sampled Quantization Noise 548
20.3.1 Bandlimited Gaussian Noise 550
20.3.2 Sine Wave 554
20.3.3 A Uniform Condition for White Noise Spectrum 556
20.4 Summary 560
20.5 Exercises 562

Part VI Quantization of System Parameters

21 Coefficient Quantization 565
21.1 Coefficient Quantization in Linear Digital Filters 566
21.2 An Example of Coefficient Quantization 569
21.3 Floating-Point Coefficient Quantization 572
21.4 Analysis of Coefficient Quantization Effects by Computer Simulation 574
21.5 Coefficient Quantization in Nonlinear Systems 576
21.6 Summary 578
21.7 Exercises 579

APPENDICES

A Perfectly Bandlimited Characteristic Functions 589
A.1 Examples of Bandlimited Characteristic Functions 589
A.2 A Bandlimited Characteristic Function Cannot Be Analytic 594
A.2.1 Characteristic Functions that Satisfy QT I or QT II 595
A.2.2 Impossibility of Reconstruction of the Input PDF when QT II is Satisfied but QT I is not 595

B General Expressions of the Moments of the Quantizer Output, and of the Errors of Sheppard's Corrections 597
B.1 General Expressions of the Moments of the Quantizer Output 597
B.2 General Expressions of the Errors of Sheppard's Corrections 602
B.3 General Expressions for the Quantizer Output Joint Moments 607

C Derivatives of the Sinc Function 613

D Proofs of Quantizing Theorems III and IV **617**
D.1 Proof of QT III 617
D.2 Proof of QT IV 618

E Limits of Applicability of the Theory – Caveat Reader **621**
E.1 Long-time vs. Short-time Properties of Quantization 621
E.1.1 Mathematical Analysis 624
E.2 Saturation effects 626
E.3 Analog-to-Digital Conversion: Non-ideal Realization of Uniform Quantization 628

F Some Properties of the Gaussian PDF and CF **633**
F.1 Approximate Expressions for the Gaussian Characteristic Function 634
F.2 Derivatives of the CF with $\mathrm{E}\{x\} \neq 0$ 635
F.3 Two-Dimensional CF 636

G Quantization of a Sinusoidal Input **637**
G.1 Study of the Residual Error of Sheppard's First Correction 638
G.2 Approximate Upper Bounds for the Residual Errors of Higher Moments 640
G.2.1 Examples 642
G.3 Correlation between Quantizer Input and Quantization Noise 643
G.4 Time Series Analysis of a Sine Wave 645
G.5 Exact Finite-sum Expressions for Moments of the Quantization Noise 648
G.6 Joint PDF and CF of Two Quantized Samples of a Sine Wave 653
G.6.1 The Signal Model 653
G.6.2 Derivation of the Joint PDF 654
G.6.3 Derivation of the Joint CF 657
G.7 Some Properties of the Bessel Functions of the First Kind 660
G.7.1 Derivatives 660
G.7.2 Approximations and Limits 661

H Application of the Methods of Appendix G to Distributions other than Sinusoidal **663**

I A Few Properties of Selected Distributions **667**
I.1 Chi-Square Distribution 667
I.2 Exponential Distribution 670
I.3 Gamma Distribution 672
I.4 Laplacian Distribution 674
I.5 Rayleigh Distribution 676
I.6 Sinusoidal Distribution 677

I.7 Uniform Distribution 679
I.8 Triangular Distribution 680
I.9 "House" Distribution 682

J Digital Dither 685
J.1 Quantization of Representable Samples 686
J.1.1 Dirac Delta Functions at $q/2 + kq$ 688
J.2 Digital Dither with Approximately Normal Distribution 689
J.3 Generation of Digital Dither 689
J.3.1 Uniformly Distributed Digital Dither 690
J.3.2 Triangularly Distributed Digital Dither 693

K Roundoff Noise in Scientific Computations 697
K.1 Comparison to Reference Values 697
K.1.1 Comparison to Manually Calculable Results 697
K.1.2 Increased Precision 698
K.1.3 Ambiguities of IEEE Double-Precision Calculations 698
K.1.4 Decreased-Precision Calculations 700
K.1.5 Different Ways of Computation 700
K.1.6 The Use of the Inverse of the Algorithm 702
K.2 The Condition Number 703
K.3 Upper Limits of Errors 705
K.4 The Effect of Nonlinearities 707

L Simulating Arbitrary-Precision Fixed-Point and Floating-Point Roundoff in Matlab 711
L.1 Straightforward Programming 712
L.1.1 Fixed-point roundoff 712
L.1.2 Floating-Point Roundoff 712
L.2 The Use of More Advanced Quantizers 713
L.3 Quantized DSP Simulation Toolbox (QDSP) 716
L.4 Fixed-Point Toolbox 718

M The First Paper on Sampling-Related Quantization Theory 721

Bibliography 733

Index 742

Addenda Published on this Book's Website Only[1]

N Comparison of the Characteristic Function Method and Sheppard's Approach **W3**
N.1 The Euler-Maclaurin Summation Formula W3
N.2 Derivation of Sheppard's Corrections W6
N.3 Approximations in the Derivation of Sheppard's Corrections W8
N.4 Sheppard's Corrections and the Characteristic Function Method W10

O Interpolation of the Cumulative Distribution Function from the Histogram and Numerical Reconstruction of the Input PDF **W23**
O.1 Sampling Theorems for Cumulative Distribution Functions W23
O.2 Convergence of the Interpolation Formula W26
O.3 Numerical Differentiation W28

P Small Bit-Number Correlation **W29**
P.1 Hybrid Sign Correlator W29
P.2 Polarity Coincidence Correlator W30

Q Noise Shaping and Sigma-delta Modulation **W33**

R Second-order Statistical Properties of a Triangle-Wave Signal **W37**
R.1 The Signal Model W37
R.2 Derivation of the Joint PDF W38
R.3 Derivation of the Joint CF W41
R.4 Exercises W48

S Characteristic Functions of Quantities Involved when Using Dither **W49**
S.1 Calculation of the Joint Characteristic Functions of ξ, x and $(x+d)'$ W49
S.2 Calculation of the Joint Characteristic Functions of Input/Output Quantities, Dither, and Quantization Errors W51
S.3 A General Theorem Concerning Subtractive Dither W55

T Kind Corrections **W57**
T.1 Numerical Evaluation of the Residual Errors of Sheppard's Corrections for a Gaussian Input W57
T.2 Kind Corrections for One Variable W62
T.2.1 General Expressions for the Centralized Moments of the Quantizer Output, and for the Errors of the Kind Corrections W64

[1] http://www.mit.bme.hu/books/quantization/

T.2.2 General Expressions for the Centralized Moments of the Quantizer Output W65
T.2.3 General Expressions for the Errors of the Kind Corrections W72
T.3 Centralized Moments of the Quantizer Output for Gaussian input W76
T.4 Kind Corrections for Two Variables W81
T.5 Exercises W81

U Comparison of the Engineers' Fourier transform and Definition of the Characteristic Function W85

V A Few Papers from the Literature of Quantization Theory W89