Downloaded from orbit.dtu.dk on: Aug 22, 2022

DTU Library

- Modulators - Stability Analysis and Optimization

Risbo, Lars

Publication date:
1995

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Risbo, L. (1995). *- Modulators - Stability Analysis and Optimization*. Technical University of Denmark.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Σ-Δ Modulators—Stability Analysis and Optimization

Lars Risbo
Electronics Institute, Bldg. 349,
Technical University of Denmark, DK-2800 Lyngby, Denmark
e-mail lrisbo@eiffel.ei.dth.dk

June 16, 1994

Contents

List of Figures v

List of Tables xi

Preface xiii

Abstract xv

List of Symbols xvii

1 Introduction 1
- 1.1 Oversampled One-bit Signal Encoding 1
- 1.2 The Linear Decoding Constraint 2
- 1.3 Σ-Δ Modulation 2
- 1.4 Σ-Δ Modulation with Nonlinear Decoding 4
- 1.5 Overview 4

I Fundamental Analysis 7

2 The Σ-Δ Modulator as a Dynamical System 9
- 2.1 Introduction 9
- 2.2 Introducing Dynamical Systems 9
- 2.3 Basic Properties of Dynamical Systems 10
  - 2.3.1 Limit Sets 11
  - 2.3.2 The Stability of Fixed and Periodic points 11
  - 2.3.3 Chaos 12
- 2.4 The Σ-Δ Modulator 12
- 2.5 Σ-Δ Modulator Equivalences 14
- 2.6 The Noninvertible Region 18
- 2.7 Local Stability Analysis and Limit Cycle Identification 19
- 2.8 Summary 22

3 The Stability of Chaotic Σ-Δ Modulators 23
- 3.1 Introduction 23
- 3.2 Chaotic Σ-Δ Modulators 23
- 3.3 Boundary Crisis 24
- 3.4 Finding the Basin of Attraction 27
- 3.5 The Escape Rate 28

3.6 The Structure of L and B_L . . . 31
3.7 Reverse Time Simulation . . . 35
3.8 Unreliable Modulators . . . 43
3.9 Nonchaotic Modulators . . . 47
3.10 Summary . . . 48

4 Stability Analysis using Symbolic Dynamics 51
4.1 Introduction . . . 51
4.2 Symbolic Dynamics . . . 51
4.3 Analytical Determination of the Escape Rate . . . 52
4.4 Counting Limit Cycles . . . 55
4.5 Considerations for Nonhyperbolic Modulators . . . 59
4.6 Summary . . . 60

5 The Stability of Non-chaotic Modulators 63
5.1 Introduction . . . 63
5.2 Third Order Examples . . . 63
5.3 Comparison with Chaotic Modulators . . . 70
5.4 Summary . . . 71

II Modeling, Design and Optimization 73

6 Quasilinear Modeling 75
6.1 Introduction . . . 75
6.2 Loop Analysis with Linearized Quantizer Model . . . 75
6.3 Quasilinear Modeling . . . 77
6.4 Estimating the Quantizer Parameters . . . 79
6.5 Equilibrium and Stability . . . 82
6.6 The Gaussian Stability Criterion . . . 85
6.7 Loop Filter Design Using NFT-prototypes . . . 86
6.8 Predicting the Quantizer Input pdf . . . 92
6.9 Summary . . . 96

7 Optimizing Feedback Filters 99
7.1 Introduction . . . 99
7.2 The Choice of Parameter Space . . . 100
7.3 BIBO stability . . . 100
7.4 Reliability . . . 103
7.5 Performance Prediction . . . 112
7.6 Optimization Strategies . . . 114
7.7 Considerations for Higher Order Designs . . . 117
7.8 Optimization of Unusual Topologies . . . 119
7.9 Summary . . . 120

8 Designing Tone Free Modulators 123
8.1 The Tone Problem . . . 123
8.2 Methods for Tone Suppression . . . 125
8.3 Measures of Information Loss . . . 128

8.4 Limit Cycle Analysis of Dithered Modulators . . . 129
8.5 Summary . . . 132

9 Σ-Δ Modulation using Vector Quantization 133
9.1 Introduction . . . 133
9.2 Specification of the VQ . . . 135
9.3 The Back-Step Algorithm . . . 135
9.4 summary . . . 136

10 Conclusions 139

A Appendix for Chapter 8 141
A.1 Introduction . . . 142
A.2 Feedback Filter Design from NTF-prototypes . . . 143
A.3 Chaotic Sigma-Delta Modulators . . . 143
A.4 NTF-prototypes with All-pass Terms . . . 144
A.5 Design and Simulation Examples . . . 145
A.6 Conclusions . . . 146

B Appendix for Chapter 9 149

C Implementation of an Audio DAC 155
C.1 Introduction . . . 156
C.2 Converter Overview . . . 157
C.3 Sigma Delta Modulator . . . 157
  C.3.1 Design and Analysis . . . 157
  C.3.2 Modulator Implementation . . . 158
C.4 Analog Section . . . 160
  C.4.1 One-bit DAC overview . . . 160
  C.4.2 Spectral DAC Properties . . . 160
  C.4.3 Clock Jitter Sensitivity . . . 161
  C.4.4 Sensitivity to Unequal Rise and Fall Times . . . 162
  C.4.5 Deterministic Components Near $f_s/2$ . . . 162
  C.4.6 Pulse DAC Implementation . . . 163
  C.4.7 Current Switch NRZ DAC Implementation . . . 163
  C.4.8 Analog Post Filter . . . 163
  C.4.9 Jitter Attenuator . . . 164
C.5 Measurement Results . . . 164
C.6 Conclusions . . . 165

Bibliography 173

Danish Summary 178

List of Figures

1.1 Oversampled one-bit signal transmission 1
1.2 Δ modulator. The block marked $\int$ is a discrete-time delaying integrator. . . 3
1.3 First order Σ-Δ modulator. The block marked $\int$ is a discrete-time delaying integrator. 3

2.1 Σ-Δ modulator. 13
2.2 Second order Σ-Δ modulator example. 15
2.3 Σ-Δ modulator with feedback filter scaling. 15
2.4 Generic Σ-Δ modulator. 16
2.5 Generic Σ-Δ modulator implemented as a usual modulator with a prefilter. 17
2.6 Second-order multiple feedback Σ-Δ modulator. 18

3.1 First order map $\mathcal{F}(e)$ according to Eq. (3.1) for zero input and $a > 1$. . . . 25
3.2 Bifurcation diagram of the first order modulator cf. Eq. (3.1) with zero input. The two hyperbolas are the loci of the two unstable fixed points for $a > 1$, cf. Eq. (3.3). 26
3.3 Basin of attraction B_L and limit set L for the modulator cf. Eq. (3.5) with $h_1 = 0.8$ and $h_2 = -1.5$. 27
3.4 Basin of attraction B_L and limit set L for the modulator cf. Eq. (3.5) with $h_1 = 0.8$ and $h_2 = -1.65$. 28
3.5 Estimate of the basin of attraction B_L for the modulator cf. Eq. (3.5) with $h_1 = 0.8$, $h_2 = -1.8$ and maximum iteration number $n_{max} = 100$. 29
3.6 Estimate of the basin of attraction B_L for the modulator cf. Eq. (3.5) with $h_1 = 0.8$, $h_2 = -1.8$ and maximum iteration number $n_{max} = 500$. 29
3.7 Survival probability $P(k)$ as a function of the iteration number k. 30
3.8 The geometry of the multiple feedback modulator from Example 3.4 for $a < 0$ and $b < \sqrt{-a}$. 32
3.9 The geometry of the multiple feedback modulator from Example 3.4 for $a < 0$ and $b > \sqrt{-a}$. 34
3.10 Reverse time simulation revealing the chaotic repeller of a stable modulator cf. Eq. (3.5) with $h_1 = 0.8$ and $h_2 = -1.5$. Compare the plot to Fig. 3.3. . 36
3.11 Reverse time simulation revealing the chaotic repeller of an unstable modulator cf. Eq. (3.5) with $h_1 = 0.8$ and $h_2 = -1.8$. 37
3.12 Survival probability $P(k)$ vs. iteration number k for the chaotic repeller on ∂B_L for the modulator cf. Eq. (3.5) with $h_1 = 0.8$ and $h_2 = -1.5$. The plot indicates an escape rate $\gamma = 0.175$. 38
3.13 Basin of attraction and limit set for the modulator cf. Eq. (3.18) with zero input and $b = 0.45$. 39

3.14 Magnified plot of the basin of attraction shown in Fig. 3.13 The center of the spiral is a point on the $(\overline{6,6})$ limit cycle. 40
3.15 Basin of attraction and limit set for the modulator cf. Eq. (3.18) with zero input and $b = 0.5$. The $(\overline{7,7})$ limit cycle is plotted with circles. 41
3.16 Reverse time simulation of the modulator cf. Eq. (3.18) with zero input and $b = 0.56$. The (6,6) and (7,7) limit cycles are plotted with crosses and circles, respectively. 41
3.17 Minimum quantizer input magnitudes of a number of $(\overline{p,p})$ limit cycles for the second order modulator of Eq. (3.18) as a function of b. Magnitudes under the dashed line indicates that the corresponding limit cycle enters the noninvertible region U. 42
3.18 Chaotic repeller obtained by reverse time simulation for the modulator cf. Eq. (3.18) with zero input and $b = 1.1$. 42
3.19 Basin of attraction and limit set for the modulator cf. Eq. (3.18) with zero input and $b = 1.7$. Notice the behavior of the limit set around the repelling fixed points at $e^* = \pm 0.7/0.15 = \pm 4\frac{2}{3}$. 44
3.20 Basin of attraction for the modulator cf. Eq. (3.18) with zero input and $b = 0.26$. 45
3.21 Limit set for the modulator cf. Eq. (3.18) with zero input and $b = 0.26$. . . 45
3.22 Magnified plot of the basin of attraction for the modulator cf. Eq. (3.18) with zero input and $b = 0.26$. The center of the plot is a point on the $(\overline{4,4})$ limit cycle. 46
3.23 Magnified plot of the basin of attraction for the modulator cf. Eq. (3.18) with zero input and $b = 0.26$. The center of the plot is a point on the $(\overline{3,3})$ limit cycle. 46
3.24 A gray-scale plot of the mean survival time of the modulator of Example 3.4 with $a = -0.05$. The white region has a mean survival time of a minimum of 1000 time-steps. Darker regions have lower survival time. 48

4.1 Limit cycles of period up to 12 for the modulator cf. Eq. (4.10) with $a = 2.3$. 55
4.2 Number of periodic points l_n with period up to 16 versus the period length n for the modulator cf. Eq. (3.5) with with $h_1 = 0.8$, $h_2 = -1.5$ and zero input. The dashed line corresponds to the function $\Lambda^n = |h_2|^n$. 57
4.3 Limit cycles with period up to 16 for the modulator cf. Eq. (3.5) with $h_1 = 0.8$, $h_2 = -1.5$ and zero input. 57
4.4 Reverse time simulation of the system cf. Eq. (4.14) with zero input. . . . 58
4.5 The hatched area is the set of survivors Ω_n for the nonhyperbolic system cf. Eq. (4.18) 60

5.1 Transient leading to an overload cycle for the modulator cf. Eq. (5.1) with $m = 0.7$, $v = 0.09$ and constant input 0.4. Quantizer input $e(k)$ is plotted versus k 64
5.2 The onset of an overload cycle for the modulator cf. Eq. (5.1) with $m = 0.7$, $v = 0.09$ and constant input 0.4. The plot is the same as in Fig. 5.1 but with different axes. 65
5.3 Survival plot for the modulator cf. Eq. (5.1) with $m = 0.7$, $v = 0.09$ and constant input 0.4. 65
5.4 Example of a granular limit cycle. 66

5.5 Survival plot for the modulator cf. Eq. (5.1) with $m = 0.7$, $v = 0.09$ and constant input 0.5. 67
5.6 Maximum length of unstable transients versus constant input for the third order modulator cf. Eq. (5.1) with $m = 0.7$ and $v = 0.09$. 68
5.7 The fraction of survivors (i.e. initial conditions leading to granular cycles) for the modulator cf. Eq. (5.1) with $m = 0.7$, $v = 0.09$. 69
5.8 Maximum length of unstable transients versus constant input for the third order modulator cf. Eq. (5.1) with $m = 0.75$ and $v = 0.09$. 69
5.9 Maximum length of unstable transients versus constant input for the third order modulator cf. Eq. (5.1). The poles are mirrored outside the unit circle in order to get a chaotic system and the parameters $m = 0.9$ and $v = 0.11$ were used. 71

6.1 Generic Σ-Δ modulator with linearized quantizer model 75
6.2 A(K)-curves for the type I, II and III loop filter examples in Table 6.1. . . 80
6.3 σ_q^2 versus $m_y > 0$ for three types of probability distributions, cf. Eq. (6.15), (6.16) and (6.17). 81
6.4 Noise amplification factor A versus m_y for these three types of probability distributions, cf. Eq. (6.15), (6.16) and (6.17). 82
6.5 Noise amplification factor A versus K for the first order feedback filter $H(z) = z^{-1}/(1 - 2z^{-1})$. The A$_{min}$-point is shown with a dot. 84
6.6 Estimated quantizer input pdf for the first order modulator with $H(z) = z^{-1}/(1 - 2z^{-1})$. Obtained using 200 uniformly spaced bins from -1 to 1 and 1 mill. samples. 85
6.7 Scaling and A$_{min}$ versus Rs (in dB) for fifth order Chebychev II filters with fb = 1/64. 88
6.8 Simulated maximum stable input amplitude versus Rs for fifth order Chebychev II filters with fb = 1/64. The smooth curve is the Gaussian A$_{min}$-criterion of Eq. (6.20). 88
6.9 Simulated and theoretical (dashed lines) A(K)-curves for filter A, B and C, cf. Table 6.2. 90
6.10 Noise amplification factor A versus dc-input found by simulations of filter A, B and C, cf. Table 6.2. Theoretical curves for Gaussian and uniform quantizer input pdf are shown with dashed lines. 90
6.11 Quantizer gain K versus dc-input found by simulations of filter A, B and C, cf. Table 6.2. 91
6.12 Power spectra of the modulator output $y(k)$ and the quantization noise $q(k)$ for a 10 million time step simulation of filter B, cf. Table 6.2 with a dc-input of 1/256. Note the frequency scale normalized in respect to half the sample rate. 92
6.13 Quantization noise q versus quantizer input e for the quasilinear quantizer model, cf. Eq. (6.23) with displacement factor d> 0. 93
6.14 Predicted (smooth curve) and simulated quantizer input pdf for a modulator with filter B, cf. Table 6.2 and zero input. 95
6.15 Predicted (smooth curve) and simulated quantizer input pdf for a modulator with feedback filter B, cf. Table 6.2 with $m_e = 0.2$ corresponding to $m_y = 0.2443$. 96

6.16 The impulse response $\text{etf}_K(n)$ for the feedback filter $H(z) = z^{-1}/(1-2z^{-1})$ and $K = 1.5$. 97

7.1 A Noise Shaper with a linear prefilter. The system is equivalent to a usual Σ-Δ modulator. 101

7.2 $S(K)$-curves for three feedback filters: $H_1(z) = z^{-1}/(1-2z^{-1})$, $H_2(z) = (z^{-1} - 0.5z^{-2})/(1-2z^{-1}+z^{-2})$ and $H_3(z) = (z^{-1} - 0.6z^{-2})/(1-2z^{-1}+z^{-2})$ 103

7.3 Contours of equal S_{min} and A_{min} for $t_3 = 0.7$. The parameters t_1, t_2 and t_3 determine the feedback filter zeros, cf. Eq. (7.1). The pole locations are given by Eq. (7.12). The S_{min} contours are shown for increments of 0.1 from 2.9 to 3.7. The A_{min}-contours are for 2.2, 2.4 and 2.6. Three points (A,B and C) are marked on the $\mathrm{A}_{min} = 2.4$ contour. 104

7.4 Plot of maximum transient length versus constant input for the modulator A, cf. Fig. 7.3. For simulation parameters see Example 7.2. 105

7.5 Plot of maximum transient length versus constant input for the modulator B, cf. Fig. 7.3. For simulation parameters see Example 7.2. 105

7.6 Plot of maximum transient length versus constant input for the modulator C, cf. Fig. 7.3. For simulation parameters see Example 7.2 106

7.7 Fraction of initial conditions surviving 2000 time steps for the modulators B and C, cf. Fig. 7.3. For simulation parameters see Example 7.2 108

7.8 Plot of the $\text{ntf}_K(k)$ impulse responses which achieves the S_{min}-values for modulator A(+), B(o) and C($\times$), cf.Fig. 7.3. The plot shows only the tails for $k \geq 2$. 109

7.9 Magnitude responses of the $F_{min}(z)$ filters with minimum one-norm for modulator A, B and C, cf. Fig. 7.3. 110

7.10 Power spectrum estimate for the simulated output of modulator C, cf. Fig. 7.3. A constant input of 1/64 was used. Notice the linear axes and the peak near $2f/f_s \approx 0.095$ 111

7.11 Contours of constant F_{max} for $t_3 = 0.7$. The $\mathrm{A}_{min} = 2.4$ contour is plotted and the points A,B and C, cf. Fig. 7.3, are marked. 111

7.12 Contours of predicted σ_b^2 for $t_3 = 0.7$. The $\mathrm{A}_{min} = 2.4$ contour is also plotted and the points A,B and C, cf. Fig. 7.3, are marked. 113

7.13 Maximum transient plot of an optimized modulator with the constraints $\mathrm{A}_{min} < 2.4$, $\mathrm{F}_{max} < 1.8$ and $\mathrm{S}_{min} < 3.5$. 115

7.14 Estimated output power spectrum for the optimized modulator with the constraints $\mathrm{A}_{min} \leq 2.4$, $\mathrm{F}_{max} \leq 1.8$ and $\mathrm{S}_{min} \leq 3.5$. 116

7.15 Estimated power spectra for the two optimized 8th order modulators in Example 7.11. The dotted line represents the modulator optimized without an STF magnitude constraint. 119

7.16 A proposed extension of the usual multiple feedback modulator (the figure is a modified version of a topology suggested in [43]). The topology allows an equivalent feedback filter with more zeros than poles. 120

8.1 Estimated power spectrum of the decimated, high-pass filtered and squared modulator output based on 100.000 decimated samples simulated with a constant input of $1/(64 \cdot 128)$. 124

8.2 Decimated and high-pass filtered output signals of the fourth order modulator of Example 7.8 fed with a constant input of $1/(64 \cdot 512)$. The upper graph is decimated 64 times and the lower is decimated 32 times. 125

8.3 Σ-Δ modulator with scaled feedback filter $H(z)$ and dither source $d(k)$. . . 126

8.4 Equivalent noise-shaper topology of a Σ-Δ modulator with scaled feedback filter and dither source (see Fig. 8.3. The filter $F_K(z)$ is given by Eq. (7.7). 127

8.5 An Minimum dither magnitude (MDM) histogram for the set of period 16 prime cycles of the modulator with feedback filter cf. Eq. (8.4) with zero input. A total of 4080 prime cycles of length 16 exist cf. Table 4.1. 130

8.6 Number of periodic points versus period length for the modulator cf. Eq. (8.4) with zero input. The parameter on each plot is the peak dither magnitude $\|d(k)\|_\infty$. 132

9.1 Σ-Δ modulator with vector quantization 133

9.2 Σ-Δ modulator with vector quantization implemented as a deterministic dither signal 134

9.3 Simulated power spectrum estimate (500.000 samples) for an eighth-order modulator obtained from a Chebychev II prototype with the parameters Rs= 108 dB and fb= $1/32 \cdot 20/24$. The modulator was simulated using the back-step algorithm with a $|e(k)| \leq 1.065$ bound and a constant input of 1/256. 137

A.1 Sigma-Delta modulator with feedback filter $H(z)$ 146

A.2 Power spectrum from a simulation of a sixth order modulator with Rs = 120 dB. 147

A.3 Power spectrum from a simulation of a sixth order modulator with Rs = 90 dB and an NTF-zero scaling of 1.06. 147

A.4 Power spectrum from a simulation of a sixth order modulator with Rs = 95 dB using dither uniformly distributed over the interval [-0.5,0.5]. 148

A.5 Power spectrum from a simulation of a fifth order modulator with Rs = 80 dB extended to sixth order using an all-pass term with $\alpha = -1.25$. . . . 148

C.1 D/A converter overview showing three digital and three analog building blocks. 165

C.2 Sigma-Delta modulator with multiple feedback topology for even order N. $x(k)$ is the input and $p(k)$ is the output. The blocks marked $\int$ are time-discrete delaying integrators with transfer function $1/(z+1)$. 166

C.3 Power density spectra estimated using averaged 8k FFTs with Kaiser-Bessel window on bit-exact 24b simulations of the modulator output (0 dB corresponds to a full-scale sinewave). Upper graph: Full spectrum of modulator with DC-input 0.2 (2 mill. samples). Lower graph: 32 times decimated modulator output with DC-input 1/512 (500.000 decimated samples). . . . 166

C.4 In-band modulator output RMS noise versus DC-output measured by bit-exact 24b simulations and decimation. Each of the approx. 512 data points is based on decimation of 1.6 mill. modulator samples. The spikes at the end of the plot are the onset of instability. 167

C.5 FPGA implementation of a second order multiple feedback modulator section. Four of these sections are cascaded to produce an 8th order modulator. The 'SignOut' signal of the last section is the output, $p(k)$, which is fed back to the ADD/SUB control line. 167
C.6 Modulator power spectrum for sinusoidal input with frequency $0.025f_s/2$ and amplitude 0.2. Notice the side-bands at $2f/f_s = 0.975$, 0.95, 0.925 etc. 168
C.7 The implemented pulse type one-bit DAC. The circuit is controlled by the modulator output $p(k)$ and two non-overlapping clocks, φ_1 and φ_2. 168
C.8 The implemented NRZ current switch one-bit DAC. The output current is either $+i_{\text{ref}}$ or $-i_{\text{ref}}$ depending on the digital input code $p(k)$. The adjustable voltage source adjusts the skewness of the low to high and high to low transitions relative to the clock. 169
C.9 Amplitude response for the analog post filters. 169
C.10 The implemented passive and active analog post filters. 170
C.11 Output spectra of the entire converter with 0dB 1kHz sinewave input quantized to 18 bit using triangular dither. The spectra are measured using a notch filter and 64 times averaged 16k FFTs at sample rate 44.1 kHz. Upper graph: NRZ DAC and passive post filter. Lower graph: Pulse DAC and active filter. 170
C.12 Level linearity of the pulse type DAC and active filter with 18-bit triangular dithered sine input at 500 Hz. 171

List of Tables

4.1 The number p_k of prime cycles of length k for $k = 1..16$ 56

4.2 The number p_k of existing prime cycles of length k for $k = 1..16$ for the modulator cf. Eq. (3.5) with zero input and $h_1 = 0.8$, $h_2 = -1.5$. 56

6.1 A(K)-curves 79

6.2 Parameters for three feedback filter examples 87

Preface

The present thesis has been submitted in partial fulfillment of the requirements for the Ph.D. degree at the Technical University of Denmark. The work has been carried out at the Electronics Institute Bldg. 349 during the period Nov. 1 1991 to April 31 1994.

My interest in the field of Σ-Δ modulation originates from the fall of 1990 where Compact Disc players with bitstream digital-to-analog converters became widespread. From the beginning of the era of the Compact Disc, the manufacturers were competing on higher and higher numbers of bits in claimed resolution, and suddenly a new type of converters were introduced using only a single bit.

I grew very fond of these new strange single-bit converters due to my background in the field of digital signal processing combined with my affection for music and systems for sound reproduction.

In february 1991, I started my masters project at the Electronics Institute. My topic was how to design stable high-order Σ-Δ modulators. The result of the project was a number of design tools and an approximate stability criterion based on quasilinear analysis. One of the chapters in this thesis includes extensions of the key results.

I received a scholarship from the Technical University of Denmark in november 1991, which enabled me to proceed with a Ph.D. study. My intention was partly to implement a Σ-Δ digital-to-analog converter with the best possible sound quality and partly to make further investigations on the stability problems concerned with Σ-Δ modulation.

I would like to thank a number of persons for their help and contributions. First of all, I would like to thank my advisors assoc. professors Peter K. Møller and John Aasted Sørensen for their never failing support and their efforts in making a pleasant and stimulating climate in the digital signal processing group. Peter A. Toft, Jan Larsen and Carsten Knudsen are thanked for a critical proof reading and review of the manuscript.

I am also very grateful for a number of persons with whom i have had enlightening discussions on the theoretical work: Robert Adams (Analog Devices USA), Truong-Thao Nguyen (Columbia Univ.), HongMo Wang (Bell Labs.), the “Junta” at the Electronics Institute (Jan Larsen, Peter A. Toft, Kim V. Hansen), Lars K. Hansen (EI) and Carsten Knudsen (MIDIT/CATS).

The practical implementation of the audio DAC was carried out in cooperation with Thomas Mørch (LIE) who contributed with a part of the printed circuit design. Jørgen Krogh (Analog Devices, Denmark) contributed with chip samples for evaluation in the DAC. Concerning the analog design, i received a lot of good advice from Peter Shah (EI/Imperial College, UK), Gudmundur Bogason (EI) and Michael Smedegaard Pedersen (LIE/Harman Marketing Europe).

Abstract

Σ-Δ modulation is widely used in high-performance data converters. The modulation process encodes a high-resolution discrete-time signal into a two-level (i.e, one-bit) signal with an increased sample rate. This provides speed to be traded for resolution. In the spectral domain, this effect is known as noise-shaping, i.e., the error of the one-bit signal is removed from a low-frequency band and concentrated at high frequencies.

A Σ-Δ modulator consists of a one-bit quantizer embedded in a feedback loop with a linear loop filter. Consequently, these modulators are highly nonlinear dynamical systems. For some types of loop filters, the modulators may be chaotic. One of the major drawbacks of Σ-Δ modulators is the potential risk of instability. This is especially the case for modulators with high-order loop filters. Instability is normally considered as the major problem of Σ-Δ modulator design.

The stability problem is pursued in this thesis. This is initially done from a nonlinear dynamics point of view. One of the conclusions is that chaotic modulators loose the stability due to a boundary crisis emerging when the attractor collides with the associated basin of attraction. The degree of instability is characterized by the so-called escape rate.

The stability is also investigated using *symbolic dynamics*. It is shown that the stability is closely related to the number of admissible limit cycles.

Some modulators are defined as being *unreliable*, i.e., the transition to the unstable regime is not a well defined point. Such modulators can be seemingly stable using simulations of moderate length. The instability will only appear in real-time implementations or very long simulations.

The thesis presents constrained optimization methods which are capable of finding reliable modulators with optimum signal-to-noise ratio. The method combines exact state-space analysis with approximate quasilinear modeling. The optimization tools have been used for the design of a 32 times oversampling 8th order Σ-Δ audio D/A-converter implemented using Field Programmable Gate Array (FPGA) circuits

Methods for suppression of the predominant tones near half the sample rate are investigated. The use of dithering and chaos are found to be equally efficient and equivalent in many respects. A new method for designing chaotic modulators is also introduced, i.e., the use of all-pass terms in the noise transfer function.

Finally, a new class of modulators using one-bit vector quantization is proposed. It is shown that the introduction of a vector quantizer can enhance the stability. Furthermore, the vector quantizer seems to reduce the amplitude of the tone near half the sample rate.

List of Symbols

n, k	Discrete time indices.
f_s	The sample rate.
$x(k)$	Modulator input signal.
$y(k)$	Modulator output signal.
$e(k)$	Quantizer input signal (error signal).
σ_e^2	Variance of $e(k)$.
$\mathcal{F}(\cdot)$	System map of a discrete-time dynamical sytem.
$\mathrm{D}\mathcal{F}(\cdot)$	Functional matrix (Jacobian) of $\mathcal{F}$.
$\mathcal{F}^{-1}(C)$	Denotes the preimage of the set C with respect to $\mathcal{F}$, i.e., the elements for which $\mathcal{F}(x) \in C$.
L	Limit set, defined in Definition 2.2.
B_L	Basin of attraction for the limit set L, defined in Definition 2.3.
∂B_L	Boundary of the basin of attraction for the limit set L.
$\boldsymbol{A}$	Transition matrix of a linear filter described in the state-space.
$\boldsymbol{x}_n$	State vector at time-step n.
S	The state-space of a modulator.
S_+	The part of the state-space for which a positive code is produced.
S_-	The part of the state-space for which a negative code is produced.
W	A bounding set which usually is supposed to surround a limit set.
U	The noninvertible region of a map $\mathcal{F}$ associated with a modulator.
V	The preimage of U.
$\Pr\{\cdot\}$	Probability.
γ	The escape rate, defined in As. 3.2.

$(\overline{p,k})$	A periodic output code sequence with period $p+k$ composed of p positive codes followed by k negative codes.
Λ	The expansion factor given as the product of the moduli of the system eigenvalues outside the unit circle, defined in Eq. (4.7).
$\|\cdot\|_1$	One-norm, i.e., sum of magnitudes.
$\|\cdot\|_\infty$	Infinity-norm, i.e., maximum magnitude.
$\|\cdot\|_2$	Two-norm, i.e., the square root of the sum of squares.
$H(z)$	Modulator feedback (loop) filter.
K	Linearized quantizer gain.
$q(k)$	Quantization noise of a quantizer.
σ_q^2	Variance of $q(k)$.
σ_b^2	Modulator output noise power within the base-band.
$\mathrm{NTF}(z)$	Noise Transfer Function from quantization noise source to modulator output, defined in Eq. (6.3). $\mathrm{ntf}(k)$ is the associated impulse response.
$\mathrm{STF}(z)$	Signal Transfer Function from modulator input to modulator output, defined in Eq. (6.2).
$\mathrm{ETF}(z)$	Error Transfer Function from quantization noise to quantizer input, defined in Eq. (6.26).
$\mathrm{A}(K)$	Noise amplification factor as a function of the quantizer gain, defined in Eq. (6.7)
A_{min}	Minimum of $\mathrm{A}(K)$.
$\mathrm{V}\{x\}$	Variance of x.
$\mathrm{Cov}\{x,y\}$	Covariance of x and y.
m_y	Mean value of the quantizer output $y(k)$.
pdf	Probability density function.
$\mathrm{S}(K)$	One-norm of $\mathrm{ntf}(k)$ versus the quantizer gain K, defined in Eq. (7.9).
S_{min}	Minimum of $\mathrm{S}(K)$.
MSA	Maximum Stable Amplitude. The maximal constant input amplitude for which a modulator is stable.

Chapter 1

Introduction

1.1 Oversampled One-bit Signal Encoding

When a signal is to be transmitted through a channel, a signal encoding scheme or a modulation is needed. The encoder accommodates the information of the signal into a form suitable for the channel. It is very common that the channel accepts a sequence of binary symbols, i.e., a one-bit signal.

A general signal transmission system using a one-bit channel is shown in Fig. 1.1. The signal source is in this case a discrete-time continuous amplitude signal which is low-pass filtered, i.e., the encoder input $x(n)$ is oversampled by the factor R= $\frac{f_s}{2f_c}$ where f_s is the sample rate and f_c is the filter cut-off frequency. The encoder transforms this oversampled signal into a one-bit signal ready for the channel. The decoder produces an estimate $\hat{x}(n)$ which approximates $x(k)$.

The performance of a signal transmission system can be assessed using a distortion measure such as the Mean Square Error (MSE), i.e., the mean value of $(x(n) - \hat{x}(n))^2$. It is also common to express the maximal signal-to-noise ratio (SNR), i.e., the maximum ratio between the signal power and the MSE. The SNR s typically measured for sinusoidal input.

In terms of information theory, the information capacity (rate) of the channel is one bit per sample. For a given signal source and information rate, the minimum attainable MSE is given by the so-called *Rate-Distortion* curve of the source [6]. The performance given

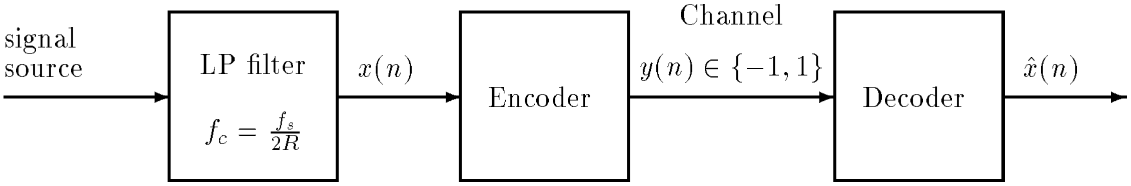

Figure 1.1: Oversampled one-bit signal transmission

by the rate-distortion curve is an upper limit, which is only attained using an optimum choice of encoder and decoder.

In a Pulse Code Modulation (PCM) system, the oversampled signal $x(n)$ is downsampled R times and sampled using a quantizer with 2^{R} levels, i.e., an R-bit quantizer. For a uniform and 'busy' quantizer, the MSE is approx. $\Delta^2/12$ where Δ is the quantization step size [19]. When a system is claimed to have R-bits of resolution, the MSE is similar to that of an R-bit PCM system. In general, resolution is traded for speed, i.e., lower MSE requires a higher oversampling factor R.

The decoding process in a PCM system is generally nonlinear: the incoming bits must be grouped together in order to form R-bit binary words. The significances or weights of the bits depend on the bit position within the word and this makes the decoder nonlinear.

1.2 The Linear Decoding Constraint

The decoder can be designed as a linear system, i.e., the decoder is a filter characterized by a transfer function or an impulse response. The use of linear decoding limits obviously the attainable performance of the transmission system, i.e., the encoder must take the linear decoding into account. The subject of this dissertation is encoders designed for use with an ideal lowpass filter as decoder. The linear decoding constraint makes it impossible to obtain R-bits of performance for R times oversampling.

One-bit encoding for use with linear decoding is very useful for the construction of data converters [9]. If a one-bit encoder is inserted in a general D/A converter, only a simple switch producing a two-level analog signal is needed. The two-level signal is then 'decoded' using an analog reconstruction low-pass filter. This approach simplifies the analog circuitry needed for a converter at the expense of a higher sample rate (R times oversampling) and a more complex digital encoding scheme. In an A/D converter, an analog one-bit encoder can be used as a front-end and the high-rate digital one-bit signal is then decoded using a digital decimation filter. Also in this case, the analog circuitry is simplified at the expense of more complex digital post-processing. Thanks to the rapid progress of digital VLSI technology, the use of oversampling one-bit converters has become widespread.

1.3 Σ-Δ Modulation

The designation Σ-Δ modulation (or Δ-Σ modulation) covers a class of one-bit encoders which is well suited for linear decoding (see [8] and [25] for a general introduction). The first member of this class was the Δ modulator which only needs an integrator for decoding (see Fig. 1.2). A Δ modulator compares the input signal with the integrated encoder output signal and if the input is higher than the integrator output, a $+1$ code is generated. In the opposite case, a -1 code is generated. This primitive type of encoder can only encode the difference (Δ) from sample to sample, hence the name Δ modulation.

The Δ modulator has very poor performance at low frequencies due to the differencing and therefore an integrator could profitably be placed before the Δ modulator. The overall system was named a Σ-Δ modulator due to the Δ modulator being preceded by an integrator (Σ). Using a simple system transformation, the Σ-Δ modulator can be constructed using only one integrator which accumulates the difference between the modulator input and output. This system is shown in Fig. 1.3. The Δ modulator actually contains a

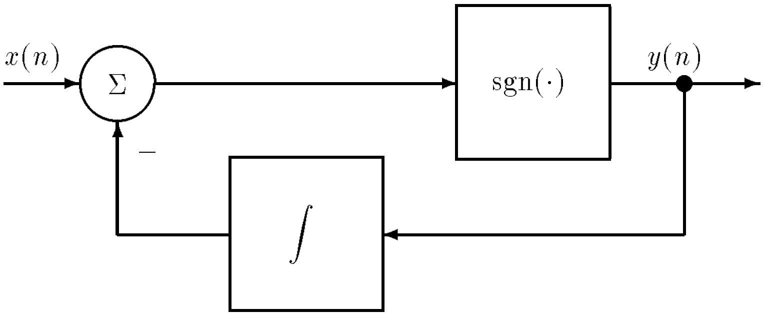

Figure 1.2: Δ modulator. The block marked ∫ is a discrete-time delaying integrator.

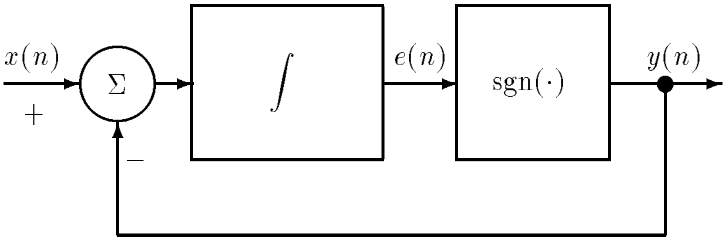

Figure 1.3: First order Σ-Δ modulator. The block marked ∫ is a discrete-time delaying integrator.

decoder (i.e., an integrator) and the output code decision is based on a comparison of the input and the decoded output, i.e., the encoder is constructed from a decoder using a feedback loop.

The simple discrete-time delaying integrator of Fig. 1.3 with transfer function $1/(z-1)$ can be replaced by a more complex higher order feedback filter with transfer function $H(z)$. The rôle of the filter is unchanged: the filter locally decodes the output code and error feedback is used to generate the output code. The feedback filter output $e(n)$ is the filtered difference between the modulator input and output. The feedback action tries to minimize the magnitude of this error signal. Consequently, if $H(z)$ is a low-pass filter, the coding error will be low for low frequencies due to the feedback. This phenomenon is known as *noise shaping*, i.e., the coding error (quantization noise) is concentrated at higher frequencies where the decoder (a low-pass filter) removes most of the error. For a first order loop filter with a pole at $z = 1$, the MSE is reduced 9 dB for each doubling of the oversampling ratio R [8].

The use of a more complex second order feedback filter was suggested in [8]. This allows a 15 dB reduction in MSE for each doubling of R, i.e., a more selective noise shaping. However, the author of [8] warned that feedback filters containing more than two integrators resulted in unstable modulators.

The stability problem was (and is) a severe obstacle to practical use of higher order Σ-Δ modulators. A major step forward was made by the authors of [10, 34] who proposed a general topology for high-order modulators which was shown to comprise practically useful and stable modulators. There are currently many commercial data converters using high-order modulators [1, 17] based on the pioneering work of [34]. The widespread scepticism about the stability of high-order modulators seems to prevail even despite such commercial devices. However, there are still many issues of the stability problem which need to be clarified.

The general (high-order) Σ-Δ modulator is a useful family of one-bit signal encoders that are constrained by a demand for linear (LP filter) decoding. The demand for stability is another encoder constraint which limits the attainable performance. An obvious question is of course: is it possible to find a better class of encoders which do not suffer from the stability problem?

1.4 Σ-Δ Modulation with Nonlinear Decoding

It has recently been discovered that the use of nonlinear decoding in conjunction with a Σ-Δ modulator as encoder can reduce the MSE compared to linear decoding [25, 39]. The proposed algorithms utilize a more specific knowledge of the non-linear encoding process.

Nonlinear decoding can be useful in A/D data converters where the decoding process is digital. A general nonlinear decoder may also compensate for circuit nonidealities in the analog encoder. However, for D/A converters, the decoding process takes place in the analog domain and this virtually excludes the possibility of nonlinear decoding.

1.5 Overview

The objective of the present dissertation is to clarify the complex issue of modulator stability and present methods for systematic and automated design of practically useful modulators. Everywhere, it is assumed that the used decoding is linear.

The dissertation is divided into two parts. Part I (Ch. 2–5) is devoted to fundamental stability analysis. Ch. 2 introduces the Σ-Δ modulator as a dynamical system using a rather mathematical description. Subsequently, basic properties of dynamical systems are discussed (e.g., *chaos, limit cycles etc.*). The purpose of the mathematical description is to isolate the modulator functionality from the more technological aspects and to provide a solid foundation for later stability analysis. In order to provide a good overview of the many different modulator topologies, the concept of modulator equivalence is introduced.

Ch. 3 focuses on chaotic modulators and various aspects of the onset of instability. The theory is accompanied by numerous examples and a measure of the degree of instability is introduced. Chaotic modulators are interesting for at least two reasons: they produce nonperiodic output and the use of chaos can help to suppress possible spurious tones in the modulator output spectrum. Secondly, the question of stability in conjunction with chaotic modulators is conceptually simpler than for nonchaotic modulators.

Ch. 4 introduces a framework of stability analysis based on symbolic dynamics, i.e., the study of the possible code sequences produced by a modulator.

Ch. 5 focuses on the stability of nonchaotic modulators. It is shown that nonchaotic high-order modulators are very complex and difficult to describe as either stable or unstable.

Part II (Ch. 6–9) presents methods and tools for systematic modulator design and optimization.

Ch. 6 introduces an analysis framework called quasilinear modeling, i.e., the use of linearized models. This chapter is an extension of [49, 51]. The chapter leads to the introduction of an approximate stability criterion.

Ch. 7 presents methods for systematic and automated modulator optimization. Useful constraints are found which ensures the design of stable and reliable modulators. The analysis in this chapter is a synthesis of the results obtained in part I and Ch. 6.

Ch. 8 is devoted to the problem of spurious modulator tones and noise modulation which are highly objectionable adverse effects in audio systems. A comparison of dithering (i.e., the addition of random noise) and the use of chaos for tone suppression is presented in an accompanying paper reprinted in App. A. The similarities between dithering and chaos are pointed out using results from information theory and symbolic dynamics (from Ch. 4).

Ch. 9 introduces a novel modulator topology which comprises a general vector quantizer. This is an attempt to answer the question: is it possible to find better one-bit encoders than Σ-Δ modulators? The use of vector quantization is used to enhance the stability of conventional modulators. This is shown in an accompanying paper reprinted in App. B.

App. C is an independent paper which is based on the methods presented in this dissertation. The paper describes the practical implementation of an entire eighth-order audio D/A converter for 32 times oversampling. Details of both the analog and digital circuitry are presented along with theory concerning possible error sources in D/A conversion (e.g., clock jitter and inter symbol interference).

Part I

Fundamental Analysis

Chapter 2

The Σ-Δ Modulator as a Dynamical System

2.1 Introduction

Σ-Δ modulators in analog-to-digital or digital-to-analog converters can be implemented in many ways; there are especially many different approaches related to different technologies: some designs are based on continuous time circuits and others are based on discrete-time switched capacitor or even switched current techniques. In addition, many different topologies are used.

This diversity of real implementations makes it at first sight very difficult to describe and analyze Σ-Δ modulators as one class of systems. In order to be able to understand and analyze such complex systems it is absolutely necessary to use a description which is technology independent. Therefore, a more mathematical approach, i.e., the concept of *nonlinear dynamical systems* will be introduced in the next sections. This will probably be regarded as just another way to introduce confusion by many readers not familiar with such mathematical description. The real purpose is, however, to give the reader a better overview and ability to see through different implementations. For instance, many seemingly different circuits can in fact be equivalent, i.e., they do the same signal processing but are implemented differently.

The mathematical description is especially useful for the design of Σ-Δ modulators: at first the modulator is optimized for functionality in the mathematical domain and, subsequently, the design is transferred to real circuit elements.

2.2 Introducing Dynamical Systems

In this section, a brief introduction to time-discrete dynamical systems will be given.

A given discrete-time dynamical system is characterized by a state-space S and two mappings $\mathcal{F} : S \times I \to S$ and $\mathcal{O} : S \to O$. The state-space is the set of possible states of the system and the $\mathcal{F}$ mapping describes the next state as a function of the current state and an independent system input taken from the set I. The output mapping $\mathcal{O}$ generates the system output as a function of the current system state and input:

$$\begin{array}{lcll} \boldsymbol{x}_{n+1} & = & \mathcal{F}(\boldsymbol{x}_n, \boldsymbol{i}_n) & , \mathcal{F} : S \times I \to S \\ \boldsymbol{o}_n & = & \mathcal{O}(\boldsymbol{x}_n, \boldsymbol{i}_n) & , \mathcal{O} : S \times I \to O \end{array} \tag{2.1}$$

where $\boldsymbol{x}_n \in S$, $\boldsymbol{i}_n \in I$ and $\boldsymbol{o}_n \in O$ are the system state, input and output at time-step n, respectively.

An Nth order system can be described by N independent state-coordinates which form an N-dimensional state-space. The chosen state-coordinates will generally describe physical properties of the system, e.g., voltages in circuit nodes, physical displacements, etc. The choice of state-coordinates for a given system is ambiguous, i.e., a multitude of different state-space descriptions of the same system can be given corresponding to different choices of independent state-variables. All of these state-space descriptions result in the same output signal sequence for the same input signal sequence. Generally, two state-space description of the same system have state-coordinates which are related through a one-to-one (bijective) transform $\mathcal{T}$. This fact motivates following definition:

Definition 2.1 *Two dynamical systems described by the maps* $\mathcal{F}_1 : S_1 \times I \to S_1$, $\mathcal{O}_1 : S_1 \times I \to O$, $\mathcal{F}_2 : S_2 \times I \to S_2$ *and* $\mathcal{O}_2 : S_2 \times I \to O$, *are defined as being* equivalent *if and only if a one-to-one map* $\mathcal{T} : S_2 \to S_1$ *exists satisfying:*

$$\begin{aligned} \mathcal{F}_2(\boldsymbol{y}_n, \boldsymbol{i}_n) &= \mathcal{T}^{-1}(\mathcal{F}_1(\mathcal{T}(\boldsymbol{y}_n), \boldsymbol{i}_n)) \,,\ \boldsymbol{y}_n \in S_2 \\ \mathcal{O}_2(\boldsymbol{y}_n, \boldsymbol{i}_n) &= \mathcal{O}_1(\mathcal{T}(\boldsymbol{y}_n), \boldsymbol{i}_n) \end{aligned} \tag{2.2}$$

The dynamical systems discussed so far, as described by Eq. (2.1), have been operating with an independent input signal sequence. A certain subset of the dynamical systems are the so-called autonomous systems which operate without input.

The sequence of states corresponding to time step n for an autonomous system is thus given by applying a mapping $\mathcal{F} : S \to S$ iteratively on an initial state $\boldsymbol{x}_0 \in S$:

$$\boldsymbol{x}_n = \mathcal{F}^n(\boldsymbol{x}_0) \,,\ n \in \mathbb{N} \tag{2.3}$$

Consequently, the study of the dynamics of autonomous systems is equivalent to the study of the iterations of a mapping. In fact, Eq. (2.3) defines a mapping between the initial state and the output state sequence which is called the *orbit* corresponding to the initial state $\boldsymbol{x}_0$, i.e., the sequence $\{\boldsymbol{x}_n\}_{n=0}^{\infty}$.

If the input signal of a given nonautonomous system can be generated from an autonomous system, the nonautonomous system and the input generating system can be merged into a single autonomous system. Especially, systems with constant input signal can generally be described as autonomous when the constant input is 'built' into the $\mathcal{F}$ mapping. This procedure will be used extensively in the following.

2.3 Basic Properties of Dynamical Systems

The discussion on dynamical systems will be focused on autonomous time-discrete systems cf. Eq. (2.3). Recall, that the initial condition $\boldsymbol{x}_0$ defines the orbit of the system. The *steady state* of a system refers to the asymptotic behavior as $n \to \infty$. The difference between the orbit and the steady state is called the *transient*. The following definitions are from [46]:

2.3.1 Limit Sets

Definition 2.2 *A point* $\boldsymbol{y}$ *of* $\boldsymbol{x}_0$ *for the map* $\mathcal{F}$ *is called a* limit point *of* $\boldsymbol{x}_0$ *if, for every neighborhood* O *of* $\boldsymbol{y}$*, the orbit* $\mathcal{F}^n(\boldsymbol{x}_0)$ *repeatedly enters* O *as* $n \rightarrow \infty$*.*

The set $L(\boldsymbol{x}_0)$ *of all limit points of* $\boldsymbol{x}_0$ *is called the* limit set *of* $\boldsymbol{x}_0$*.*

The limit set L of an initial condition $\boldsymbol{x}_0$ is the set in state-space the orbit visits frequently in steady state. In other words: a limit set characterizes a steady-state solution. A fundamental property of linear systems is that there is either one limit set or infinitely many limit sets. Nonlinear system may have a finite number of coexisting limit sets.

Definition 2.3 *A limit set* L *is* attracting *or* asymptotically stable *if there exists an open neighborhood* O *of* L *such that* $L(\boldsymbol{x}) = L$ *for all* $\boldsymbol{x} \in O$*. The union of all such neighborhoods* O *is called the* basin of attraction B_L *of an attracting limit set* L*.*

The basin of attraction is the set of points which are asymptotically attracted to the limit set. Stable linear systems have only one limit set. Furthermore, the limit set is attracting and the basin of attraction is the entire state-space, i.e., the limit set of a stable linear system is *globally attracting.*

Definition 2.4 *A* fixed point *or a* equilibrium point $\boldsymbol{x}^*$ *of a map* $\mathcal{F}$ *is a point for which* $\mathcal{F}(\boldsymbol{x}^*) = \boldsymbol{x}^*$*. The limit set of a fixed point is the fixed point itself.*

Stable linear systems with constant input have precisely one fixed point and this fixed point is the only limit set.

Definition 2.5 *A* periodic point $\boldsymbol{x}^p$ *of a map* $\mathcal{F}$ *is a point for which a number* $k \in \mathbb{N}$ *exists satisfying* $\mathcal{F}^k(\boldsymbol{x}^p) = \boldsymbol{x}^p$*. The least number* K *for which* $\mathcal{F}^K(\boldsymbol{x}^p) = \boldsymbol{x}^p$*, is called the* prime period *of the periodic point. A periodic point with prime period* K *is called a period-*K *point. The closed orbit* $\{\boldsymbol{x}^p, \mathcal{F}(\boldsymbol{x}^p), ...\mathcal{F}^K(\boldsymbol{x}^p)\}$ *is called a* limit cycle *or a* periodic orbit *which is the limit set of the period-*K *point.*

2.3.2 The Stability of Fixed and Periodic points

The stability of fixed points and periodic points can be studied by linearization. Let $\boldsymbol{x}^*$ be a fixed point of a map $\mathcal{F}$. The orbit $\boldsymbol{x}_n$ corresponding to the initial condition $\boldsymbol{x}^* + \delta\boldsymbol{x}_0$ can then be approximated to first order by:

$$\boldsymbol{x}_n = \mathcal{F}^n(\boldsymbol{x}^* + \delta\boldsymbol{x}_0) \approx \boldsymbol{x}^* + (\mathrm{D}\mathcal{F}|_{\boldsymbol{x}^*})^n \delta\boldsymbol{x}_0 \tag{2.4}$$

where $\mathrm{D}\mathcal{F}|_{\boldsymbol{x}^*}$ is the functional (Jacobian) matrix of $\mathcal{F}$, i.e. a matrix with the $N \times N$ partial derivatives evaluated at $\boldsymbol{x} = \boldsymbol{x}^*$.

The stability of the orbit, cf. Eq. (2.4), can in some cases be determined by the eigenvalues $\{\lambda_i\}$ of the linearized system, i.e. the eigenvalues of $\mathrm{D}\mathcal{F}|_{\boldsymbol{x}^*}$. If $|\lambda_i| < 1$ for all i then all sufficiently small perturbations will tend toward zero as $n \rightarrow \infty$, and the fixed point is attracting or asymptotically stable. Conversely, if any of the eigenvalues has modulus greater than unity, then perturbations in certain directions will grow with time, and the fixed point is consequently unstable.

Since a period-K point $\boldsymbol{x}^p$ of a map $\mathcal{F}$ is a fixed point of the map $\mathcal{F}^K$, the stability of limit cycles can be determined from the eigenvalues of $\mathrm{D}\mathcal{F}^{\mathrm{K}}(\boldsymbol{x})$. Using the chain rule:

$$\mathrm{D}\mathcal{F}^{\mathrm{K}}|_{\boldsymbol{x}^{\mathrm{p}}} = \mathrm{D}\mathcal{F}|_{\boldsymbol{x}_{\mathrm{K}-1}} \cdot \mathrm{D}\mathcal{F}|_{\boldsymbol{x}_{\mathrm{K}-2}} \cdots \mathrm{D}\mathcal{F}|_{\boldsymbol{x}_0} \tag{2.5}$$

where $\boldsymbol{x}_i = \mathcal{F}^i(\boldsymbol{x}^p)$ is the corresponding orbit or limit cycle.

2.3.3 Chaos

Some nonlinear dynamical systems have asymptotic orbits that are completely nonperiodic. This can be the case when the orbit can be dissolved into a countable sum of periodic functions each of whose periods are an integer combination of frequencies taken from a finite base set [46]. Such asymptotic behavior is called *quasiperiodic* and the orbits are nonperiodic but predictable on any time scale. For instance, the sum of two sine waves with incommensurable frequencies is nonperiodic but easily predictable. Even a single discrete-time sinusoidal signal with irrational frequency is nonperiodic.

In some cases a nonlinear dynamical system can exhibit another nonperiodic steady state behavior called *chaos*. Chaos can loosely be defined as a nonperiodic motion that is not quasiperiodic, i.e., a chaotic orbit cannot be described by a finite number base frequencies and the orbit cannot be predicted on long time scales. There is no commonly accepted definition for chaos but most definitions have the same ingredients stated in different ways. Most definitions require at least the first two of the following three properties to be fulfilled: A system map $\mathcal{F}$ is chaotic on a domain L if

1. There is a sensitivity to initial conditions, i.e., two nearby orbits diverge and will eventually become uncorrelated. This makes long term prediction impossible.

2. The system map is topologically transitive on L, i.e., points within any given neighborhood of L can eventually hit any point in L. This basically ensures that the entire domain L is visited repeatedly, i.e., L is actually a limit set, cf. Definition 2.2.

3. The periodic points are dense in L, i.e., every neighborhood of a point in L contains at least one periodic point.

Such mathematical definitions do not guarantee the existence of a nonperiodic solution. However, for most nice systems, fulfilling the above requirements, there will be a nonperiodic *dense orbit* that comes arbitrarily close to any point in the limit set.

A chaotic limit set is from a practical point of view just the union of infinitely many unstable limit cycles. A given initial point can always be perturbed slightly in order to get a periodic orbit of some period length. A chaotic limit set cannot contain any stable limit cycles since each stable limit cycle forms a stable limit set with a corresponding basin of attraction where there is no sensitivity to initial conditions.

A chaotic orbit is always close to many limit cycles but the orbit will be nonperiodic since every limit cycle is repelling. The unstable limit cycles will therefore form the skeleton of a chaotic limit set: the short cycles shows the basic structure and the longer cycles add more details [13]. Chaotic limit sets are often very complex geometric objects with a fractal dimension. A stable and fractal limit set is often called a *strange attractor*.Conversely, an asymptotically unstable chaotic and fractal limit set is a *strange repeller*.

2.4 The Σ-Δ Modulator

A Σ-Δ modulator is a linear filter with a binary feedback signal formed as the negated sign of the filter output (see Fig. 2.1). A linear and causal Nth order filter is a dynamical system as described by Eq. (2.1) for which the state-space is the Euclidean space, $S = \mathbb{R}^N$, and both of the mappings $\mathcal{F}$ and $\mathcal{O}$ are linear. The linear filter $H(z)$ in Fig. 2.1 can thus

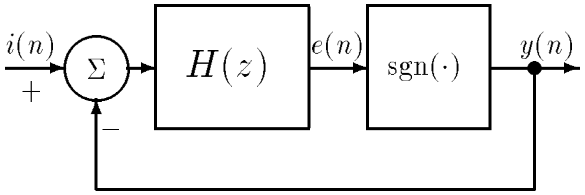

Figure 2.1: Σ-Δ modulator.

be described in state-space as follows [18]:

$$\begin{aligned} \boldsymbol{x}_{n+1} &= \boldsymbol{A}\boldsymbol{x}_n + \boldsymbol{b}i(n) \\ e(n) &= \boldsymbol{c}^\top \boldsymbol{x}_n + di(n) \end{aligned} \tag{2.6}$$

Where $i(n)$ and $e(n)$ are the (scalar) filter input and output signals, respectively. The transition matrix $\boldsymbol{A}$ is an (N, N) matrix and $\boldsymbol{b}$, $\boldsymbol{c}$ and the state vector $\boldsymbol{x}_n$ are column vectors of length N. The parameter d determines obviously the impulse response value at time zero, i.e., $h(0) = d$. The transfer function $H(z)$ of the filter can be found from Eq. (2.6) using the z-transform and linear algebra:

$$\begin{aligned} z\boldsymbol{X}(z) &= \boldsymbol{A}\boldsymbol{X}(z) + \boldsymbol{b}\boldsymbol{I}(z) \;\Rightarrow \\ \boldsymbol{X}(z) &= -(\boldsymbol{A} - \boldsymbol{E}z)^{-1}\boldsymbol{b}I(z) \\ E(z) &= \boldsymbol{c}^\top \boldsymbol{X}(z) + dI(z) \\ &= H(z)I(z) = (d - \boldsymbol{c}^\top(\boldsymbol{A} - \boldsymbol{E}z)^{-1}\boldsymbol{b})I(z) \end{aligned} \tag{2.7}$$

Where $\boldsymbol{X}(z)$, $E(z)$ and $I(z)$ are the z-transforms of the state vector sequence $\boldsymbol{x}_n$ and the filter output and input, respectively. The matrix $\boldsymbol{E}$ is the (N, N) unit matrix.

Eq. (2.7) shows that the poles of $H(z)$ are the z-values for which Eq. (2.7) becomes singular, i.e., the poles of $H(z)$ are equal to the roots of the characteristic polynomial of $\boldsymbol{A}$. In other words, the eigenvalues of $\boldsymbol{A}$ are poles of $H(z)$. The zeros of $H(z)$ are generally determined by $\boldsymbol{b}$, $\boldsymbol{c}$ and d.

For a given state-space description of a filter, the transfer function will be uniquely determined by Eq. (2.7); however, a multitude of state-space descriptions result in a given transfer function. A well known result from linear system theory is that two filters with the same transfer function generate identical output for identical input when proper initial conditions are used, i.e., filters with identical transfer function $H(z)$ are equivalent, cf. Definition 2.1.

Using Eq. (2.6), an Nth order Σ-Δ modulator corresponding to Fig. 2.1 can be characterized in state-space matrix-vector notation:

$$\begin{aligned} \boldsymbol{x}_{n+1} &= \boldsymbol{A}\boldsymbol{x}_n + \boldsymbol{b}(i(n) - \mathrm{sgn}(\boldsymbol{c}^\top \boldsymbol{x}_n)),\; \boldsymbol{x}_n \in \mathbb{R}^N \\ y(n) &= \mathrm{sgn}(\boldsymbol{c}^\top \boldsymbol{x}_n) \end{aligned} \tag{2.8}$$

where $y(n)$ is the modulator output sequence and $\mathrm{sgn}(\cdot)$ denotes the signum function, i.e., $\mathrm{sgn}(x) \triangleq 1$ for positive x and $\mathrm{sgn}(x) \triangleq -1$, otherwise.

Using the system description of Eq. (2.1), this general Nth order Σ-Δ modulator can be characterized by:

$$\begin{aligned} \boldsymbol{x}_{n+1} &= \mathcal{F}(\boldsymbol{x}_n, i(n)) = \boldsymbol{A}\boldsymbol{x}_n + \boldsymbol{b}(i(n) - \mathcal{O}(\boldsymbol{x}_n)) \\ y(n) &= \mathcal{O}(\boldsymbol{x}_n) = \operatorname{sgn}(\boldsymbol{c}^\top \boldsymbol{x}_n) \end{aligned} \tag{2.9}$$

The feedback loop of Fig. 2.1 including the signum function cannot be realized if the filter is delay free, i.e., the filter impulse response must have the property $h(0) = 0$, where $h(n)$ is the impulse reponse of $H(z)$. This is the reason why the parameter d from Eq. (2.6) is omitted in Eq. (2.8).

For most Σ-Δ modulator topologies, the feedback filter transfer function $H(z)$ is not immediately observable. In these cases, the modulator has to be described in state-space at first before $H(z)$ can be derived using Eq. (2.7).

Example 2.1 Consider the second order modulator of Fig. 2.2 with three parameters, b_1, b_2 and a. The feedback filter is composed of a cascade of discrete-time integrators with distributed input. This approach is extensively used for construction of high-order modulators [1, 55, 17] and the resulting coefficient sensitivity is low compared to other structures [49]. The two variables $x_1(n)$ and $x_2(n)$ are chosen as state-coordinates, i.e., the state-vector is $\boldsymbol{x}_n = \begin{bmatrix} x_1(n) & x_2(n) \end{bmatrix}^\top$. This choice results in the following state-space description:

$$\begin{aligned} \boldsymbol{x}_{n+1} &= \boldsymbol{A}\boldsymbol{x}_n + \boldsymbol{b}\left(i(n) - \operatorname{sgn}\left(\boldsymbol{c}^\top \boldsymbol{x}_n\right)\right) \\ \boldsymbol{A} &= \begin{bmatrix} 1 & 1 \\ -a & 1 \end{bmatrix} \\ \boldsymbol{b} &= \begin{bmatrix} b_1 & b_2 \end{bmatrix}^\top \\ \boldsymbol{c} &= \begin{bmatrix} 1 & 0 \end{bmatrix}^\top \end{aligned} \tag{2.10}$$

The feedback transfer function $H(z)$ can now be derived using Eq. (2.7):

$$\begin{aligned} H(z) &= -\boldsymbol{c}^\top (\boldsymbol{A} - \boldsymbol{E}z)^{-1} \boldsymbol{b} \\ &= \begin{bmatrix} 1 & 0 \end{bmatrix} \frac{-1}{z^2 - 2z + 1 + a} \begin{bmatrix} 1-z & -1 \\ a & 1-z \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \\ &= \frac{b_1 z^{-1} + (b_2 - b_1) z^{-2}}{1 - 2z^{-1} + (1+a) z^{-2}} \end{aligned} \tag{2.11}$$

Notice that b_1 and b_2 determine the zero of $H(z)$ and that $H(z)$ has the required unit sample delay. The poles of $H(z)$ are determined alone by the internal feedback factor a □

2.5 Σ-Δ Modulator Equivalences

As stated in Sec. 2.4, two linear dynamical systems are equivalent if they have identical transfer functions. Consequently, two Σ-Δ modulators are equivalent if the corresponding feedback filters have identical transfer functions. Furthermore, Σ-Δ modulators are

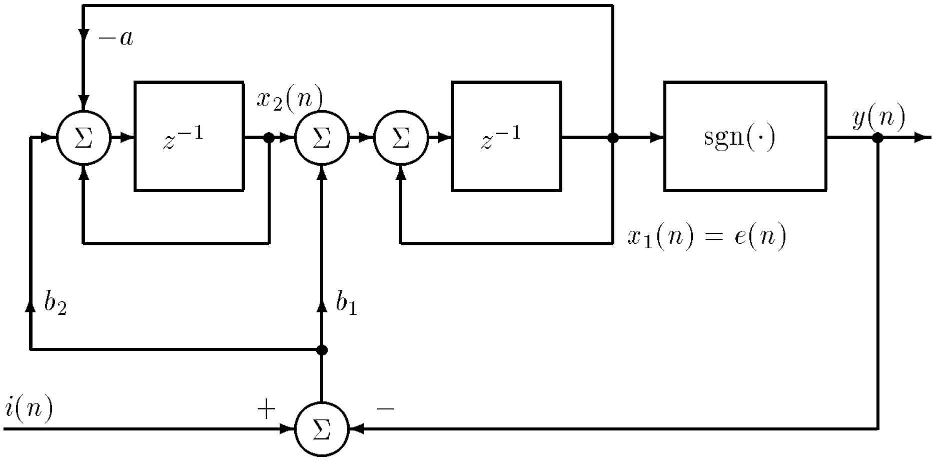
$-a$
Σ
z^{-1}
$x_2(n)$
Σ
Σ
z^{-1}
sgn(·)
$y(n)$
$x_1(n) = e(n)$
b_2
b_1
$i(n)$
+
−
Σ

Figure 2.2: Second order Σ-Δ modulator example.

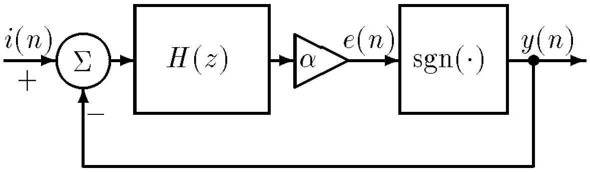
$i(n)$
+
Σ
−
$H(z)$
α
$e(n)$
sgn(·)
$y(n)$

Figure 2.3: Σ-Δ modulator with feedback filter scaling.

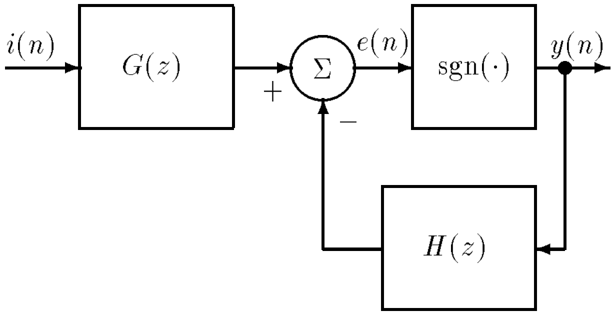

Figure 2.4: Generic Σ-Δ modulator.

invariant to scaling of the feedback filter $H(z)$. This fact can be realized from Fig. 2.3 where the filter output is scaled by the factor α before the signum function. This system produces obviously the same output for any positive value of α due the fact that the signum function is invariant to argument scaling. This justifies the following theorem:

Theorem 2.1 *Two Σ-Δ modulators cf. Fig. 2.1 are equivalent if and only if the corresponding feedback filters $H_1(z)$ and $H_2(z)$ have the property:*

$$H_2(z) = \alpha H_1(z) \ , \ \alpha > 0 \tag{2.12}$$

The theorem states that the poles and zeros of $H(z)$ defines two equivalence classes of modulators giving the same output for the same input, i.e., one class for negative and one class for positive scalings of $H(z)$.

> **Example 2.2** Consider the modulator in Fig. 2.2 discussed in Example 2.1. Since the modulator is invariant to scalings of $H(z)$, it is observed from Eq. (2.11) that the parameter b_1 can be fixed to unity without loss of generality, i.e., the real valued zero of $H(z)$ can be determined alone by the remaining parameter b_2. Scaling of both b_1 and b_2 will only affect the internal signal levels within the integrators. The actual scaling can be decided when the circuit is to be implemented □

So far, the modulators considered have been restricted to a certain class as defined by Fig. 2.1 which does not comprise every known modulator topology. Fig. 2.4 shows a more general generic Σ-Δ modulator which is characterized by two transfer functions, i.e., the feedback filter $H(z)$ between the quantizer output and input and the input transfer function $G(z)$ between the input and the quantizer input. The generic modulator of Fig. 2.4 can be rearranged into an equivalent system consisting of a usual modulator cf. Fig. 2.1 preceded by a linear prefilter as shown in Fig. 2.5. Consequently, generic

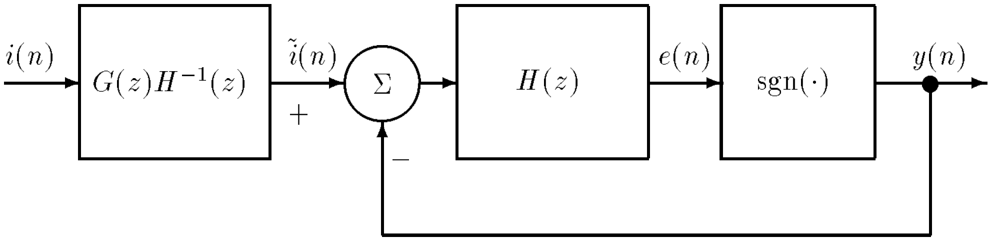

Figure 2.5: Generic Σ-Δ modulator implemented as a usual modulator with a prefilter.

modulators cf. Fig. 2.4 with the same feedback filter $H(z)$ are equivalent when the input signal is preprocessed by a proper linear filter. Since linear filtering is a very well known operation, the study of the nonlinear dynamics of Σ-Δ modulators can be restricted to the sub-class cf. Fig. 2.1. Notice that generic modulators are generally not invariant to scalings of $H(z)$ since the the prefilter $G(z)H^{-1}(z)$ scales inversely with $H(z)$.

Example 2.3 Consider the second order modulator in Fig. 2.6. This modulator topology is a modification of the modulator shown in Fig. 2.2 and is commonly know as the *multiple feedback* modulator [55]. This example will demonstrate that the multiple feedback modulator is comprised by the generic model, cf. Fig. 2.4.

Obviously, the modulators of Fig. 2.2 and Fig. 2.6 share the same feedback filter $H(z)$, cf. Eq. (2.11):

$$H(z) = \frac{b_1 z^{-1} + (b_2 - b_1) z^{-2}}{1 - 2z^{-1} + (1+a) z^{-2}} \tag{2.13}$$

The input transfer function, $G(z)$, from the input node to the quantizer input node is found:

$$\begin{aligned} G(z) &= -\boldsymbol{c}^{\mathsf{T}} (\boldsymbol{A} - \boldsymbol{E} z)^{-1} \begin{bmatrix} 0 & 1 \end{bmatrix}^{\mathsf{T}} \\ &= \frac{z^{-2}}{1 - 2z^{-1} + (1+a) z^{-2}} \end{aligned} \tag{2.14}$$

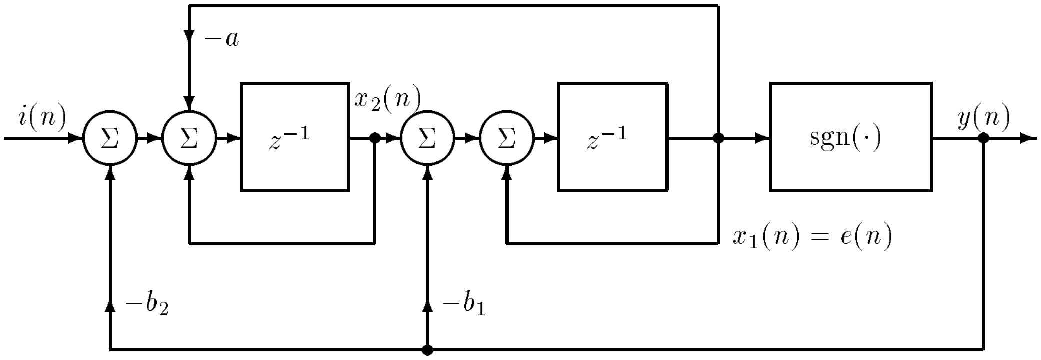

Figure 2.6: Second-order multiple feedback Σ-Δ modulator.

Consequently, the multiple feedback modulator, cf. Fig. 2.6 is equivalent to the modulator in Fig. 2.2 preceded by the prefilter:

$$G(z)H^{-1}(z) = \frac{z^{-2}}{b_1 z^{-1} + (b_2 - b_1)z^{-2}} = \frac{1/b_1 z^{-1}}{1 + (b_2/b_1 - 1)z^{-1}} \tag{2.15}$$

The prefilter is a first order recursive filter with a one sample delay. Notice also that scaling of both b_1 and b_2 just scales the gain of the prefilter accordingly.

The multiple feedback modulator is often used in a transposed form called the *feedforward* (FF) modulator [55, 17]. The FF modulator is, like the topology in Fig. 2.2, equivalent to a usual Σ-Δ modulator, i.e., $G(z) = H(z)$ for these topologies □

The subject of equivalences among different modulator topologies has also been discussed in [63, 66, 68].

2.6 The Noninvertible Region

Real physical dynamical systems (e.g., systems ruled by Newton's laws of motion) are reversible, i.e., the axis of time can be reversed. Even when such continuos-time systems a modeled in discrete-time, the system maps are generally invertible and it is possible to go back and forth in time. This section will show that the signum function or one-bit quantizer of a Σ-Δ modulator causes the system map to be noninvertible in a certain region in the state-space.

The nonlinear map $\mathcal{F}$ of a Σ-Δ modulator with constant input i can be characterized by two linear maps defined on two distinct domains:

$$\mathcal{F}(\boldsymbol{x}_n) = \begin{cases} \mathcal{F}_+(\boldsymbol{x}_n) & , \boldsymbol{x}_n \in S_+ \\ \mathcal{F}_-(\boldsymbol{x}_n) & , \boldsymbol{x}_n \in S_- \end{cases}$$

$$\begin{aligned} S_+ &= \{\boldsymbol{x} \in \mathbb{R}^N | \mathcal{O}(\boldsymbol{x}) = 1\} \\ &= \{\boldsymbol{x} \in \mathbb{R}^N | \boldsymbol{c}^\top \boldsymbol{x} > 0\} \\ S_- &= \{\boldsymbol{x} \in \mathbb{R}^N | \mathcal{O}(\boldsymbol{x}) = -1\} \\ &= \{\boldsymbol{x} \in \mathbb{R}^N | \boldsymbol{c}^\top \boldsymbol{x} \leq 0\} \\ \mathcal{F}_+(\boldsymbol{x}_n) &= \boldsymbol{A}\boldsymbol{x}_n + \boldsymbol{b}(i-1) \\ \mathcal{F}_-(\boldsymbol{x}_n) &= \boldsymbol{A}\boldsymbol{x}_n + \boldsymbol{b}(i+1) \end{aligned} \tag{2.16}$$

The piecewise linearity enables the existence of a region U in which the system map is noninvertible, i.e., U is the set of elements $\boldsymbol{y}$ which gives two solutions to the equation $\mathcal{F}(\boldsymbol{x}) = \boldsymbol{y}$. Since both of the linear maps $\mathcal{F}_+$ and $\mathcal{F}_-$ are generally invertible, the set U can be derived:

$$\begin{aligned} U &= \mathcal{F}_+(S_+) \bigcap \mathcal{F}_-(S_-) \\ &= \{\boldsymbol{y} \in \mathbb{R}^N | \mathcal{F}_+^{-1}(\boldsymbol{y}) \in S_+ \wedge \mathcal{F}_-^{-1}(\boldsymbol{y}) \in S_-\} \\ &= \{\boldsymbol{y} \in \mathbb{R}^N | \boldsymbol{c}^\top \boldsymbol{A}^{-1}\boldsymbol{b}(i-1) < \boldsymbol{c}^\top \boldsymbol{A}^{-1}\boldsymbol{y} < \boldsymbol{c}^\top \boldsymbol{A}^{-1}\boldsymbol{b}(i+1)\} \end{aligned} \tag{2.17}$$

Both the sets $V_+ = \mathcal{F}_+^{-1}(U)$ and $V_- = \mathcal{F}_-^{-1}(U)$ are mapped on U by $\mathcal{F}$, i.e., the set $V = V_+ \bigcup V_-$ is the preimage of the noninvertible region U. The set V has a remarkably simple structure:

$$\begin{aligned} V_+ &= \{\boldsymbol{x} \in \mathbb{R}^N | \boldsymbol{c}^\top \boldsymbol{A}^{-1}\boldsymbol{b} < \boldsymbol{c}^\top \boldsymbol{A}^{-1}\left(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}(i-1)\right) < \boldsymbol{c}^\top \boldsymbol{A}^{-1}\boldsymbol{b}(i-1)\} \\ &= \{\boldsymbol{x} \in \mathbb{R}^N | 0 < \boldsymbol{c}^\top \boldsymbol{x} < 2\boldsymbol{c}^\top \boldsymbol{A}^{-1}\boldsymbol{b}\} \\ V_- &= \{\boldsymbol{x} \in \mathbb{R}^N | -2\boldsymbol{c}^\top \boldsymbol{A}^{-1}\boldsymbol{b} < \boldsymbol{c}^\top \boldsymbol{x} < 0\} \\ V &= \{\boldsymbol{x} \in \mathbb{R}^N | \, |\boldsymbol{c}^\top \boldsymbol{x}| < 2\boldsymbol{c}^\top \boldsymbol{A}^{-1}\boldsymbol{b}\} \end{aligned} \tag{2.18}$$

Recall from Sec. 2.4 that the quantizer input e is equal to $\boldsymbol{c}^\top \boldsymbol{x}$ and that the transfer function of the feedback filter is given by $H(z) = -\boldsymbol{c}^\top(\boldsymbol{A} - \boldsymbol{E}z)^{-1}\boldsymbol{b}$. Consequently, the elements in V satisfy the very simple relation:

$$|e| < -2H(z=0) \tag{2.19}$$

Notice that this relation is very general, i.e., V is independent of the modulator input signal and V is defined only by the observable quantizer input e and by the feedback filter $H(z)$. It can be concluded that if a modulator operates exclusively in the noninvertible region U, it must operate exclusively in V as well, i.e., the quantizer input magnitude will be bounded by $-2H(z=0)$. Notice furthermore that U is nonempty for $H(z=0) < 0$. It will be demonstrated later that the noninvertible region plays an important rôle for the stability of a modulator.

2.7 Local Stability Analysis and Limit Cycle Identification

This section focuses on the structure and stability of the limit cycles of Σ-Δ modulators. The fact that a Σ-Δ modulator is a linear filter with nonlinear and binary feedback simplifies both the identification and stability analysis of possible limit cycles.

According to Eq. (2.16), the functional (Jacobian) matrix of a modulator map is simply given as:

$$\mathrm{D}\mathcal{F}(\boldsymbol{x}) = \boldsymbol{A} \, , \; \boldsymbol{c}^\top \boldsymbol{x} \neq 0 \tag{2.20}$$

This means that the eigenvalues of the transition matrix $\boldsymbol{A}$ determines the stability of fixed and periodic points. Recall from Sec. 2.4 that the eigenvalues of $\boldsymbol{A}$ are equal to the poles of the modulator feedback filter $H(z)$. Consequently, modulators with one or more poles outside the unit circle have unstable fixed points and limit cycles.

The limit cycles of a Σ-Δ modulator with known (periodic or constant) input can be identified by opening the modulator loop. This is again a consequence of the feedback path being a linear filter: Consider the modulator of Fig. 2.1 where the loop is opened and the quantizer is replaced by a signal generator which produces a periodic binary signal $y(n)$ with the two discrete amplitude values 1 and -1. The output $e(n)$ of the linear feedback filter $H(z)$ (which formerly was the quantizer input) will then have a periodic steady-state solution and if $\mathrm{sgn}(e(n)) = y(n)$ for all n, a limit cycle exists with $y(n)$ as periodic output code. This means that there is a close relationship between the periodic output code sequence and the limit cycle in state-space.

The periodic steady-state quantizer input signal $e(n)$ can be found as the solution to a set of linear equations for any period k code sequence $y(n)$ and there will thus exist a linear mapping between $y(n)$ and $e(n)$ which depends on $H(z)$ and on the modulator input.

Let $H(z)$ be given as a rational transfer function $C(z)/D(z)$ where $C(z) = c_1 z^{-1} + c_2 z^{-2} \cdots c_N z^{-N}$ and $D(z) = 1 + d_1 z^{-1} + d_2 z^{-2} \cdots d_N z^{-N}$. For a given input signal $i(n)$ and output signal $y(n)$, the loop filter input is given by $v(n) = i(n) - y(n)$. In the z-domain, the loop filter output $e(n)$ is given by:

$$E(z) = \frac{C(z)}{D(z)} V(z) \tag{2.21}$$

or equivalently:

$$D(z)E(z) = C(z)V(z) \tag{2.22}$$

Let $v(n)$ be periodic with period k. The steady-state $e(n)$ is then also periodic with period k and $e(n)$ can be found from the following linear equation:

$$\begin{bmatrix} 1 & 0 & \cdots & d_N & \cdots & d_2 & d_1 \\ d_1 & 1 & 0 & \cdots & d_N & \cdots & d_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & d_N & \cdots & d_3 & d_2 & 1 \end{bmatrix} \begin{bmatrix} e(0) \\ e(1) \\ \vdots \\ e(k-1) \end{bmatrix} = \begin{bmatrix} 0 & \cdots & c_N & \cdots & c_2 & c_1 \\ c_1 & 0 & \cdots & c_N & \cdots & c_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & c_N & \cdots & c_2 & c_1 & 0 \end{bmatrix} \begin{bmatrix} v(0) \\ v(1) \\ \vdots \\ v(k-1) \end{bmatrix} \tag{2.23}$$

Or using the symbolic notation:

$$\boldsymbol{D}_k \boldsymbol{e} = \boldsymbol{C}_k \boldsymbol{v} \tag{2.24}$$

The solution is then given by:

$$\boldsymbol{e} = \boldsymbol{D}_k^{-1} \boldsymbol{C}_k \boldsymbol{v} \tag{2.25}$$

The existence of the periodic sequence $y(n)$ as a limit cycle can then be tested using the condition $y(n) = \mathrm{sgn}(e(n))$ for all n.

For certain $H(z)$ pole configurations the set of equations (Eq. (2.23)) may become singular, e.g., if $H(z)$ has a pole a $z = 1$ the filter has infinite dc-gain and the mean value of the filter input must be zero. In this case the mean value of the steady-state filter output is undetermined and may be considered as a free parameter, i.e., there are infinitely many steady-state solutions for $e(n)$. The code sequence will then exist as a limit cycle for the mean value of the filter output belonging to a whole interval where the condition $\mathrm{sgn}(e(n)) = y(n)$ is satisfied for all n. Generally, the steady-state filter output solution is nonunique when one or more $H(z)$ poles are on the unit circle at the locations $\exp(j\pi l/k)$ which correspond to the possible discrete frequencies of the spectra of signals with period k [25]. In these cases, there will either be none or infinitely many solutions.

In order to simplify notation, code sequences will be written as binary numbers, i.e., sequences of the symbols '1' and '0' corresponding to quantizer output amplitudes 1 and -1, respectively. A periodic repetition of a code sequence is indicated by overlining, e.g., the notation $\overline{1100}$ means the (infinite) code sequence 110011001100.... This means that the set of 2^k k-digit binary numbers is equivalent to the set of output sequences with period k and that each possible periodic point can be identified with a binary number. However, normally only a fraction of the 2^k period k sequences will exist as limit cycles.

Code sequence segments can also be described by counting the maximum number of consecutive '1' and '0' symbols. The notation (2,2) means a code segment of two '1' symbols followed by two '0' symbols and the notation $(\overline{2,2})$ is an alternative notation for the periodic sequence $\overline{1100}$.

Example 2.4 Some periodic code sequences with period 4 will be tested for existence as limit cycles for the second-order modulator of Fig. 2.2. The feedback filter $H(z)$ for this modulator is given as (cf. Eq. (2.11)):

$$H(z) = \frac{b_1 z^{-1} + (b_2 - b_1) z^{-2}}{1 - 2z^{-1} + (1+a) z^{-2}} \tag{2.26}$$

The matrix $\boldsymbol{D}_4$ is:

$$\boldsymbol{D}_4 = \begin{bmatrix} 1 & 0 & 1+a & -2 \\ -2 & 1 & 0 & 1+a \\ 1+a & -2 & 1 & 0 \\ 0 & 1+a & -2 & 1 \end{bmatrix} \tag{2.27}$$

Note that this matrix becomes singular for $a = 0$ (the sum of the rows and columns is zero). In this case there is a double pole at $z = 1$.

The matrix $\boldsymbol{C}_4$ is:

$$\boldsymbol{C}_4 = \begin{bmatrix} 0 & 0 & b_2 - b_1 & b_1 \\ b_1 & 0 & 0 & b_2 - b_1 \\ b_2 - b_1 & b_1 & 0 & 0 \\ 0 & b_2 - b_1 & b_1 & 0 \end{bmatrix} \tag{2.28}$$

Let $b_1 = 1$, $b_2 = 0.5$ and $a = 0.1$. For zero modulator input, the periodic code sequence $\overline{10} = \overline{1010}$ gives $\boldsymbol{v} = [\ -1 \quad 1 \quad -1 \quad 1\ ]^{\mathsf{T}}$. Using Eq. (2.25), the steady-state filter output is $\boldsymbol{e} = [\ 0.366 \quad -0.366 \quad 0.366 \quad -0.366\ ]^{\mathsf{T}}$. The $\overline{10}$ limit cycle exists since $\mathrm{sgn}(e(n)) = y(n)$. This limit cycle is very persistent for most modulators and causes most modulators to have a tone at half the sample rate for zero input (see Ch. 8).

The periodic code sequence $\overline{1000}$ results in $\boldsymbol{e} = [\ 3.19 \quad 2.09 \quad 2.17 \quad 2.54\ ]^{\mathsf{T}}$; hence, this limit cycle does not exist for zero input. However, for a suitable negative constant modulator input, this limit cycle will exist: the dc-gain of $H(z)$ is 5; hence, for a constant dc-input in the interval $[-3.19/5, -2.54/5]$ the $\overline{1000}$ limit cycle exists □

2.8 Summary

This chapter has introduced Σ-Δ modulators as general discrete-time nonlinear dynamical systems. It was shown that the nonlinear dynamics of a modulator is only determined by the normalized feedback filter transfer function. Different modulator topologies sharing the same feedback filter are all equivalent when proper linear input filters are applied. Since linear filters are well known, it is only necessary to focus on the feedback filter inside the nonlinear loop around the quantizer.

The piecewise linearity of these systems simplifies the local stability analysis: the system behaves like a linear system in two disjoint domains and the poles of the feedback filters determine the local stability of fixed points and limit cycles. It was also demonstrated that the maps of Σ-Δ modulators have a noninvertible region where the map is two-to-one. Furthermore, it was shown that this region can be characterized generally from the quantizer input and the feedback filter transfer function evaluated at $z = 0$.

Finally, it was shown how the limit cycles of a modulator can be found by opening the loop. Furthermore, the output code sequences of a limit cycle is uniquely characterizing the corresponding orbit for most systems. Consequently, there is a one-to-one correspondence between limit cycles and the binary numbers.

Chapter 3

The Stability of Chaotic Σ-Δ Modulators

3.1 Introduction

The term ‘stability’ is often mentioned in connection with Σ-Δ modulators. Since a large number of stability definitions exist, a more explicit formulation of the stability of Σ-Δ modulators must be given. The practical problem is that a modulator operating with a given input suddenly might ‘explode’, i.e., the quantizer input starts oscillating with a rapidly increasing amplitude. These oscillations may become unbounded for the purely mathematical modulator described in Ch. 2; however, for real modulators, non-linearities such as arithmetic overflow and circuit saturation will limit the amplitude.

The onset of uncontrolled oscillations causes the signal encoding properties of the modulator to virtually disappear and the resulting signal-to-noise ratio decreases dramatically. Obviously, combinations of input signals and feedback filter parameters which cause instability should be avoided. The goal of the Σ-Δ modulator stability analysis is to answer the question: is a given modulator stable for a reasonably large set of initial conditions and with a given class of input signals?

This chapter focuses on the stability of chaotic modulators. The question in this case is: Does the system have a stable attractor and what is its basin of attraction?

It will appear from the numerous examples in this chapter that it is extremely difficult to derive an easily applied, general and not too conservative stability criterion.

Due to the usually high oversampling ratio, the input signals for real modulators will normally change slowly. In order to simplify the stability analysis, the input signal in the following is restricted to the class of constant input signals, i.e., the modulators can be described as an autonomous dynamical systems.

3.2 Chaotic Σ-Δ Modulators

When one or more poles of $H(z)$ are outside the unit circle, every fixed or periodic points is unstable; however, this does not necessarily mean that the orbits of the system become unbounded with time as it is the case for linear systems. Due to the possible existence of a non-invertible region U (see Sec. 2.6), the orbit may stay bounded. The reason for this is that the sets V_+ and V_- are both mapped on top of each other on the non-invertible region U. This folding mechanism seems to be necessary in order to bound the orbit in steady state. The existence of a folding mechanism in combination with state-space stretching, i.e., one or more eigenvalues are outside the unit circle, indicates the presence of chaos, i.e., the contracting effect of the folding mechanism counterbalances the state-space divergence. This motivates following assumption:

Assumption 3.1 *A Σ-Δ modulator with at least one $H(z)$ pole outside the unit circle and with a non-empty non-invertible region is assumed to be chaotic, i.e., a bounded limit set exists on which the modulator is chaotic. Furthermore, the intersection between the limit set and the non-invertible region is non-empty, i.e., the non-invertible set is visited repeatedly.*

3.3 Boundary Crisis

As stated in Section 3.1, the question of stability of chaotic modulators is: When is the chaotic limit set attracting and what is the corresponding basin of attraction? Generally, it is a very difficult task to answer this question, since the stability obviously depends on global properties rather than local stability analysis. To illustrate this: modulators can be made stable despite the fact that local state-space perturbations grow in all directions, i.e., all $H(z)$-poles are outside the unit circle. When a chaotic modulator is stable, a basin of attraction B_L with a non-zero measure exist, i.e., B_L is the set of initial conditions which are attracted asymptotically to the chaotic limit set. It is concluded from simulations that a chaotic modulator has at most one stable limit set which must chaotic since every limit cycle is unstable. Consequently, the initial conditions outside B_L will generally result in unbounded orbits; however, initial conditions which are exactly on other unstable limit points, e.g., points on limit cycles, will give bounded orbits which are not attracted to the chaotic limit. Therefore, unstable limit points not belonging to the chaotic attractor must either be on the boundary of B_L, ∂B_L or belong to the exterior of B_L. However, some of the limit points on the boundary of B_L can be points of accumulation of B_L, i.e., any sufficiently small non-zero perturbation gives an orbit which is attracted to the chaotic limit set. For limit points belonging to the exterior of B_L, any sufficiently small perturbation causes the orbit to become unbounded. Furthermore, ∂B_L must be invariant to the system map, i.e., every point on ∂B_L is mapped to another point on ∂B_L.

Obviously, an attracting chaotic limit set L must be contained in its corresponding basin of attraction, i.e., $L \subset B_L$. When the modulator input or filter parameters is changed such that the chaotic modulator becomes unstable, the basin of attraction B_L disappears. A way to interpret the onset of instability is to imagine that L 'grows' out of B_L or in general that L collides with the boundary of B_L. This type of bifurcation is often called a *boundary crisis* [22]. It is characteristic to this type of bifurcation that B_L suddenly disappears at the onset of instability rather than it gradually shrinks towards zero measure. Furthermore, the limit set does not 'jump' in size just before loosing stability. This explains why modulators may 'blow up' suddenly without any preceding warning.

Notice that after the onset of instability, a chaotic but unstable limit set survives. This means that there still exist initial conditions which give bounded orbits, namely the limit points; however, even small perturbations cause the modulator to 'blow up', i.e., the set of initial conditions giving bounded orbits has zero measure. Hence, the stability concept of a chaotic limit set is in good agreement with the practical modulator stability definition as defined in Section 3.1. In Section 3.5, the so-called escape rate will be defined which quantifies the instability of a repelling chaotic limit set.

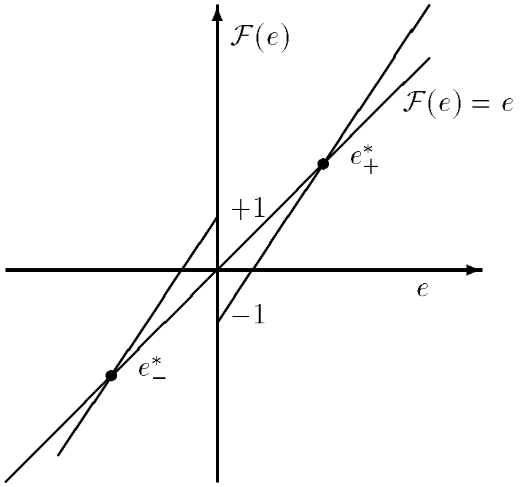

Figure 3.1: First order map $\mathcal{F}(e)$ according to Eq. (3.1) for zero input and $a > 1$.

Example 3.1 Consider the general first order Σ-Δ modulator with constant input given by the map:

$$e(n+1) = \mathcal{F}(e(n)) = ae(n) + i - \operatorname{sgn}(e(n)) \tag{3.1}$$

where i is the constant input and a is a parameter.

The map $\mathcal{F}$ is schematically shown in Fig. 3.1 for $a > 1$ and $i = 0$.

The corresponding feedback filter transfer function is:

$$H(z) = \frac{z^{-1}}{1 - az^{-1}} \tag{3.2}$$

The feedback filter has a pole at a. When a is less than unity, the map corresponds to a first order modulator with a leaky integrator [16, 45]. For $a = 1$, the system is equivalent to the well known circle map (rotation of the circle) [68] and this system is quasiperiodic for irrational constant input [15] . For $a > 1$ the modulator is generally chaotic [15], i.e., the pole at a is outside the unit circle.

Eq. (3.1) shows that for $a > 1$, the map has to fixed points given by:

$$\begin{aligned} e_+^* &= \frac{1-i}{a-1} \\ e_-^* &= \frac{-1-i}{a-1} \end{aligned} \tag{3.3}$$

The noninvertible set is the interval $U =]-1+i, 1+i[$ and the preimage of U is the interval $V =]2H(0), -2H(0)[=]\frac{-2}{a}, \frac{2}{a}[$.

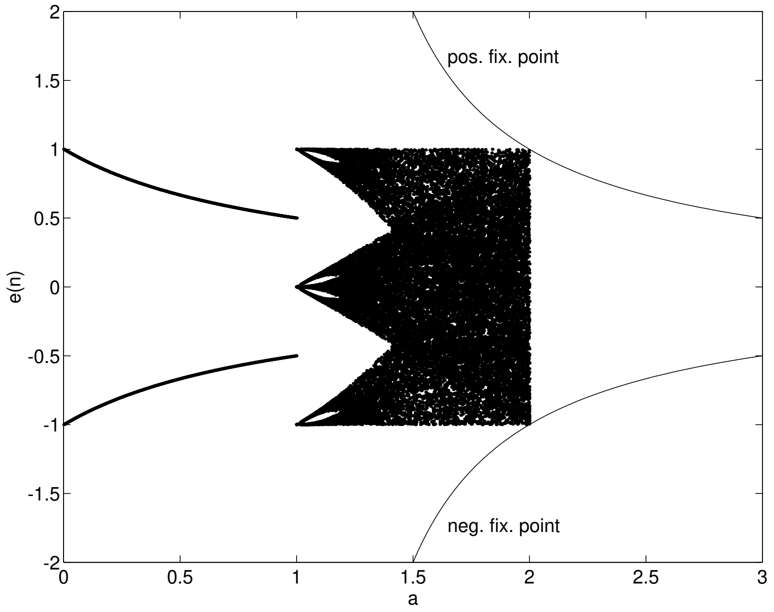

Figure 3.2: Bifurcation diagram of the first order modulator cf. Eq. (3.1) with zero input. The two hyperbolas are the loci of the two unstable fixed points for $a > 1$, cf. Eq. (3.3).

The unstable fixed points separate points which iterate to infinity from points which iterate towards the signum function threshold at $e = 0$. Points in a neighborhood of $e = 0$ are either mapped into a neighborhood of $e = 1 + i$ or $e = -1 + i$, i.e., the end-points of U. Consequently, the modulator is assumed to be stable if U is fully inside the interval between the unstable fixed points which is fulfilled when:

$$|i| < \frac{2}{a} - 1 \qquad (3.4)$$

Hence, the stable input range decreases as a is increased.

When this condition is satisfied, the basin of attraction must be the interval between the unstable fixed points, i.e, $B_L =]e^*_-, e^*_+[$, and the chaotic limit set L must be contained in U, cf. Ass. 3.1.

Fig. 3.2 shows a bifurcation diagram for the first order modulator with zero input (see also [15, 59]). The modulator was simulated with a slowly increasing a-parameter and the obtained quantizer input $e(n)$ is plotted versus a. For $a < 1$, the stable $\overline{10}$ limit cycle is the only existing limit cycle (at most one limit cycle can exist at a time for fixed i and a for $a < 1$, cf. [15]). For $a > 1$, a chaotic limit set emerges including some more complex limit cycles like $\overline{1100}$. However, the now unstable $\overline{10}$ limit cycle is not included in the chaotic limit set. Notice that the chaotic limit set becomes unstable at $a = 2$ due to a collision with the unstable fixed points which form the boundary of B_L □

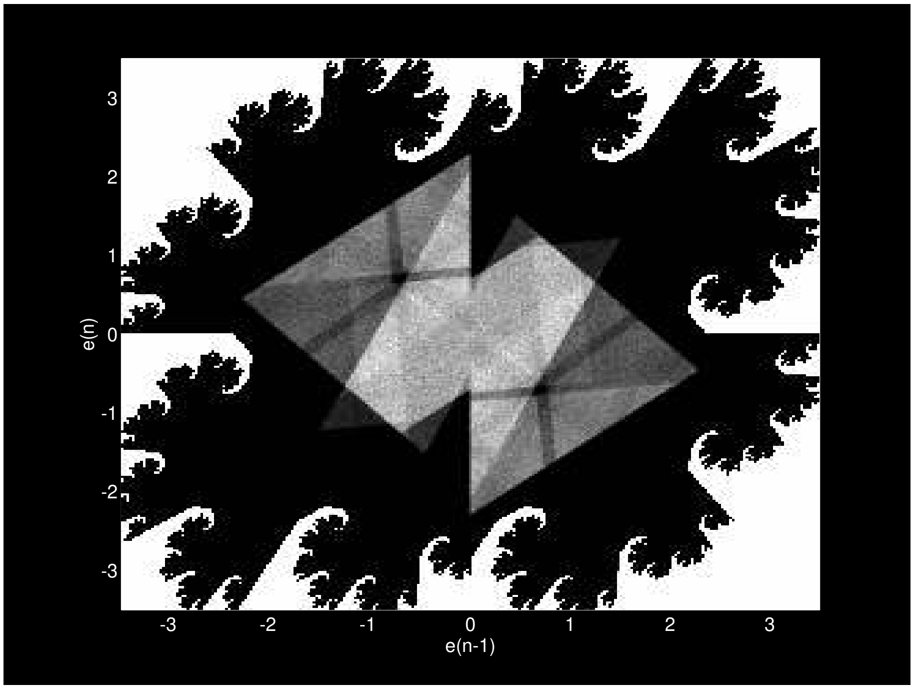

Figure 3.3: Basin of attraction B_L and limit set L for the modulator cf. Eq. (3.5) with $h_1 = 0.8$ and $h_2 = -1.5$.

3.4 Finding the Basin of Attraction

This section will focus on the onset of instability for chaotic modulators from a more practical point of view. Especially the often very complex structure of the chaotic limit set and the basin of attraction will be studied by simulations. The examples are based on second order modulators in order to enable graphical illustrations.

The basin of attraction, B_L of a modulator can in general not be determined analytically; hence, B_L must be estimated numerically. For all practical purposes, B_L is the set of initial conditions which give bounded orbits. The state-space must thus be sampled in a large number of grid points and the points giving orbits which stay bounded by a certain bound up to a certain maximum number of iterations, n_{max}, are marked as members of B_L. A stable chaotic limit set L can be approximated simply as the union of all points on an orbit starting within B_L; however, the initial transient of the orbit before steady state behavior is reached should be omitted.

Example 3.2 Fig. 3.3 and 3.4 shows B_L for two modulators with feedback filter

$$H(z) = \frac{h_1 z^{-1} + h_2 z^{-2}}{1 - h_1 z^{-1} - h_2 z^{-2}} \tag{3.5}$$

where h_1 is 0.8 and h_2 is -1.5 and -1.65.

This class of modulators has been investigated empirically as well as analytically in [60].

For each point in Fig. 3.3 and Fig. 3.4, up to 500 iterations were performed and points giving an quantizer input signal $|e(n)| < 10$ are plotted in black and

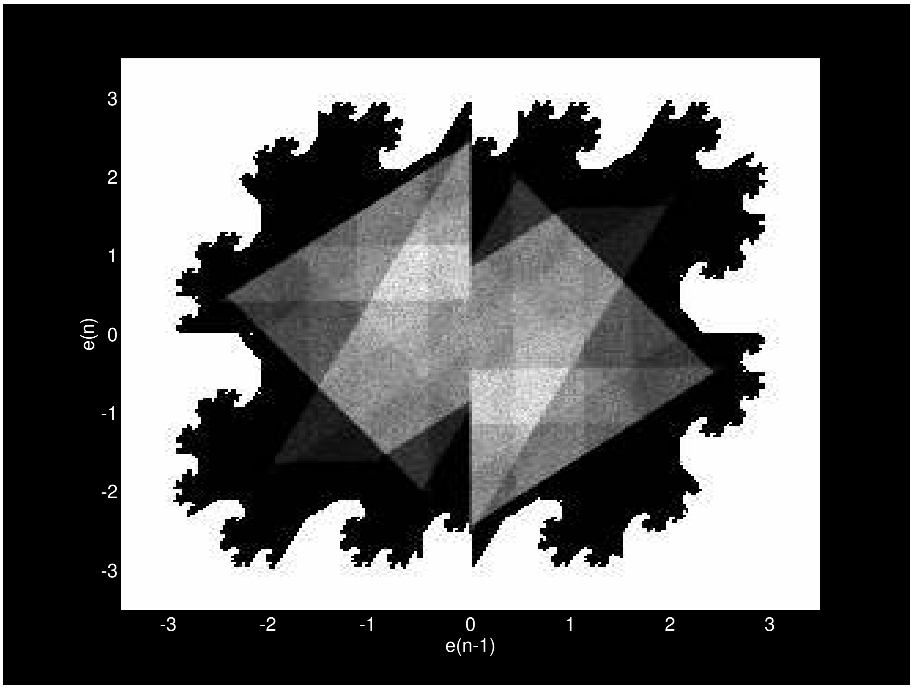

Figure 3.4: Basin of attraction B_L and limit set L for the modulator cf. Eq. (3.5) with $h_1 = 0.8$ and $h_2 = -1.65$.

the rest are white. The state-vector used is $\boldsymbol{x}_n = [e(n-1), e(n)]^\top$, i.e., the plots show the quantizer input $e(n)$ versus $e(n-1)$. In addition the stable chaotic limit sets are plotted using a gray scale; limit points which are frequently visited are lighter than limits points rarely visited. The figures indicate that both modulators are stable, but the space between the boundary of B_L and L reduces as h_2 is lowered to -1.65.

The modulator with $h_2 = -1.5$ was proven to be stable in [57]: it was shown that the regions of the state-space square with $|e| < 2$ which is mapped outside the square will return to the square again within a finite number of iterations, hence, this proves that the quantizer input is bounded. This approach works obviously only when the used square is fully inside B_L and this is not the case for $h_2 = -1.65$ (see Fig. 3.4).

When h_2 is approx. -1.7, L collides with ∂B_L and the modulator becomes unstable. This can be observed from Fig. 3.5 and Fig. 3.6 for $h_2 = 1.8$ where B_L has been estimated using n_{max} equal to 100 and 500. Observe that the estimated B_L sets becomes 'thinner' for higher n_{max}, i.e., the number of grid points surviving (i.e., staying below a bound) n_{max} iterations decreases with n_{max} □

3.5 The Escape Rate

The observations done in Example 3.2 inspire to the assumption that a certain fraction of orbits stemming from random initial conditions will not survive the $(k+1)$'th iteration for an unstable system. Consequently, it is assumed that the number of survivors decreases exponentially with k [13] and thereby the so-called *escape rate* γ is defined:

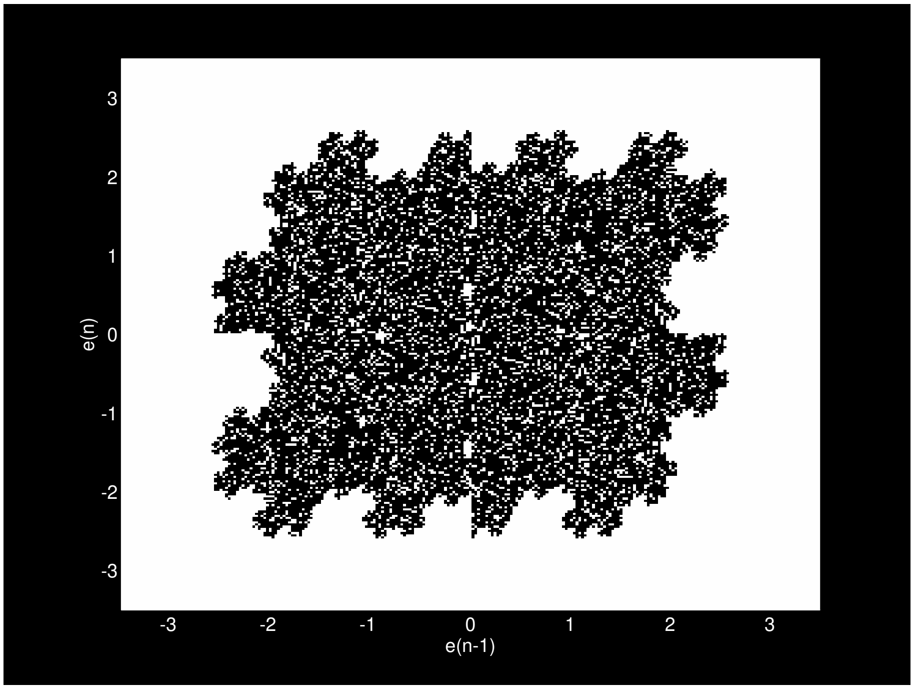

Figure 3.5: Estimate of the basin of attraction B_L for the modulator cf. Eq. (3.5) with $h_1 = 0.8$, $h_2 = -1.8$ and maximum iteration number $n_{max} = 100$.

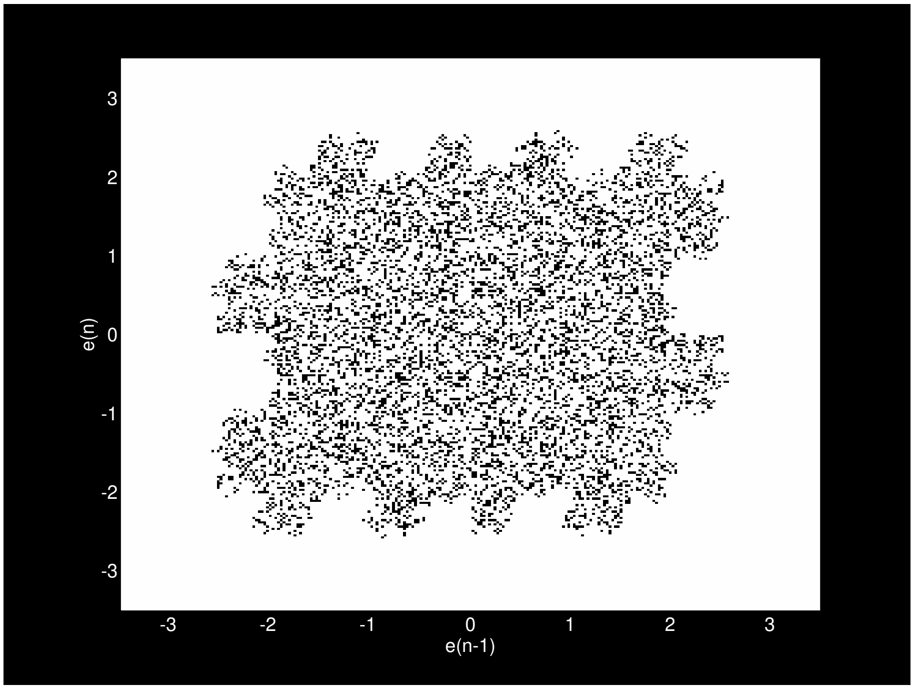

Figure 3.6: Estimate of the basin of attraction B_L for the modulator cf. Eq. (3.5) with $h_1 = 0.8$, $h_2 = -1.8$ and maximum iteration number $n_{max} = 500$.

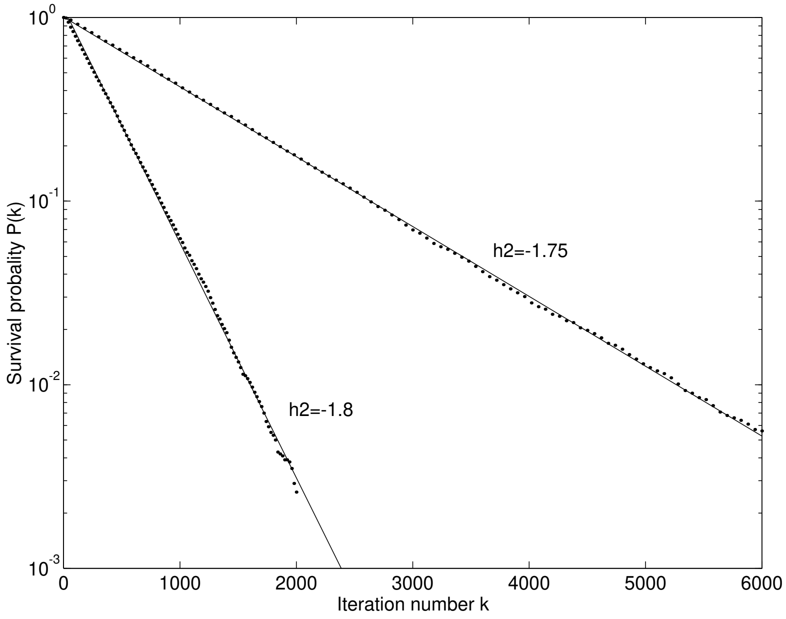

Figure 3.7: Survival probability $P(k)$ as a function of the iteration number k.

Assumption 3.2 *The probability $P(k)$ that the orbit of an unstable chaotic modulator with random initial condition stays inside a suitable bounding set W for k iterations is assumed to be:*

$$P(k) = \Pr\{\forall_{n \le k} : \boldsymbol{x}_n \in W\} = P_0\, e^{-\gamma k} \qquad (3.6)$$

where P_0 depends on the probability distribution of initial conditions $\boldsymbol{x}_0$ and γ is the escape rate. The mean survival time is defined as $T = 1/\gamma$.

In order to give a meaningful escape rate, the unstable repeller must be contained in the bounding set W used in Eq. (3.6). When an orbit leaves this set, the orbit should never return into W again, i.e., every point outside W should escape to infinity or another attractor. In Example 3.2 the bounding set W was the set of state-vectors with quantizer input magnitude less than some maximum value. When P_0 is unity, then the mean survival time $T = 1/\gamma$ is the average number of time steps the system 'lives' before it escapes. Generally P_0 is the fraction of initial points that do not escape within the first time steps. If the distribution of initial points covers far more than the limit set then P_0 will be small and it will take some time steps before the orbits starts to escape. The assumed exponential decay of $P(k)$ is thus only accurate for large k.

The existence of an escape-rate necessitates that the bounding set W contains a number of limit cycles or fixed points, i.e., for suitably chosen initial conditions, it is possible to stay forever inside W. This condition is normally fulfilled, since a chaotic limit set has infinitely many limit cycles.

The exponential decay is similar to the decay of radioactive material.

Example 3.3 The exponential decay of survivors, can be investigated numerically. The system is simulated with a large number of suitably distributed random initial conditions and the iteration number for which the system exceeds a given bound is estimated. Subsequently, the fraction of survivors are plotted on log-scale versus the iteration number on lin-scale yielding a straight line with slope equal to the escape rate. Fig. 3.7 shows this type of plot for the second order modulator cf. Eq. (3.5) for $h_1 = 0.8$. Two plots are shown: $h_2 = -1.75$ giving $\gamma = 8.8{\cdot}10^{-4}$ and $h_2 = -1.8$ giving $\gamma = 2.9{\cdot}10^{-3}$. Both plots are based on 10.000 random initial conditions and the orbits are considered as survivors until the quantizer input magnitude exceeds 10. The asymptotic slope of the survival plots (i.e., the escape rate) does naturally not depend on the choice of bounding set as long as the repeller is fully included herein □

3.6 The Structure of L and B_L

Generally, the shapes of the stable limit set L and the basin of attraction B_L are very complicated and impossible to describe analytically except for a very few cases (like Example 3.1). This makes it almost impossible to find a comprehensive analytical stability criterion for the general class of modulators. In the following a number of examples will demonstrate the complexity of especially the basin of attraction.

The next example demonstrates that a certain subset of second order modulators have a very simple B_L set which can be described analytically and in addition, the chaotic limit set can be bounded to a certain region. This fact allows the derivation of a sufficient stability criterion.

Example 3.4 Consider the second order multiple feedback modulator discussed in Example 2.3 (see Fig. 2.6) with constant input i. In order to normalize the feedback transfer function $H(z)$, the parameter b_1 is set to unity and the parameter b_2 is substituted for b. This normalization reflects the fact that there is only one degree of freedom for the choice of the zero of $H(z)$. The multiple feedback modulator is equivalent to an usual Σ-Δ modulator preceded filtered by a prefilter. The prefilter has in this case the following first order transfer function, cf. Eq. (2.14):

$$G(z)H^{-1}(z) = \frac{z^{-1}}{1 + (b-1)z^{-1}} \tag{3.7}$$

Since only constant input is considered, the prefilter scales the input by factor $1/b$. Consequently, scaling of the input for the multiple feedback modulator by a factor b makes the multiple feedback modulator equivalent to a usual Σ-Δ modulator for constant input. Using this scaling, the system can be characterized by the following map:

$$\boldsymbol{x}_{n+1} = \mathcal{F}(\boldsymbol{x}_n) = \begin{bmatrix} 1 & 1 \\ -a & 1 \end{bmatrix} \boldsymbol{x}_n - \operatorname{sgn}\left(\begin{bmatrix} 1 & 0 \end{bmatrix} \boldsymbol{x}_n\right) \begin{bmatrix} 1 \\ b \end{bmatrix} + \begin{bmatrix} 0 \\ bi \end{bmatrix} \tag{3.8}$$

The state vector is given by $\boldsymbol{x}_n = \begin{bmatrix} x_1(n) & x_2(n) \end{bmatrix}^\top$ where $x_1(n)$ is the quantizer input.

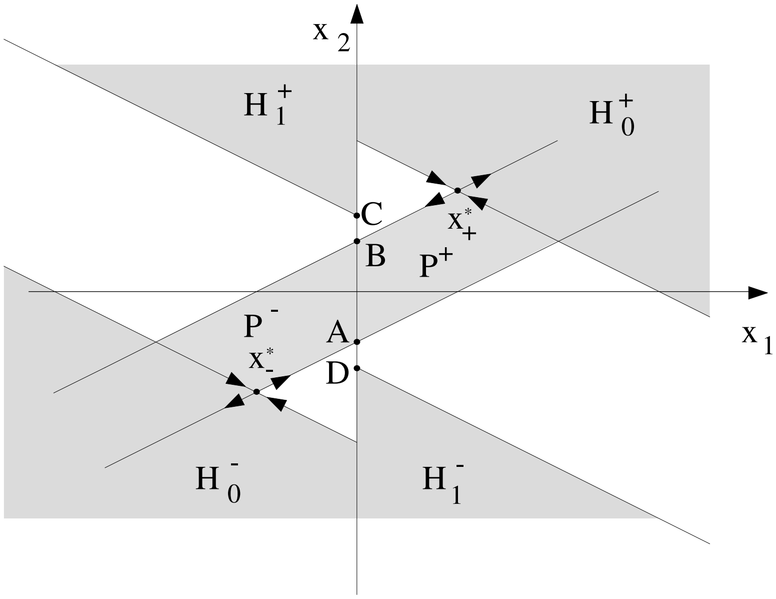

Figure 3.8: The geometry of the multiple feedback modulator from Example 3.4 for $a < 0$ and $b < \sqrt{-a}$.

For negative a, the system has two fixed points:

$$\boldsymbol{x}_{+}^{*} = \begin{bmatrix} \frac{b(1-i)}{-a} \\ 1 \end{bmatrix}, \boldsymbol{x}_{-}^{*} = \begin{bmatrix} \frac{-b(1+i)}{-a} \\ 1 \end{bmatrix} \tag{3.9}$$

The system has the eigenvalues or poles of $H(z)$:

$$z = 1 \pm \sqrt{-a} \tag{3.10}$$

For $-4 < a < 0$ there are two real valued eigenvalues; one outside and one inside the unit circle. Consequently, each of the two fixed points are *saddle points* and they have a stable and an unstable manifold [46]. The manifolds are just straight lines in the eigenvector directions and the manifolds are invariant to the system map. The stable manifold of $\boldsymbol{x}_{+}^{*}$ corresponding to the eigenvalue $z = 1 - \sqrt{-a}$ is given by:

$$\begin{aligned} \mathcal{F}(\boldsymbol{x} - \boldsymbol{x}_{+}^{*}) &= (1 - \sqrt{-a})(\boldsymbol{x} - \boldsymbol{x}_{+}^{*}) \Leftrightarrow \\ \left(\begin{bmatrix} 1 & 1 \\ -a & 1 \end{bmatrix} - (1 - \sqrt{-a})\boldsymbol{E} \right) \boldsymbol{x} &= \left(\begin{bmatrix} 1 & 1 \\ -a & 1 \end{bmatrix} - (1 - \sqrt{-a})\boldsymbol{E} \right) \boldsymbol{x}_{+}^{*} \Leftrightarrow \\ \begin{bmatrix} \sqrt{-a} & 1 \end{bmatrix} \boldsymbol{x} &= \frac{b(1-i)}{\sqrt{-a}} + 1 \end{aligned} \tag{3.11}$$

Similarly, the unstable manifold of $\boldsymbol{x}_{+}^{*}$ corresponding to the eigenvalue $1 + \sqrt{-a}$ is given by:

$$\begin{bmatrix} -\sqrt{-a} & 1 \end{bmatrix} \boldsymbol{x} = \frac{-b(1-i)}{\sqrt{-a}} + 1 \tag{3.12}$$

The stable manifold of $\boldsymbol{x}_+^*$ separates points with $x_1 > 0$ which iterate towards infinity from points which iterate towards the quantizer threshold at $x_1 = 0$. The geometry is shown in Fig. 3.8 for $b < \sqrt{-a}$ where the shaded area H_0^+ is the points which escape to infinity along the unstable manifold. Similarly, the stable manifold of $\boldsymbol{x}_-^*$ delimits the shaded area H_0^- of points with $x_1 < 0$ iterating towards infinity. Consequently, both fixed points and parts of their stable manifolds are on the boundary of B_L if the system is stable. The shaded set H_1^+ is the preimage of H_0^+ with respect to $\mathcal{F}_-$, i.e., $H_1^+ = \mathcal{F}_-^{-1}(H_0^+)$; furthermore, this set must be fully outside B_L for the same reasons as for H_0^+. the wedge H_1^+ is delimited by the line $x_1 = 0$ and by a line parallel to the stable manifold of $\boldsymbol{x}_+^*$ which intersects the line $x_1 = 0$ at the point $\mathrm{C} = \begin{bmatrix} 0 & x_{2,c} \end{bmatrix}^\top$ in Fig. 3.8. Furthermore, it can be concluded that the point C must be on ∂B_L since any neighborhood of C is mapped on both sides of the stable manifold of $\boldsymbol{x}_+^*$. The point C can be found from the condition that $\mathcal{F}_-(C)$ must be on the stable manifold of $\boldsymbol{x}_+^*$. Using Eq. (3.11) one arrives at:

$$\begin{aligned} \begin{bmatrix} \sqrt{-a} & 1 \end{bmatrix} \mathcal{F}_-\left(\begin{bmatrix} 0 \\ x_{2,c} \end{bmatrix}\right) &= \tfrac{b(1-i)}{\sqrt{-a}} + 1 \Leftrightarrow \\ \sqrt{-a}(x_{2,c}+1) + x_{2,c} + b(1+i) &= \tfrac{b(1-i)}{\sqrt{-a}} + 1 \Leftrightarrow \\ x_{2,c} &= \tfrac{1}{1+\sqrt{-a}}\left(\tfrac{b(1-i)}{\sqrt{-a}} + 1 - b(1+i) - \sqrt{-a}\right) \end{aligned} \tag{3.13}$$

Similarly, the shaded region H_1^- in Fig. 3.8 is the preimage of H_0^- with respect to $\mathcal{F}_+$ and the region is a wedge characterized the point D which like C is on ∂B_L.

The hatched region $\mathrm{P} = \mathrm{P}^+ \bigcup \mathrm{P}^-$ in Fig. 3.8 is delimited by all the four manifolds. The chaotic limit set L must be inside P since points outside the stable manifolds iterate to infinity and points just outside the unstable manifolds but inside the stable manifolds must iterate into P. Consequently, a sufficient but not necessary stability criterion is that the modulator is stable when the set P does not collide with the boundary of B_L, i.e., when P is a so-called *trapping region* which is mapped into it self. This condition holds when the both the points C and D on ∂B_L are outside P for $b < \sqrt{-a}$. The reason is that orbits escaping the system from an initial point inside P must sooner or later enter H_1^- or H_1^+ before the orbit escapes along one of the unstable manifolds and C and D will be the first points in these sets which will be hit at a collision between L and B_L. For positive input and $b < \sqrt{-a}$ the criterion is fulfilled as long as the point C does not cross the unstable manifold of $\boldsymbol{x}_+^*$. Using Eq. (3.12) and Eq. (3.13) the stability criterion holds for positive input i when:

$$\begin{aligned} \begin{bmatrix} -\sqrt{-a} & 1 \end{bmatrix} \begin{bmatrix} 0 \\ x_{2,c} \end{bmatrix} &> \tfrac{-b(1-i)}{\sqrt{-a}} + 1 \Leftrightarrow \\ \tfrac{1}{1+\sqrt{-a}}\left(\tfrac{b(1-i)}{\sqrt{-a}} + 1 - b(1+i) - \sqrt{-a}\right) &> \tfrac{-b(1-i)}{\sqrt{-a}} + 1 \Leftrightarrow \\ i &< \tfrac{a+b}{b(1+\sqrt{-a})}\ , \ \text{for } i > 0\ ,\ b < \sqrt{-a} \end{aligned} \tag{3.14}$$

Due to the symmetry of the system, the sufficient stability criterion can be extended to:

$$|i| < \frac{a+b}{b(1+\sqrt{-a})}\ ,\ b < \sqrt{-a} \tag{3.15}$$

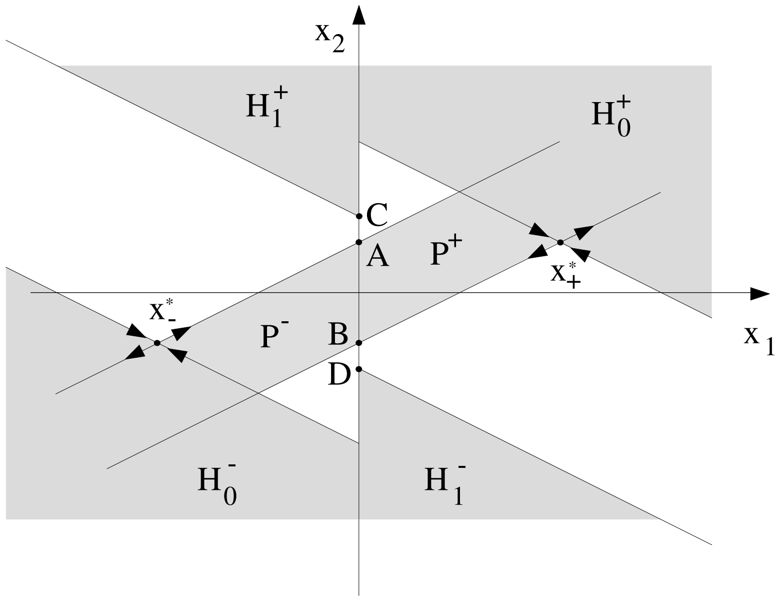

Figure 3.9: The geometry of the multiple feedback modulator from Example 3.4 for $a < 0$ and $b > \sqrt{-a}$.

For $b = \sqrt{-a}$ an interesting situation arises: the unstable manifolds of the two fixed points become identical and the set P is thus the line segment between the fixed points and P is identical to B_L. Consequently, the chaotic limit set L of the two dimensional system is embedded in a one dimensional subspace and the dynamics of the system can thus asymptotically be described by a one dimensional map. This phenomenon is also reflected by the feedback transfer function $H(z)$ from Eq. (2.11):

$$H(z) = \frac{z^{-1}\left(1-(1-b)z^{-1}\right)}{(1-(1+\sqrt{-a})z^{-1})(1-(1-\sqrt{-a})z^{-1})} \tag{3.16}$$

which shows that a real pole at $1-\sqrt{-a}$ and a real zero at $1-b$ are cancelling for $b = \sqrt{-a}$, i.e., the system is a first order modulator embedded in a second-order system. The same situation occurs for $b = -\sqrt{-a}$, but in this case, the resulting non-chaotic first order modulator is unstable due to the embedding into a diverging state-space.

The geometry of the system is shown for $b > \sqrt{-a}$ in Fig. 3.9; the set P is two-dimensional again. For this case and for positive input, the point C must not cross the unstable manifold of $\boldsymbol{x}_-^*$ in order to meet the sufficient stability criterion. Using the symmetry, Eq. (3.13) and an equation for the unstable manifold of $\boldsymbol{x}_-^*$ the criterion becomes:

$$\begin{gathered}
\begin{bmatrix} -\sqrt{-a} & 1 \end{bmatrix} \begin{bmatrix} 0 \\ x_{2,c} \end{bmatrix} > \tfrac{b(1+|i|)}{\sqrt{-a}} - 1 \Leftrightarrow \\
\tfrac{1}{1+\sqrt{-a}} \left(\tfrac{b(1-|i|)}{\sqrt{-a}} + 1 - b(1+|i|) - \sqrt{-a} \right) > \tfrac{b(1+|i|)}{\sqrt{-a}} - 1 \Leftrightarrow \\
|i| < \tfrac{1/b-1}{1+1/\sqrt{-a}} \,,\; b > \sqrt{-a}
\end{gathered} \tag{3.17}$$

> The sufficient stability criteria of Eq. (3.15) and Eq. (3.17) indicate the maximum stable constant input amplitude as a function of the modulator parameters a and b. Numerical stability analysis will validate these criteria in Example 3.10 and it will appear that the criteria can be somewhat pessimistic for some parameter values. An improved stability criterion for $b > \sqrt{-a}$ has been made in [50]. This criterion is based on a more detailed (and complex) geometric analysis □

This example allowed an analytical study primarily due to the fact that the eigenvalues were both real and one was inside the unit circle. This kind of stability analysis is very difficult to perform on general high-order modulators. Even the general second-order modulator with complex poles is virtually impossible to analyze. The traditional double-loop modulator (i.e., a second-order modulator with two poles at $z = 1$) has been investigated intensively and rigorous stability criteria have been made [67, 47, 25].

3.7 Reverse Time Simulation

When all eigenvalues are complex and outside the unit circle, the analysis gets extremely complex. The basins of attraction shown in Fig. 3.3 and 3.4 correspond to modulators of this category and the basins of attraction have indeed a very complex structure. A closer analysis shows that the basin boundaries contain a large number of limit cycles and seem to be fractal. This fact indicates that an unstable chaotic limit set or repeller exists on ∂B_L simultaneously with the stable chaotic limit set L.

Normally it is almost impossible to locate a chaotic repeller numerically; however, in this particular case a reversal of the time axis can solve the problem. When a system with all eigenvalues outside the unit circle is iterated in reverse time, all limit cycles become attracting. As mentioned in Ch. 2, the map of a Σ-Δ modulator is noninvertible in a certain region U where two solutions exist. Consequently, reverse time simulation is generally a traversal of a binary tree where each time step is a node with a corresponding point in the state-space.

When a node is in U, there will be two branches from the node corresponding to the two inverse map solutions. Every node in the tree corresponds thus to points in state-space which ends on the initial node in forward time. Some of the branches in the tree can be the beginning of a path which never enters U, i.e., there is only one branch on each of the succeeding nodes. These paths are attracted to a limit cycle which is fully outside U. For some modulators, there are no such limit cycles and almost every branch will sooner or later lead to a node in U with two branches again leading to nodes in U. In these cases a chaotic repeller exists which is attracting in reverse time. Consequently, when a path in the tree, which repeatedly enters U, is followed, the state-space points corresponding to the nodes of the path will get closer and closer to the repeller. Picking the branches randomly, the entire repeller can be sampled and plotted graphically. Empirically, this procedure has proven to work irrespectively of the initial point (i.e., the initial node). This seems to indicate that the forward time system can reach any point in the state-space using initial conditions arbitrarily close to the repeller.

> **Example 3.5** Fig. 3.10 demonstrates an example of reverse time simulation using random branches of the stable modulator used for Fig. 3.3 (i.e., feedback

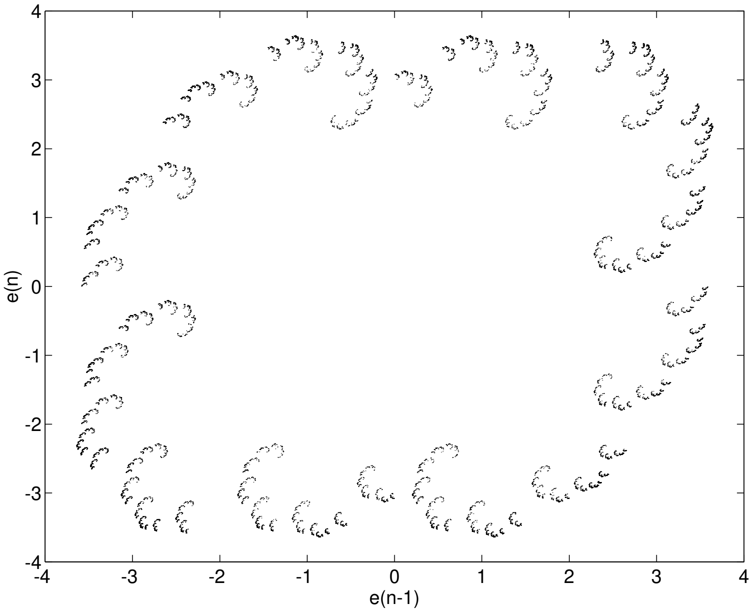

Figure 3.10: Reverse time simulation revealing the chaotic repeller of a stable modulator cf. Eq. (3.5) with $h_1 = 0.8$ and $h_2 = -1.5$. Compare the plot to Fig. 3.3.

filter cf. Eq. (3.5) with $h_1 = 0.8$ and $h_2 = -1.5$). A total of 30.000 points are plotted revealing a very complicated fractal set which fits the boundary of B_L shown in Fig. 3.3; however, this time reverse plot shows much more details □

A chaotic repeller is normally not observable since only orbits with initial conditions exactly on the repeller itself will follow the repeller. Orbits starting close to the repeller will either escape to infinity or be attracted to possible stable limit sets. In the former example the repeller lives on the boundary of the basin of attraction for a stable chaotic limit set, i.e., points close to the repeller are either escaping or being attracted to the stable chaotic limit set. The repulsion from the boundary repeller is forming a kind of 'embankment' around the stable limit set which keeps the stable limit set together: points on the inner side of the embankment stays inside and points on the outside are escaping. Whereas ∂B_L is simple in a one-dimensional system (see Example 3.1), the geometry gets more and more complex for higher order systems.

The degree on instability of a repeller on ∂B_L can, as for other unstable limit sets, be quantified, i.e., an escape rate can be found.

Example 3.6 The escape rate of the repeller shown in Fig. 3.10 can be measured by simulations when a suitable bounding set W is defined. Let W be the set of state-space points satisfying $(2 < |e(n)| < 4) \wedge (2 < |e(n-1)| < 4)$. The stable chaotic limit set is almost fully outside this set while the repeller in fully inside (see also Fig. 3.3). The survival probability $P(k)$ was measured versus the iteration number k using simulations with random initial conditions and the result was plotted in Fig. 3.12. The graph shows an almost straight

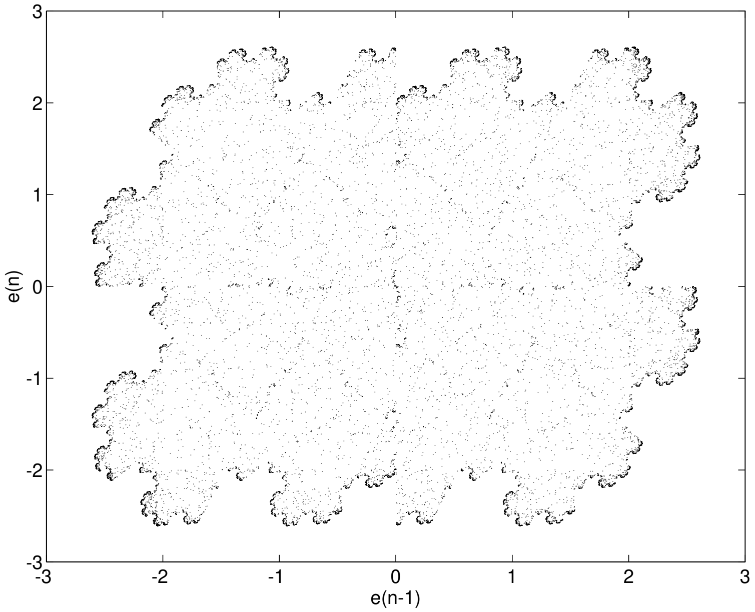

Figure 3.11: Reverse time simulation revealing the chaotic repeller of an unstable modulator cf. Eq. (3.5) with $h_1 = 0.8$ and $h_2 = -1.8$.

> line with a slope corresponding to an escape rate of 0.175 equivalent to a mean survival time of approx. 6 time steps. This is indeed a very unstable repeller □

As demonstrated previously, the modulator cf. Eq. (3.5) becomes unstable when h_2 is lowered to approx. -1.7 for $h_1 = 0.8$. At this point, the stable chaotic limit set L collides with the boundary of the basin of attraction which contains a chaotic repeller. During this collision, the chaotic repeller absorbs the formerly stable chaotic limit set leaving only a chaotic repeller. This repeller can again be seen using reverse time simulation for this type of collision between a stable and unstable chaotic limit set.

> **Example 3.7** Fig. 3.11 shows a reverse time simulation of the unstable modulator cf. Eq. (3.5) with $h_1 = 0.8$ and $h_2 = -1.8$. The plot reveals the chaotic repeller □

It is only a subset of the Σ-Δ modulators which has a chaotic repeller living at the boundary of the basin of attraction. For some parameter values, a reverse time simulation using random branches is always attracted to a dominating limit cycle with no points in the noninvertible region U (such limit cycles will trap the system in reverse time). In principle, a reverse time simulation can get arbitrarily close to any limit cycle, if certain branches are deliberately used. However, when random branches are used, the reverse time system will asymptotically be trapped by a limit cycle totally outside U or be attracted to a chaotic repeller consisting of infinitely many limit cycles with orbits entering U. In fact, simulations seem to indicate that the asymptotic behavior of the reverse time system using

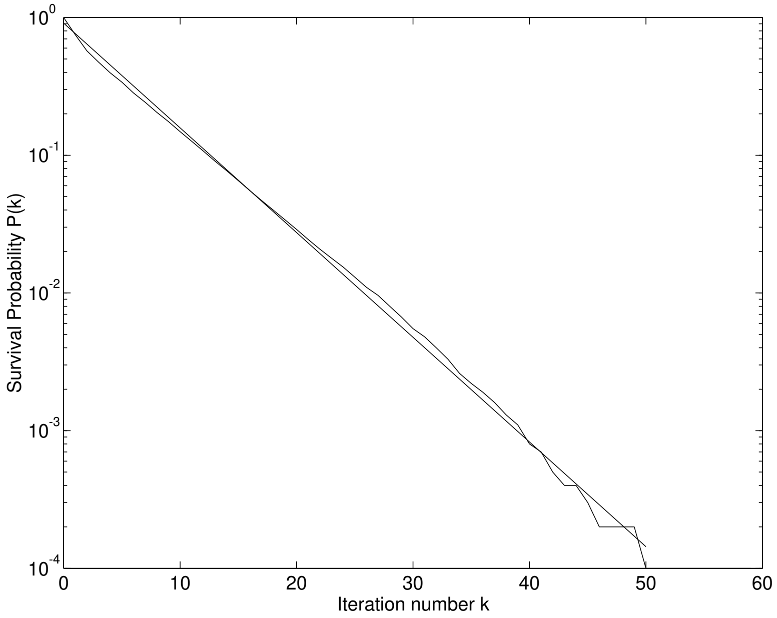

Figure 3.12: Survival probability $P(k)$ vs. iteration number k for the chaotic repeller on ∂B_L for the modulator cf. Eq. (3.5) with $h_1 = 0.8$ and $h_2 = -1.5$. The plot indicates an escape rate $\gamma = 0.175$.

random branches often is independent of the initial condition. In this case, the reverse time system is attracted to points which by means of an arbitrarily small perturbation can hit any place in the entire state-space in forward time. Such points must obviously belong to ∂B_L.

Typically, a modulator has a dominating limit cycle outside U corresponding to a periodic output signal having p positive codes followed by p negative codes which is called a $(\overline{p,p})$ limit cycle. Imagine that the modulator also has a $(\overline{p+1,p+1})$ limit cycle with an orbit entering U. The $(\overline{p+1,p+1})$ limit cycle cannot attract the reverse time system unless certain branches a used for each time step; hence, when random branches are used, it is almost certain that the system is trapped by the $(\overline{p,p})$ limit cycle. In some situations, the modulator parameters can be perturbed such that both the orbits of the dominant $(\overline{p,p})$ limit cycle and the $(\overline{p+1,p+1})$ limit cycle enter U at the same time, and at this point, a chaotic repeller might arise. This can be explained as follows: when the $(\overline{p,p})$ cycle is followed in reverse time, a $(p,p+1)$ code segment might be generated when a 'wrong' branch is used at a point in U. When the system later enters U again, two branches can be taken. The result is that the reverse time system repeatedly enters U and will produce a random output code sequence composed of the four segments (p,p), $(p,p+1)$, $(p+1,p)$ and $(p+1,p+1)$. The reverse time system is no longer attracted to a single limit cycle but to a chaotic limit set or repeller which consists of infinitely many 'hybrid' limit cycles composed of the previously mentioned four segments.

Large parts of the stable and chaotic parameter space for the second order modulator cf. Eq. (3.5) show a diversity of beautiful chaotic repellers coexisting with the stable (forward) chaotic limit set.

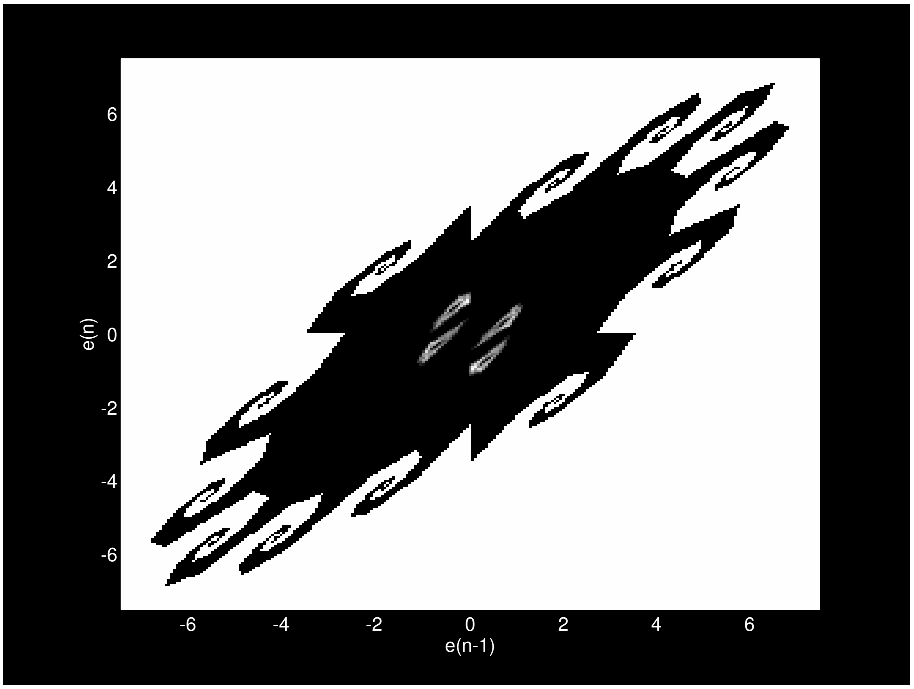

Figure 3.13: Basin of attraction and limit set for the modulator cf. Eq. (3.18) with zero input and $b = 0.45$.

Example 3.8 Consider the modulator with feedback filter

$$H(z) = \frac{z^{-1} - bz^{-2}}{1 - 1.95z^{-1} + 1.1z^{-2}} \tag{3.18}$$

with complex poles outside the unit circle. For $b = 0.45$ and zero input the system is stable and has $(\overline{p,p})$ limit cycles for $p = 1,2,3,4,5,6$. Only the $(\overline{6,6})$ limit cycle is totally outside U. This can be observed from the corresponding periodic quantizer input signal $e(n)$ which is the repeated sequence $\langle$1.8 4.2 5.5 5.7 4.4 1.8 -1.8 -4.2 -5.5 -5.7 -4.4 -1.8$\rangle$. Recall from Eq. (2.19) that modulator states with quantizer input satisfying $|e| < -2H(z = 0)$ are in the set V which is mapped into U. Consequently, the $(\overline{6,6})$ limit cycle does not enter U since $\min(|e(n)|) > -2H(z = 0) = 2b/1.1 \approx 0.82$ and the reverse time system is thus asymptotically attracted to this limit cycle. All the other $(\overline{p,p})$ limit cycles enter U. Fig. 3.13 shows the basin of attraction for the modulator including the stable (forward) limit set. As expected, the $(\overline{6,6})$ limit cycle is on the boundary of B_L which forms a spiral around all the points on the limit cycle. Fig. 3.14 is a magnified plot around one of the points on the limit cycle and this plot shows clearly that B_L seems to spiral infinitely many times around the centers.

A (7,7) limit cycle arises when b is increased to 0.5. This limit cycle enters U and the system is thus still attracted to the (6,6) limit cycle in reverse time. Fig. 3.15 shows the basin of attraction and the limit set for $b = 0.5$. Notice that the spirals around the (6,6) limit cycle are preserved but the shape of

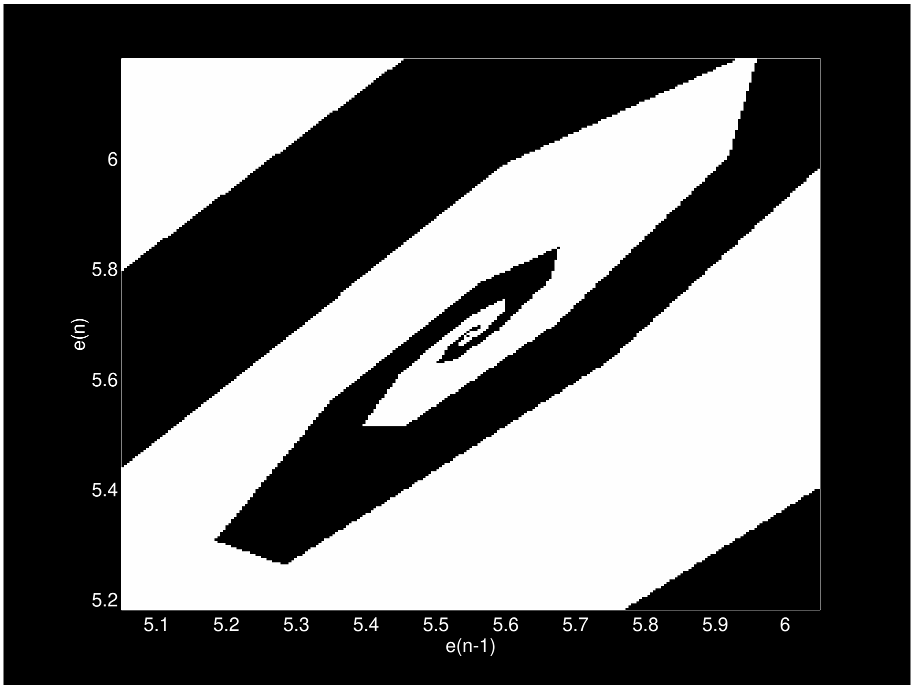

Figure 3.14: Magnified plot of the basin of attraction shown in Fig. 3.13 The center of the spiral is a point on the $(\overline{6,6})$ limit cycle.

B_L has become more complicated than for $b = 0.45$. This is probably due the onset of the (7,7) limit cycle outside B_L which is plotted with circles.

When b is increased to 0.56, both the $(\overline{6,6})$ and the $(\overline{7,7})$ limit cycles enter U which indicates the presence of a chaotic repeller on ∂B_L. This is indeed confirmed by Fig. 3.16 which is a plot of a reverse time simulation revealing a complex fractal set located in the space 'between' the $(\overline{6,6})$ and $(\overline{7,7})$ limit cycles which are shown with circles and crosses. Orbits close to this repeller will produce an output code composed of the segments $(6,6)$, $(6,7)$ $(7,6)$ and $(7,7)$.

Around $b = 0.58$ the $(\overline{7,7})$ limit cycle is totally outside U and the chaotic repeller disappears. Now the reverse time system is attracted to the $(\overline{7,7})$ limit cycle.

The bifurcations of the $(\overline{p,p})$ limit cycles can be observed in Fig. 3.17 which shows the minimum quantizer input magnitudes as a function of b. The straight dashed line shows the V-interval limit at $-2H(z = 0) = 2b/1.1$. The plot shows clearly the succession of dominant $(\overline{p,p})$ limit cycles outside U and the windows for which there exist a chaotic repeller when no $(\overline{p,p})$ limit cycle is outside U. Many of the conclusions drawn from this example can be seen from the bifurcation diagram: the small window with a chaotic repeller around $b = 0.56$ (Fig. 3.16) surrounded by regions where either the $(\overline{6,6})$ or the $(\overline{7,7})$ limit cycle is outside U and on ∂B_L.

The very nice basin of attraction seen for $b = 0.45$ is at a region where the dominant $(\overline{6,6})$ limit cycle is peaking while no other limit cycles are outside U. At slightly lower b values, both the $(\overline{5,5})$ and $(\overline{6,6})$ limit cycles are outside U and this situation will be analyzed in detail later in Example 3.9. Fig. 3.17

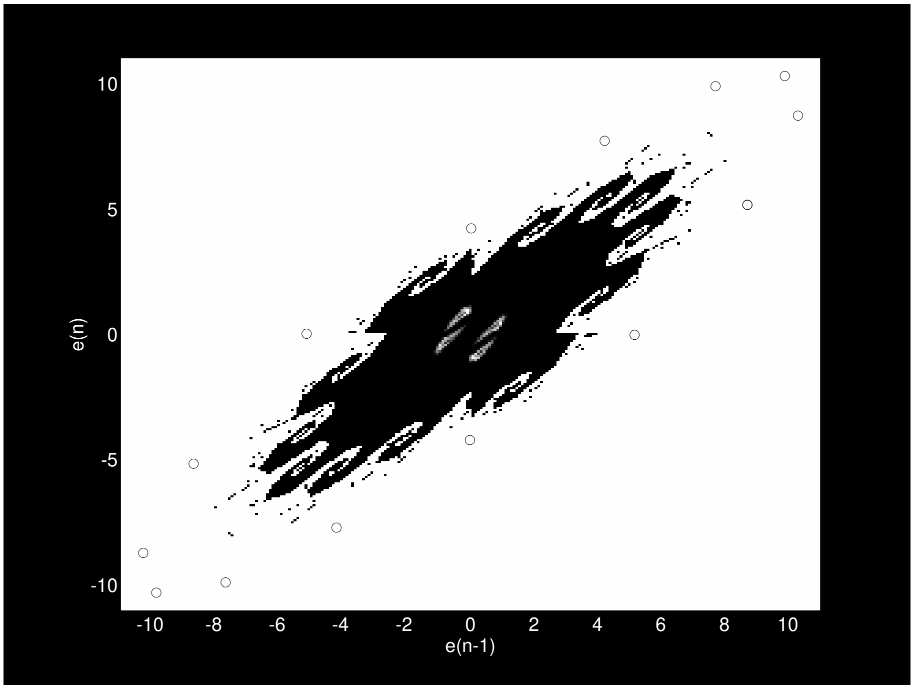

Figure 3.15: Basin of attraction and limit set for the modulator cf. Eq. (3.18) with zero input and $b = 0.5$. The $(\overline{7,7})$ limit cycle is plotted with circles.

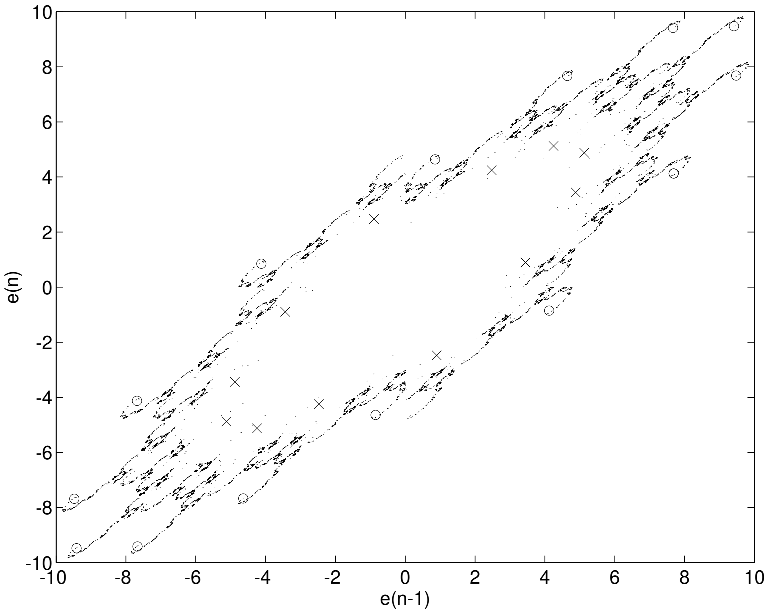

Figure 3.16: Reverse time simulation of the modulator cf. Eq. (3.18) with zero input and $b = 0.56$. The (6,6) and (7,7) limit cycles are plotted with crosses and circles, respectively.

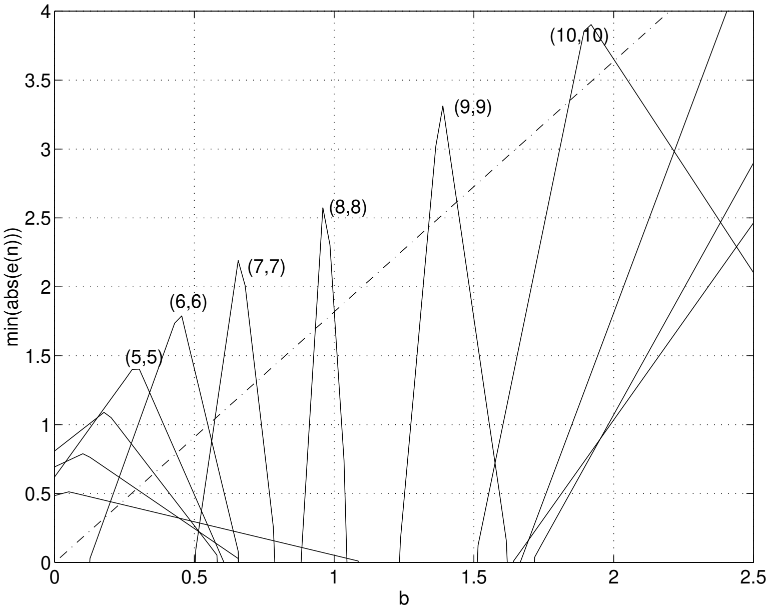

Figure 3.17: Minimum quantizer input magnitudes of a number of $(\overline{p,p})$ limit cycles for the second order modulator of Eq. (3.18) as a function of b. Magnitudes under the dashed line indicates that the corresponding limit cycle enters the noninvertible region U.

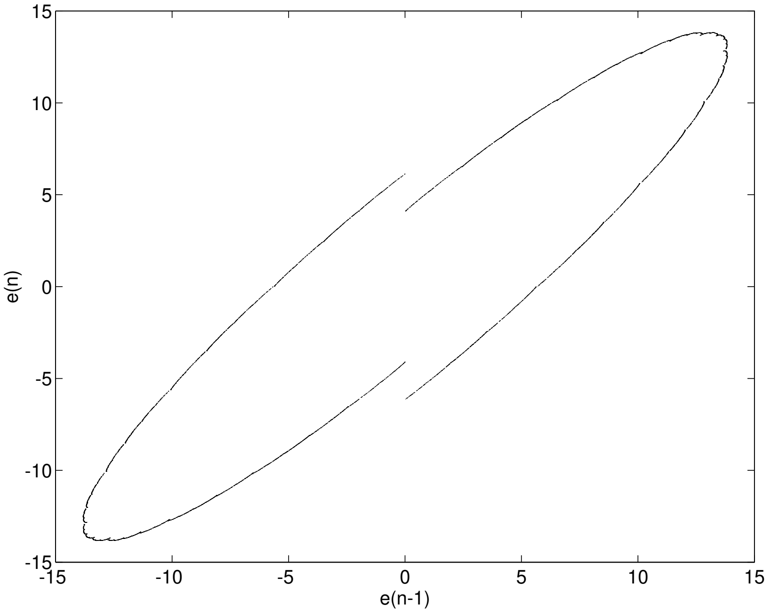

Figure 3.18: Chaotic repeller obtained by reverse time simulation for the modulator cf. Eq. (3.18) with zero input and $b = 1.1$.

also predicts the existence of a window with a chaotic repeller for b between approx. 0.75 and 0.9 which is confirmed by simulations. Later for b between approx. 1.0 and 1.3 there is another window with a chaotic repeller and for $b = 1.1$ a very interesting repeller can be observed in Fig. 3.18 using reverse time simulation. The repeller forms an almost smooth one dimensional curve with an elliptical shape. This is probably due to the fact that the $(\overline{2,2})$ limit cycle disappears for $b > 1.1$ and no $(\overline{p,p})$ limit cycles at all exist until a $(\overline{9,9})$ limit cycle appears at approx. $b = 1.25$.

For zero input and $b > 1$, two fixed points e^*_- and e^*_+ emerge. They can be located using the feedback transfer function of Eq. (3.18) evaluated at $z = 1$:

$$\begin{aligned} -\mathrm{sgn}(e^*)H(z=1) &= e^* \quad \Leftrightarrow \\ e^* &= \pm\tfrac{b-1}{0.15} \qquad , b > 1 \end{aligned} \tag{3.19}$$

These fixed points are outside V and U for:

$$\begin{aligned} &\tfrac{b-1}{0.15} > \tfrac{2b}{1.1} \quad \Leftrightarrow \\ &b > 1.375 \end{aligned} \tag{3.20}$$

At $b = 1.375$ there is also a $(\overline{9,9})$ limit cycle outside U thus giving a total of three limit sets which all can be trapping in reverse time. However, simulations have shown that only for initial conditions in a small neighborhood, the fixed points can be reached in reverse time so the $(\overline{9,9})$ limit cycle will dominate the shape of B_L, i.e., the limit cycle is on ∂B_L. To be exact: the fixed points will also be on ∂B_L if they do not belong to the forward chaotic limit set L, but in that case the fixed points are fully surrounded by B_L, i.e., a deleted neighborhood of each fixed point exists which is contained in B_L. In fact simulations show that the fixed points outside U are not included in L and it seems to be a general fact that L is strictly composed of limit cycles (and fixed points) which enter U. Fig. 3.19 shows L and B_L for $b = 1.7$ where a chaotic repeller coexists with the fixed points at $e^* = \pm 0.7/0.15 = \pm 4\frac{2}{3}$ which seem to 'blow holes' in L. Between $b = 1.8$ and $b = 2$ the system has one dominant $(\overline{10,10})$ limit cycle outside U and for $b > 2$ a region with a chaotic repeller starts. When b exceeds approx. 2.1 the system becomes unstable and reverse time simulations using random branches are asymptotically trapped by the fixed points outside U after a chaotic transient □

When a forward orbit of a modulator shows long code segments which correspond to orbits on ∂B_L, it can be concluded that the forward limit set is very close to a collision with ∂B_L. The knowledge of the limit sets on ∂B_L can thus be used to give early warnings against instability.

3.8 Unreliable Modulators

The definition of the escape rate shows that instability is a matter of degrees; if γ is very small it is very unlikely that instability occurs for even long simulations. Since the computational resources limit the maximum practical length of simulations, there might exist unstable systems with very small escape rate which cannot with certainty be classified correctly as unstable by means of simulations. This is especially problematic for

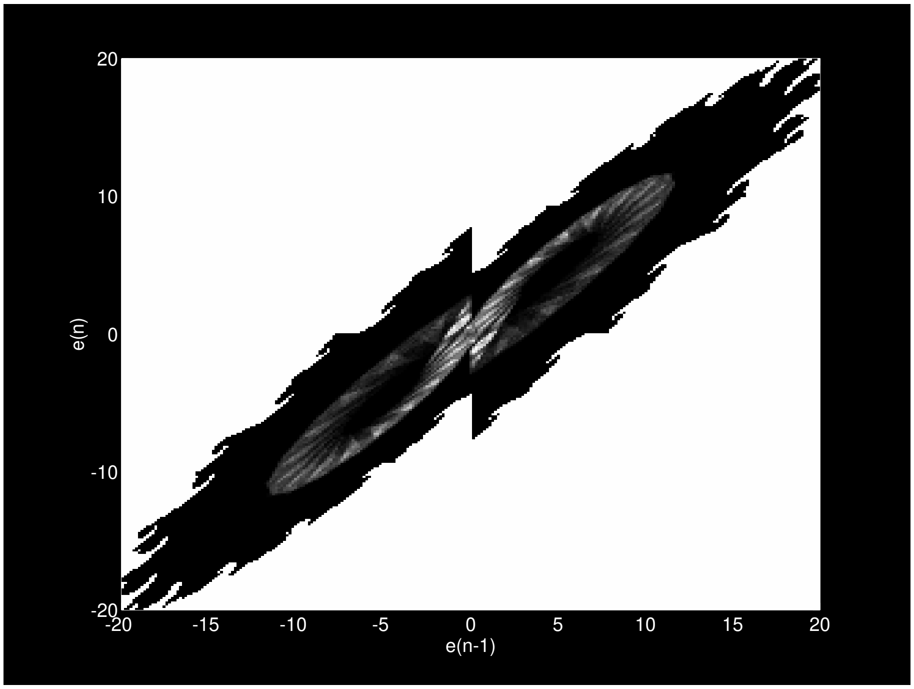

Figure 3.19: Basin of attraction and limit set for the modulator cf. Eq. (3.18) with zero input and $b = 1.7$. Notice the behavior of the limit set around the repelling fixed points at $e^* = \pm 0.7/0.15 = \pm 4\frac{2}{3}$.

systems which have an escape rate which changes slowly with parameter perturbations: if the systems show a weak instability on the longest time scale simulations allow, it is not certain that a perturbation normally improving stability, e.g., reduction of the input signal amplitude, will ensure stability. Such systems are loosely defined as being *unreliable*. On the other hand, for a *reliable* modulator the escape rate increases fast when instability occurs and stability can easily be assured by a slight perturbation in parameter space. If the perturbation is in the wrong direction, the escape rate will increase significantly and a perturbation in the opposite direction can be used.

Example 3.9 Recall the second order modulator used in Example 3.8 with feedback transfer function given in Eq. (3.18). The bifurcation diagram in Fig. 3.17 shows that a region exists for $b < 0.4$ where several $(\overline{p,p})$ limit cycles coexist outside U and several of these can be globally attracting in reverse time. In this region B_L is very complex and starts having 'holes' due to the fact that points on ∂B_L emerges at places which formerly were inside B_L.

The basin of attraction is shown in Fig. 3.20 for zero input and $b = 0.26$ and the corresponding chaotic limit set is shown of Fig. 3.21. This is very clearly an example of an unreliable modulator: the chaotic limit set is scattered about and fully surrounded by holes in B_L. A slight parameter perturbation may cause a collision between L and one of the holes in B_L resulting in a unstable system, probably with a low escape rate.

The basin of attraction forms again spirals around a limit cycle; this time the $(\overline{5,5})$ limit cycle which as expected is outside U. At the same time the

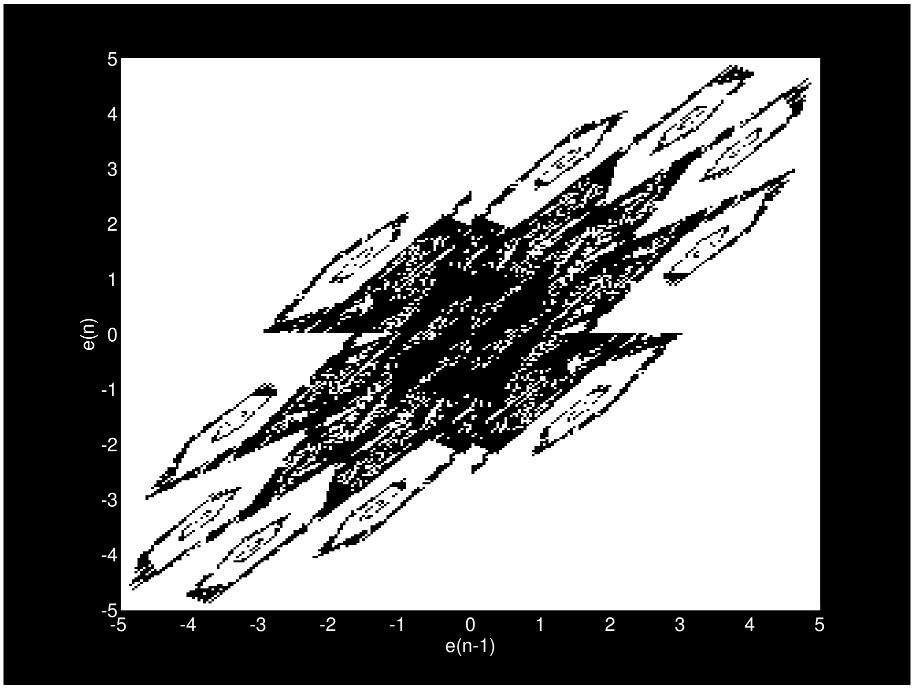

Figure 3.20: Basin of attraction for the modulator cf. Eq. (3.18) with zero input and $b = 0.26$.

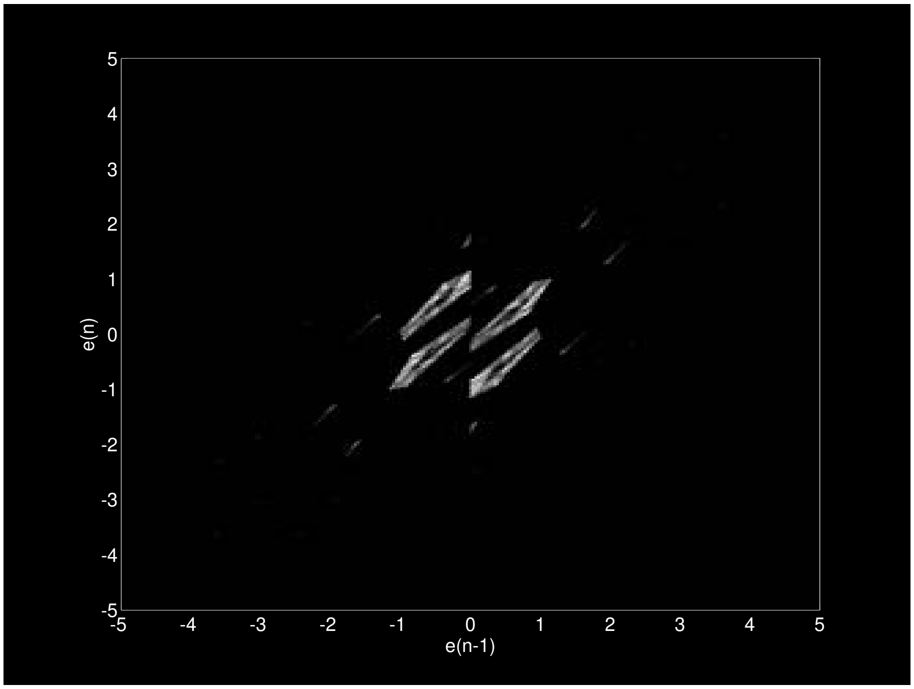

Figure 3.21: Limit set for the modulator cf. Eq. (3.18) with zero input and $b = 0.26$.

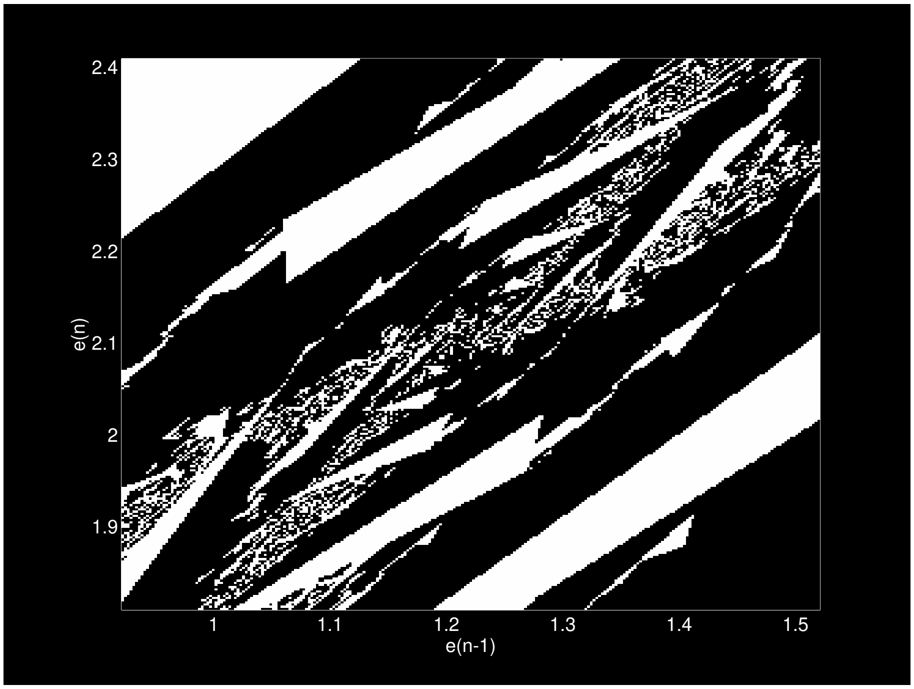

Figure 3.22: Magnified plot of the basin of attraction for the modulator cf. Eq. (3.18) with zero input and $b = 0.26$. The center of the plot is a point on the $(\overline{4,4})$ limit cycle.

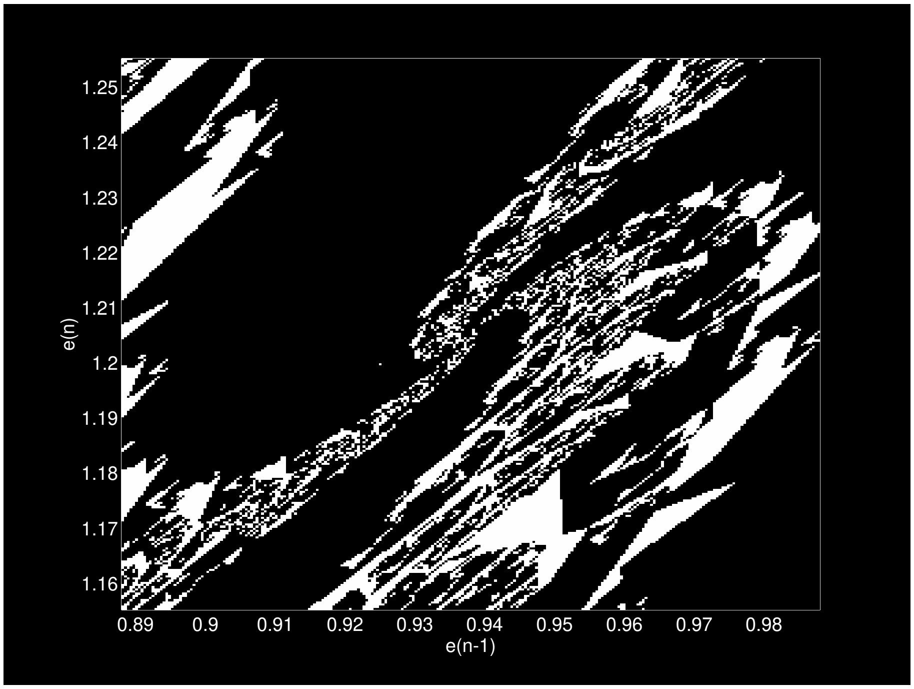

Figure 3.23: Magnified plot of the basin of attraction for the modulator cf. Eq. (3.18) with zero input and $b = 0.26$. The center of the plot is a point on the $(\overline{3,3})$ limit cycle.

limit cycles $(\overline{6,6})$, $(\overline{4,4})$ and $(\overline{3,3})$ exist and all of them are outside U as well. However, the $(\overline{5,5})$ limit cycle has the highest minimum quantizer input magnitude. A closer inspection of the magnified plots in Fig. 3.22 and Fig. 3.23 reveals that B_L also forms internal spirals around the $(\overline{4,4})$ and $(\overline{3,3})$ limit cycles and these internal spirals can explain the holes in B_L. The $(\overline{6,6})$ limit cycle is totally outside B_L indicating that this limit cycle can only be reached in reverse time from outside B_L. The strange B_L set can thus be explained by the coexistence of several limit cycles on ∂B_L outside U which all are globally attracting in a reverse time simulation using random branches □

The previous example showed clearly that stability analysis of second-order modulators is generally very complicated. The next example validates a stability criterion derived in Example 3.4 for a certain class of second-order modulators.

Example 3.10 The stability of a certain class of second-order modulators with a real pole outside and inside the unit circle was investigated in Example 3.4. It was possible to derive a sufficient stability criterion (Eq. (3.15) and Eq. (3.17). In order to validate the analysis, the parameter space (b, i) was investigated by simulations for $a = -0.05$, i.e., b and i was sampled in a uniform 512×512 grid for $0 < b < 1.5$ and $0 < i < 1$. For each point, the modulator was simulated 200 times with random initial conditions. The simulations was stopped when the stable manifolds were crossed (i.e., when the system escaped) or when a maximum of 1000 iterations was elapsed. For each parameter set, the average iteration number was determined and the result is shown in Fig. 3.24 using a gray scale. The criteria of Eq. (3.15) and Eq. (3.17) are shown as black curves. Eq. (3.15) is very accurate and the onset of stability for $\sqrt{b} < -a$ is very sudden. For $\sqrt{b} > -a$, the Eq. (3.17) criterion is very conservative and the onset of instability is not so well defined. This is clearly a more unreliable region. For b approaching unity, the plot is rippling for increasing input i, i.e., there are unstable 'ridges' with stable 'valleys' in between. When a valley is followed, the mean survival time is only increasing slowly □

3.9 Nonchaotic Modulators

When the modulator is nonchaotic, allmost all limit cycles are attracting, i.e., asymptotically stable [25]. This means that after a transient the modulator will asymptotically end on a limit cycle, regardless of the initial condition. Consequently, the basins of attraction corresponding to all the limit cycles will induce a partitioning on the entire state-space. The basin boundaries can be very complex and this gives rise to very long chaotic transients before the orbit is attracted to a limit cycle.

For high-order modulators ($N > 2$), some of the limit cycles may have extreme magnitudes and give poor signal-to-noise ratios. Consequently, when a modulator is attracted to such a limit cycle, the modulator is practically unstable. Normally, it is easy to distinguish limit cycles corresponding to stable and unstable modulator operation since limit cycles with unstable operation have extreme maximum amplitudes. The state-space can thus

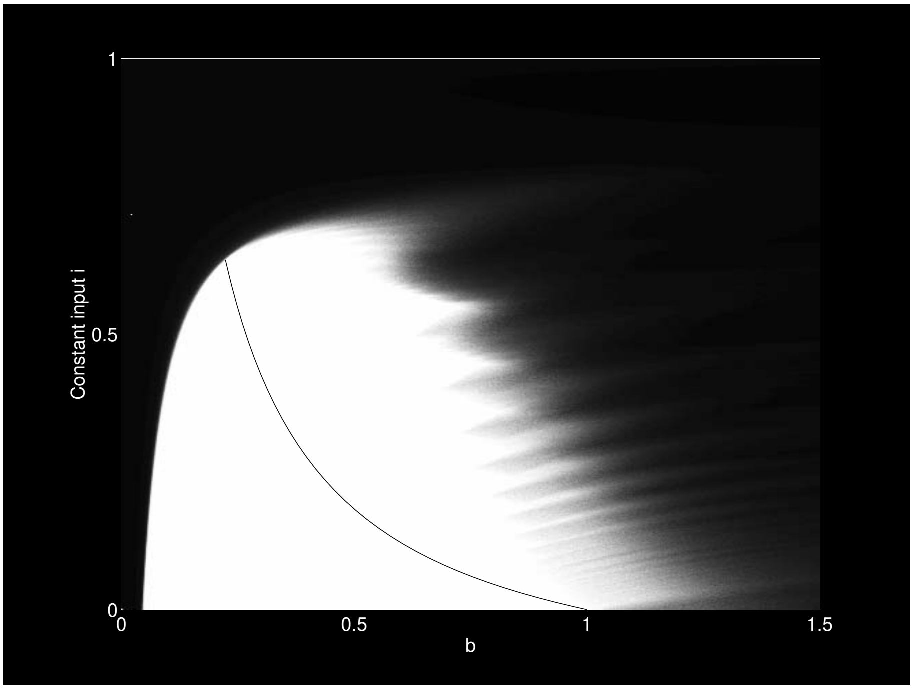

Figure 3.24: A gray-scale plot of the mean survival time of the modulator of Example 3.4 with $a = -0.05$. The white region has a mean survival time of a minimum of 1000 time-steps. Darker regions have lower survival time.

naturally be partitioned into two disjoint sets B_S and B_U, i.e., sets of points which are attracted to 'stable' and 'unstable' limit cycles, respectively. This fact emphasizes that the practical modulator stability concept must not be confused with the asymptotical stability of limit sets, cf. Definition 2.3. In fact, it is very difficult to give an adequate and comprehensive stability definition for nonchaotic modulators. This question will be discussed later in chapter 5.

3.10 Summary

This chapter analyzed basins of attraction of chaotic modulators. The onset of instability was characterized as a boundary crisis, i.e., the chaotic limit set collides with the boundary of the basin of attraction. After the collision, the chaotic attractor becomes an unstable repeller. The degree of instability was quantified through the definition of the escape-rate which is the slope of the survival probability versus iteration number in lin-log plot. There is always a positive probability that the system survives for arbitrarily long time; the closer the initial condition is on the repeller, the longer the system survives.

Some first and second-order systems were investigated in detail. The different examples showed the complexity of the basin of attraction for chaotic second-order systems. In some cases, a chaotic repeller is found on the boundary of the basin of attraction. The parameter regions for which a boundary repeller existed could be predicted using limit cycle analysis.

The complexity of even second-order modulators shows how difficult it is to derive a comprehensive, general and not to conservative stability criterion. Only in a few special cases of second-order modulators, it has been possible to derive analytical stability criteria.

It was shown that the onset of instability can occur very differently: for some systems, the escape rate only increases slowly for varying parameters. Such systems were called unreliable because the existence of parameter regions with very low but positive escape rate makes instability very difficult to detect using simulations.

Chapter 4

Stability Analysis using Symbolic Dynamics

4.1 Introduction

This chapter introduces a method for stability analysis based on studies of the unstable limit cycles of a chaotic system. The basis for this analysis is the use of *symbolic dynamics*. Symbolic dynamics has previously been used for analysis of the general first-order modulator [15] and the use of symbolic dynamics for stability analysis was proposed by the author of [13].

Recall from Sec. 2.3 that a chaotic limit set or attractor can be perceived as the union of indefinitely many unstable limit cycles, i.e., the collection of limit cycles is the skeleton of the attractor or repeller. The question is obviously: how can all these unstable limit sets form a stable limit set? The use of symbolic dynamics can provide an answer to this question and, in addition, the escape rate can be expressed analytically for some systems.

4.2 Symbolic Dynamics

In this chapter the precise state-space description of Σ-Δ modulators will be replaced by the study of possible (i.e., admissible) binary output code sequences. This kind of analysis is often referred to as *symbolic dynamics* [11, 15]. The basics of this approach has in fact been outlined in Sec. 2.2: an initial condition in the state-space is mapped into an infinite sequence of binary symbols, i.e., the output code sequence. Instead of analyzing a modulator in the N-dimensional state-space, the dynamics of the system is revealed by the admissible binary output sequences. Using iterations of the system map $\mathcal{F}$, an initial condition $\boldsymbol{x}_0$ is mapped into a sequence $\mathrm{S}(\boldsymbol{x}_0) = \{s_0, s_1, s_2 ...\} \in \Sigma$ of binary symbols. The mapping $\mathrm{S}(\cdot)$ has the property that:

$$\mathrm{S}(\mathcal{F}(\boldsymbol{x}_0)) = \sigma(\mathrm{S}(\boldsymbol{x}_0)) \tag{4.1}$$

where $\sigma(\cdot)$ is the shift map on the set of infinite binary sequences Σ:

$$\sigma(\{s_0, s_1, s_2 ...\}) = \{s_1, s_2 ...\} \tag{4.2}$$

This means that the shift map $\sigma(\cdot)$ is homeomorphic (i.e., topologically transitive) to the system map $\mathcal{F}$. The limit cycles of $\mathcal{F}$ thus correspond to periodic binary sequences of $\sigma(\cdot)$.

The sequence mapping $\mathrm{S}(\cdot)$ does not generally map the state-space onto the entire set Σ of binary sequences, i.e., a modulator can normally only produce a subset of binary sequences. Furthermore the mapping is not always one-to-one, i.e., for some modulators, several initial conditions might produce the same binary sequence (this is the case for the usual double-loop modulator with two poles at $z = 1$). However, in most cases, when the binary output sequence is known for an infinite number of time steps, it is possible to uniquely reconstruct the actual state-space orbit producing the given code sequence. This useful fact was employed in Section 2.7 to identify the possible limit cycles corresponding to infinite periodic code sequences.

4.3 Analytical Determination of the Escape Rate

Consider a dynamical system described by a map $\mathcal{F}$ and a state space S. When a bounding set W is introduced the set Ω_n of initial points surviving n time-steps is defined:

$$\Omega_n = \{\boldsymbol{x} \in S | \mathcal{F}^l(\boldsymbol{x}) \in W \text{ for } 0 \leq l \leq n\} \tag{4.3}$$

The set Ω_n of surviving initial points can naturally be partitioned into (disjoint) subsets according to the first $n+1$ symbols $s_{0..n}$ of the binary sequences generated when starting inside Ω_n:

$$\Omega_n = \bigcup_{s \in \Sigma}^{(n)} \omega_{s,0..n} \tag{4.4}$$

where $\omega_{s,0..n}$ designates the set of initial conditions $\boldsymbol{x}_0$ giving an $(n+1)$-bit code sequence $s_{0..n}$ where $s = \mathrm{S}(\boldsymbol{x}_0)$. Hence, the set union of Eq. (4.4) is formed over maximally $2^{(n+1)}$ subsets. The structure of these subsets can be investigated recursively. First the bounding set W is split into two halves according to the code produced:

$$\begin{aligned} W_+ &= W \bigcap S_+ \\ W_- &= W \bigcap S_- \end{aligned} \tag{4.5}$$

where S_+ and S_- cf. Eq. (2.16) are the regions of state-space giving positive and negative output codes, respectively.

The $\omega_{s,0..n}$ subsets can then be found from the recursion:

$$\begin{aligned} \omega_{s,n..n} &= W_{s_n} \\ \omega_{s,n-l..n} &= \mathcal{F}^{-1}_{s_{n-l}}(\omega_{s,n-l+1..n}) \bigcap W_{s_l} \end{aligned} \tag{4.6}$$

where the notation s_l means the l'th symbol of the binary sequence s; s_l used as index distinguishes between W_+ and W_- and between $\mathcal{F}_+$ and $\mathcal{F}_-$. For each step of the recursion, the code sequence corresponding to s is followed one symbol backwards. The resulting subsets $\omega_{s,l..n}$ will shrink in volume as l increases for two reasons: the inverse maps $\mathcal{F}_+^{-1}$ and $\mathcal{F}_-^{-1}$ might be contracting in some directions and everything outside either the W_+ or W_- sets is cut away for each iteration. For some code sequences this may result in empty subsets, i.e., the code sequence $s_{0..n}$ is not admissible.

If the system has eigenvalues inside the unit circle, the inverse maps are expanding in some directions, but the repeated bounding by W_+ or W_- limits the possible expansion. Consequently, it is the eigenvalues λ_p outside the unit circle and the repeated intersection with the W_+ and W_- sets which determine the volume shrinking. Recall that a modulator with eigenvalues (i.e., feedback filter poles) outside the unit circle is chaotic.

The recursion cf. Eq. (4.6) is not practical for stability analysis. Instead of using an explicit expression of the volume of Ω_n it is assumed that the exponential volume shrinking due to the eigenvalues will dominate asymptotically [13] and, consequently, that for each iteration, the average volumes of the subsets are reduced by a factor

$$\Lambda = \prod_{|\lambda_p|>1} |\lambda_p| \tag{4.7}$$

where λ_p are the eigenvalues of the modulator.

From the definition of the escape rate Eq. (3.6) it is expected that the volume Φ_n of the survivors Ω_n is exponentially decreasing with n. Using the partitioning cf. Eq. (4.4) one arrives at:

$$\Phi_n = \sum_{s\in\Sigma}^{(n)} \phi_{s,0..n} \approx \Phi_0 e^{-\gamma n} \tag{4.8}$$

where $\phi_{s,0..n}$ is the state-space volume of the subset $\omega_{s,0..n}$.

This expression can be simplified by approximating the volumes $\phi_{s,0..n}$ by $1/\Lambda^n$:

$$\Phi_n = \beta \frac{\epsilon_n}{\Lambda^n} \approx \Phi_0 e^{-\gamma n} \tag{4.9}$$

where ϵ_n is the number of non-empty subsets $\omega_{s,0..n}$, β is an unknown scaling factor and γ is the escape rate.

Eq. (4.9) shows that as n increases, the growth of the number of non-empty subsets $\omega_{s,0..n}$ must counterbalance the growth of Λ^n for a stable system. If Λ^n grows faster than ϵ_n, the volume of survivors will decrease asymptotically, i.e., the system is unstable. According to the definition of the escape rate, the decrease in the volume of survivors is expected to be exponential. Consequently, Eq. (4.9) can provide an analytical expression for the escape rate γ.

Example 4.1 Consider the first order modulator discussed in Example 3.1 with zero input:

$$e_{n+1} = \mathcal{F}(e_n) = ae_n - \operatorname{sgn}(e_n)\,,\ a > 1 \tag{4.10}$$

It was previously established that this modulator is unstable for zero input when $a > 2$. Especially for $a = 2$ it is observed that both the unit intervals]0,1[and]-1,0[are mapped on the interval]-1,1[. This means that both values of the succeeding output code of the system can be generated irrespectively of the current code, i.e., there are four intervals of initial points giving the four code sequences (00), (01), (10) and (11). Similarly, eight intervals exist giving every of the eight binary sequences of length three. As a result any code sequence can be generated by choosing a proper initial point in the interval]-1,1[and, consequently, any limit cycle exists. This property is preserved when $a > 2$. This fact is sufficient to prove that the modulator is chaotic according to the mathematical definition of chaos [30, 15]. The reason is that the shift map $\sigma(\cdot)$ of Eq. (4.2) is chaotic on Σ, i.e., the entire set of binary symbol sequences.

Since the $\mathcal{F}$ map is expanding by a factor $\Lambda = a$, it is expected that the average length of the subset $\omega_{s,0..n-1}$ giving a bounded orbit and a particular code sequence $s_{0..n}$ is proportional to $1/a^n$ and by summing over the 2^n possible code sequences with length n, one gets:

$$\Phi_n = \sum_{s}^{(n)} \phi_i = \beta \frac{2^n}{a^n} = \beta e^{-n \ln(a/2)} \tag{4.11}$$

where β is an unknown scaling factor.

The equation shows that the escape rate is $\gamma = \ln(a/2)$. For a near 2, the approximation $\gamma \approx a/2 - 1$ can be used □

Example 4.1 allowed an analytical expression for the escape rate due the fact that every possible limit cycle exists. This is indeed not a general property of Σ-Δ modulators; normally only a fraction of the possible periodic code sequences exist as limit cycles. The problem is thus to estimate the number ϕ_n of non-empty subsets $\omega_{s,0..n}$ as a function of the iteration number n.

A modulator is unstable if ϵ_n grows slower than Λ^n or equivalently: the growth of ϵ_n must outweigh the growth of Λ^n for a stable system. The question about the stability of a modulator is now turned into a matter of counting admissible sequences and as it will appear, this is almost the same as counting limit cycles.

Some of the $\omega_{s,0..n}$ subsets have binary sequences which exist as limit cycles of length n and the corresponding point on the limit cycle will be contained in the subset. Other subsets have binary sequences which are parts of limit cycles with length larger than n and some of the subsets have binary sequences which do not exist as limit cycles. The subsets are centered around points which after a number of iterations hit the chaotic limit set on a point of a limit cycle entering the noninvertible region, i.e., the centers are points that eventually become periodic. Finally, some subsets are empty, i.e., these subsets have sequences which are not admissible and therefore do not exist as limit cycles.

It is expected that the growth of the number ϵ_n of admissible binary sequences $s_{0..n}$ is proportional to the growth of the number l_n of period n points due to the close relationship between the $\omega_{s,0..n}$ subsets and the limit cycles of the system.

The limit cycles of an unstable chaotic system is the skeleton of the chaotic limit set, i.e., the limit cycles up to a certain length n is a kind of approximation to the usually very complex limit set [13]. As n increases, the approximation gets better and more detailed.

Example 4.2 The limit cycles of the first order modulator from Example 4.1 reveals the unstable chaotic limit set L for $a > 2$. Fig. 4.1 shows the limit cycles up to length 12 plotted for $a = 2.3$ where $e(k)$ is plotted as a function of $e(k-1)$. Hence, every plotted point is on the graph for the $\mathcal{F}$ map which schematically is shown in Fig. 3.1. The limit set L is fractal and is a so-called *Cantor* set. Recall from Example 3.1 that points outside the interval between the fixed points at $e^* = \frac{\pm 1}{a-1} \approx \pm 0.77$ iterates towards infinity. As an initial approximation the surviving points must be in this interval; however, in the middle of this interval there is another interval of points which are mapped

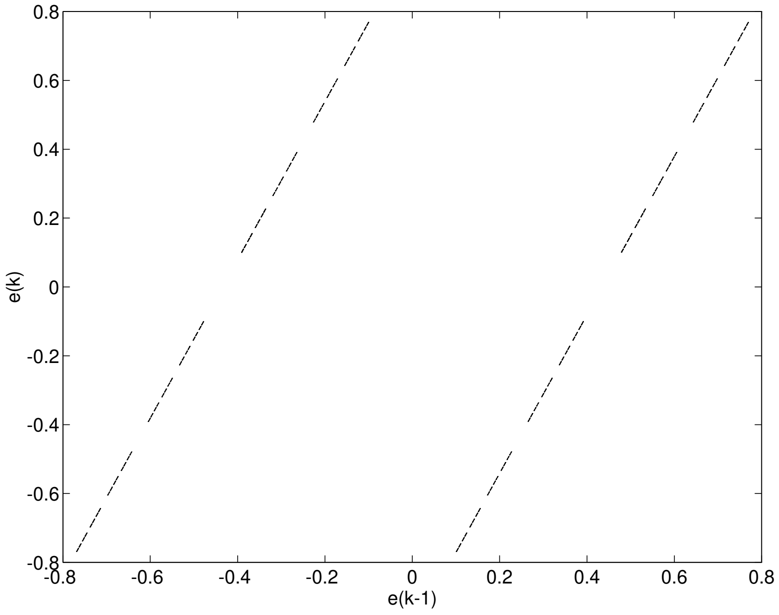

Figure 4.1: Limit cycles of period up to 12 for the modulator cf. Eq. (4.10) with $a = 2.3$.

> outside the fixed points. The next approximation to the set of surviving points is to delete the middle interval leaving two intervals which survive one iteration. Continuing this procedure, the intervals are split into two intervals while a middle interval is deleted. Every time the same fractions of the intervals are deleted. The final result is a Cantor set which can be approximated by the set of periodic points up to a suitable length. See also [15] for a limit cycle analysis of the chaotic first order modulator □

4.4 Counting Limit Cycles

The number l_n is the number of the 2^n periodic sequences of length n which exist as limit cycles. At first hand it seems necessary to test every 2^n for existence. Recall from Sec. 2.7 that the modulator loop is cut open and the periodic steady-state quantizer input is determined for each periodic binary sequence. The corresponding limit cycle exists if the quantization of the quantizer input matches the binary sequence. This makes it obviously unnecessary to test cyclic shifts of a given binary sequence since the result is the same. Furthermore, if a binary sequence has a prime period which is lower than n, the corresponding limit cycle retraces itself several times and has fewer than n distinct points. The limit cycle investigations can thus be restricted to a number of so-called *prime cycles*. A prime cycle of period n is a binary periodic sequence with prime period n, in addition, cyclic shifts of a prime cycle are equivalent. If a period n prime cycle exist as limit cycle it will consist of n distinct points. Both the fixed points $\overline{1}$ and $\overline{0}$ are prime cycles of length one. The only prime cycle of length two is $\overline{10} = \overline{01}$ since the cycles $\overline{11}$ and $\overline{00}$ are repetitions of the two prime cycles of length one. Let p_k be the number of prime cycles with length k. The total number of binary sequences with period n is 2^n which

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
p_k	2	1	2	3	6	9	18	30	56	99	186	335	630	1161	2182	4080

Table 4.1: The number p_k of prime cycles of length k for $k = 1..16$

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
p_k	2	1	0	1	2	0	4	3	6	7	6	10	20	28	32	41

Table 4.2: The number p_k of existing prime cycles of length k for $k = 1..16$ for the modulator cf. Eq. (3.5) with zero input and $h_1 = 0.8$, $h_2 = -1.5$.

must equal the sum of kp_k for every divisor k in n:

$$2^n = \sum_{k \uparrow n} kp_k \tag{4.12}$$

From this equation p_k can be found recursively. Only the prime cycles with period $k \uparrow n$ has to be examined and this reduces the computational burden significantly. Table 4.1 shows p_k for $k = 1..16$. If all periodic points with period, e.g., 10 have to be found, only $p_{10} + p_5 + p_2 + p_1 = 108$ prime cycles have to be tested. The brute force method requires $2^{10} = p_1 + 2p_2 + 5p_5 + 10p_{10} = 1024$ limit cycles to be tested.

Example 4.3 Recall the chaotic second order modulator cf. Eq. (3.5) with $h_1 = 0.8$ and $h_2 = -1.5$. It is easily shown that the product of the pole moduli is equal to $|h_2|$. Consequently, $\Lambda = |h_2|$ for $|h_2| > 1$.

The prime cycles up to length 16 was tested for existence for the system with zero input. Table 4.2 shows the number of prime cycles found.

The number l_n of period n points is given as:

$$l_n = \sum_{k \uparrow n} kp_k \tag{4.13}$$

where p_k are the existing prime cycles of length k.

Fig. 4.2 shows l_n plotted versus n on semilog. scale for the second order modulator example. The dotted line shows the function Λ^n where $\Lambda = |h_2| = 1.5$. It is seen that l_n approximately follows this exponential growth.

Fig. 4.3 shows every periodic point found with period n up to 16. Compare the plot to Fig. 3.3 and Fig. 3.10. The structure of the stable limit set L as well as the unstable repeller on the boundary of the basin of attraction, ∂B_L, can easily be recognized. The set of periodic points can be partitioned into points on L and ∂B_L. The number of periodic points on ∂B_L is obviously growing much slower than Λ^n since these limit cycles form an unstable repeller (see also Example 3.6 for the escape rate). This means that the number of periodic points in L will dominate asymptotically when the total number of periodic points is estimated □

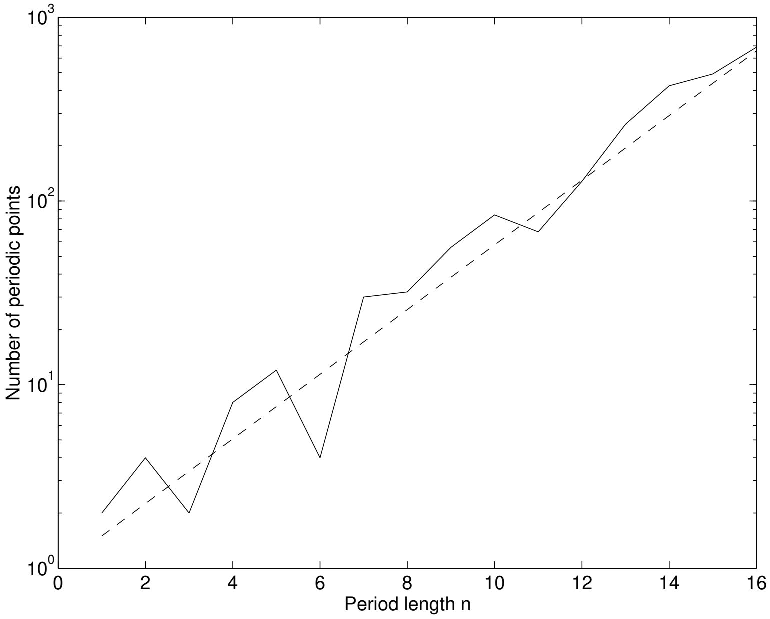

Figure 4.2: Number of periodic points l_n with period up to 16 versus the period length n for the modulator cf. Eq. (3.5) with with $h_1 = 0.8$, $h_2 = -1.5$ and zero input. The dashed line corresponds to the function $\Lambda^n = |h_2|^n$.

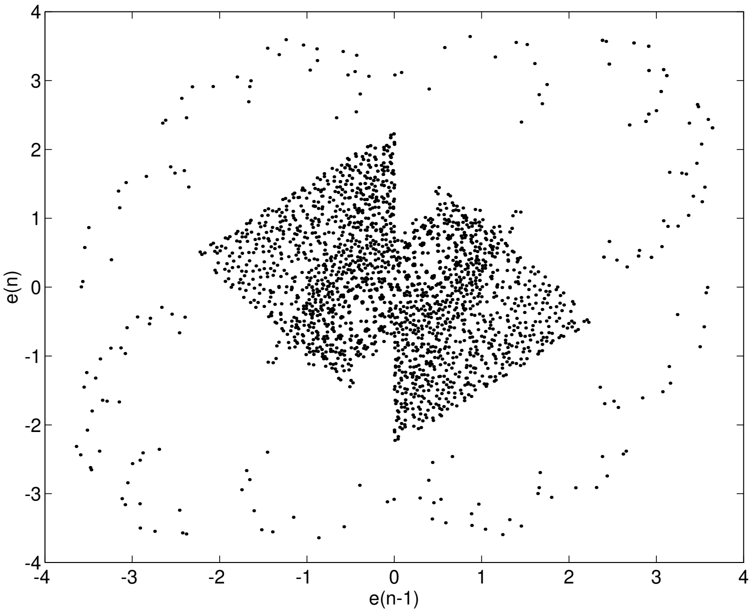

Figure 4.3: Limit cycles with period up to 16 for the modulator cf. Eq. (3.5) with $h_1 = 0.8$, $h_2 = -1.5$ and zero input.

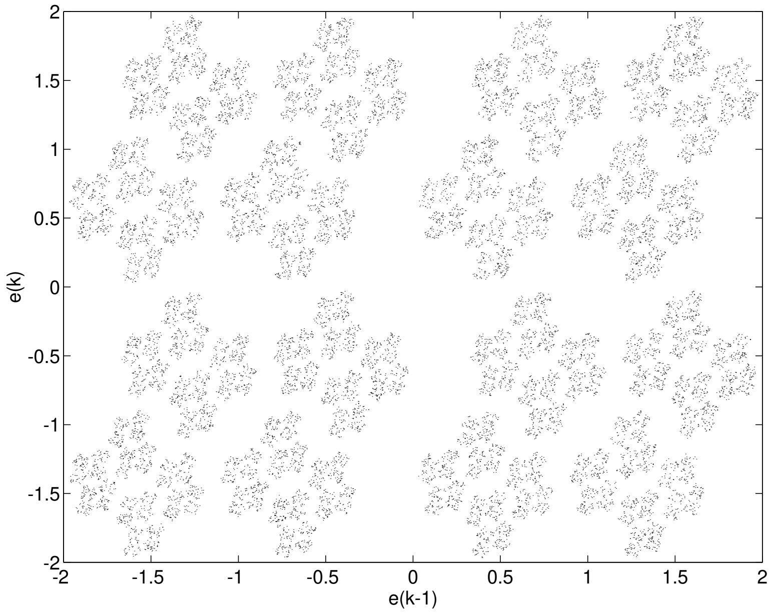

Figure 4.4: Reverse time simulation of the system cf. Eq. (4.14) with zero input.

Example 4.4 The first order system from Example 4.1 was an example of a system which has every of the maximal 2^n period n points. A necessary condition for this property is that the expansion factor $\Lambda = a$ must be greater than two. This example shall demonstrate that there also exists second order modulators with every limit cycle existing. Recall the system from Example 4.3 with a $H(z)$ given by:

$$H(z) = \frac{h_1 z^{-1} + h_2 z^{-2}}{1 - h_1 z^{-1} - h_2 z^{-2}} \tag{4.14}$$

where $h_1 = 0.8$ was used. When h_2 is decreased below approx. 2.25 every limit cycle exists. This can be verified efficiently by testing prime cycles alone up to a suitable length. For $h_2 = -2.3$ the system has every prime cycle at least up to $n = 16$, i.e., every of the prime cycles corresponding to Table 4.1.

Recall that the previously described reverse time simulation technique (see Sec. 3.7) can be used to approximate the limit set of an unstable chaotic system. This technique is usually much faster than finding and plotting every prime cycle. If the limit set found by reverse time simulation is fully inside the noninvertible region U, each reverse time step has two possible branches reflecting the fact that every code sequence is possible or, equivalently, that every limit cycle exists. Recall also that if the quantizer input fulfills $|e(k)| < -2H(z = 0) = 2$ for every k, then the orbit will be inside U. Consequently, the reverse time simulation technique can also be used to show that every limit cycle exists: Fig. 4.4 shows a reverse time simulation for $h_2 = -2.3$. Observe that $e(k)$ has a magnitude less than 2 indicating that every limit cycle exists. The plotted set has a very regular and self-similar Cantor structure.

When every limit cycle exists the escape rate can be determined analytically, cf. Eq. (4.9):

$$\gamma = \ln(\Lambda/2) = \ln(|h_2|/2) \tag{4.15}$$

□

4.5 Considerations for Nonhyperbolic Modulators

So far only modulators with one or more poles outside the unit circle have been considered. It was shown that the volume of the $\omega_{s,0..n}$ subsets are expected to shrink exponentially with n. In this section nonhyperbolic modulators will be investigated. These modulators have all feedback filter poles on the unit circle.

The traditional second-order modulator has a double pole at $z = 1$ corresponding to zero frequency. The corresponding feedback filter is usually constructed as a cascade of two discrete-time integrators. In order to investigate the stability of modulator limit cycles, a simple linear dynamical system is considered:

$$\boldsymbol{x}_{k+1} = \boldsymbol{A}\boldsymbol{x}_k, \text{ where } \boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \tag{4.16}$$

The system has a multitude of fixed points on a line in the eigenvector direction through $\boldsymbol{x} = 0$, i.e., $[\ x^* \quad 0\ ]^\top = \boldsymbol{A}[\ x^* \quad 0\ ]^\top$. It is easily seen that the powers of $\boldsymbol{A}$ are gives as:

$$\boldsymbol{A}^n = \begin{bmatrix} 1 & n \\ 0 & 1 \end{bmatrix} \tag{4.17}$$

Let $W = \{\boldsymbol{x} \in \mathrm{I\!R}^2 | \|\boldsymbol{x}\|_\infty < 1\}$ be a bounding set, i.e., both coordinates must have a magnitude less than unity. The set Ω_n cf. Eq. (4.3) of points giving bounded orbits for n iterations are then given by:

$$\Omega_n = \{[\ x_1 \quad x_2\ ]^\top \in \mathrm{I\!R}^2 \mid |x_1| < 1 \wedge |x_1 + nx_2| < 1\} \tag{4.18}$$

This set is shown schematically in Fig. 4.5 for $n > 1$. It is observed that the area of Ω_n is proportional to $1/n$, i.e., Ω_n is shrinking for increasing n. It can be concluded that every fixed point $\boldsymbol{x}^*$ is unstable since no open neighborhood of points asymptotically attracted to the fixed point exists.

Second-order Σ-Δ modulators sharing the transition matrix $\boldsymbol{A}$ of Eq. (4.16) have also the same stability properties as the simplified system cf. Eq. (4.16). Each periodic binary sequence will either not exist as a limit cycle or exist as a multitude of limit cycles in state-space with periodic points on whole line segments in the eigenvector direction. Around each 'periodic line segment' corresponding to the sequence s, the subset $\omega_{s,0..n}$ will approximately look like Ω_n in Fig. 4.5. Consequently, every limit cycle is unstable and, furthermore, there is sensitivity to initial conditions: perturbations which are not in the eigenvector direction will grow linearly with time and eventually change the output code generated. It is also expected that the volume of $\omega_{s,0..n}$ (i.e., the area in 2-dim.) is proportional to $1/n$. In order to be stable the modulator must thus have a number l_n of period n points which at least is growing proportional to n.

Generally, it is expected that Nth order modulators with N coincident poles at $z = 1$ has a number of periodic points that increases as n^{N-1}.

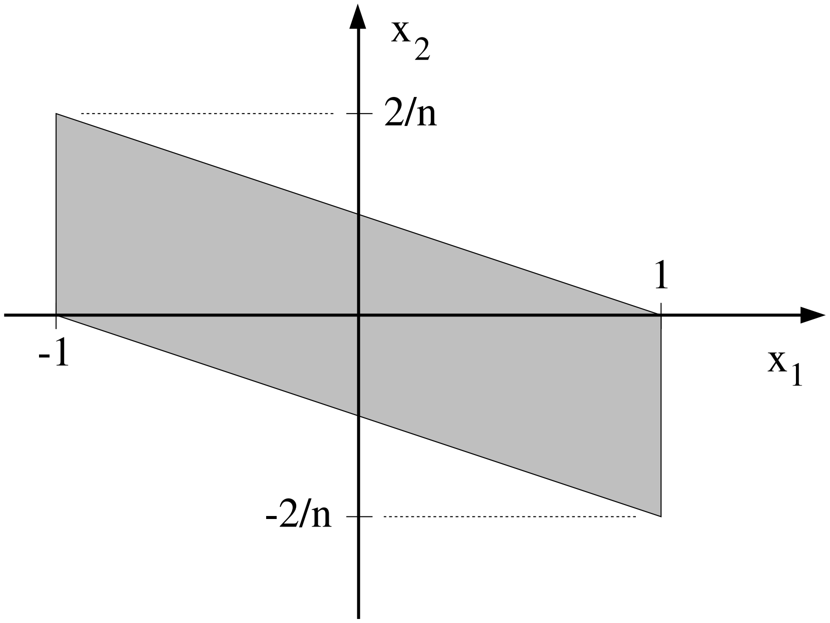

Figure 4.5: The hatched area is the set of survivors Ω_n for the nonhyperbolic system cf. Eq. (4.18)

This special class of modulators has thus two of the three conditions for being chaotic: sensitivity to initial conditions and a large and dense set of periodic points. However, the state-space divergence is linear and not exponential as for modulators with one or more poles outside the unit circle. Furthermore, simulations show that typical orbits with constant (zero) input will not 'fill out' a whole limit set but will follow one or a few limit cycles. Only if the system is perturbed by some small amplitude input noise, the orbits will reveal a limit set that looks like the union of the periodic points [68]. To be more specific, the second-order modulator is not topologically transitive. Consequently, the second order modulator with double poles at $z = 1$ cannot be characterized as being chaotic.

Modulators with distinct poles on the unit circle have limit cycles which are not repelling and nor attracting. A linear dynamical second-order system similar to Eq. (4.16) with distinct unit circle poles (i.e., eigenvalues) have orbits which are lying on closed ellipses. Hence, the Ω_n-sets have asymptotically constant volume. It is therefore expected that the number of periodic points is asymptotically constant for modulators with distinct unit circle poles.

4.6 Summary

The stability of Σ-Δ modulators was investigated using an analysis of the binary output code: the set of surviving initial conditions was partitioned according to the code sequence produced. The volume of each of the partitions was investigated using the eigenvalues of the system. For chaotic modulators, the conclusion was that the number of periodic points for a stable system must grow as fast as Λ^n where n is the period length and Λ is the expansion factor, i.e., the product of pole moduli greater than unity. If the number of periodic points grows slower, the system is unstable. This fact enabled the escape-rate γ

to be derived analytically for some systems.

The traditional double-loop modulator with a double pole at $z = 1$ was shown to have a linear volume expansion and it was argued that this modulator is not chaotic. For systems with distinct poles on the unit circle, the system map is asymptotically volume preserving and such systems are expected to have an asymptotically constant number of periodic points.

The presented framework for stability analysis thus explains why chaotic systems can produce more complex output code patterns due to the larger number of limit cycles.

Chapter 5

The Stability of Non-chaotic Modulators

5.1 Introduction

As described in Sec. 3.1 the the stability properties for non-chaotic modulators is somewhat different than for chaotic modulators. Recall that the limit cycles are attracting when every eigenvalue is inside the unit circle. Consequently, the system cannot be unbounded but will asymptotically end on a limit cycle which depends on the initial conditions. For modulators with orders higher than two and poles close to but inside the unit circle there is an amplitude gap separating two categories of limit cycles: 'stable' limit cycles with small amplitude and 'unstable' limit cycles with high amplitude oscillating at a low frequency. The latter have very poor encoding properties, i.e., the output code is a bad approximation to the input signal. Furthermore, these 'unstable' limit cycles are usually of the type $(\overline{r,s})$, i.e., r positive symbols followed by s negative symbols. The entire state-space can be split into the two disjoint sets B_S and B_U corresponding to points which are attracted asymptotically to 'stable' and 'unstable' limit cycles, respectively. The designations granular cycles and overload cycles will be used henceforward in order to avoid confusion with the usual limit cycle stability concept.

The amplitude of the overload cycles has been predicted using a describing function approximation in [25]. As a rule of thumb, the amplitudes of the overload cycles become infinite as the the poles approaches the unit circle. It has been suggested in [25] that the stability problem should be remedied by reducing the pole moduli in order to reduce the amplitude of the overload cycles to an acceptable level. However, this approach reduces the signal-to-noise ratio [25].

5.2 Third Order Examples

Throughout the rest of this chapter a certain class of third order modulators will be investigated by simulations. The feedback filter poles of this class are fixed and have all moduli equal to 0.975. The zeros constitute a complex pair with moduli given by the parameter m and zero frequency at the parameter v:

$$H(z) = \frac{z^{-1}(1 - me^{jv\pi}z^{-1})(1 - me^{-jv\pi}z^{-1})}{(1 - 0.975z^{-1})(1 - 0.975e^{0.085j}z^{-1})(1 - 0.975e^{-0.085j}z^{-1})} \tag{5.1}$$

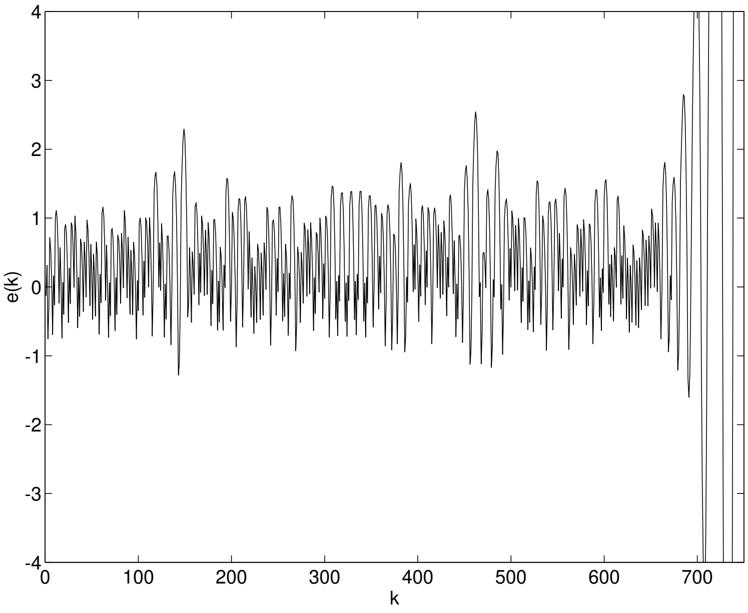

Figure 5.1: Transient leading to an overload cycle for the modulator cf. Eq. (5.1) with $m = 0.7$, $v = 0.09$ and constant input 0.4. Quantizer input $e(k)$ is plotted versus k

Modulators with this type of feedback filter will be studied with constant input. The two categories of limit cycles can easily be distinguished since the overload cycles for these modulators have usually quantizer input magnitudes exceeding 100 while the magnitudes for granular cycles rarely exceeds 3.

One common feature of high-order non-chaotic modulators seems to be the existence of very long transients before the system decides which limit cycle to be attracted to. Especially the transients ending on overload cycles are very interesting: the system may operate normally with tight bounds on the quantizer input for many hundreds of time steps until it suddenly blows up an ends on an overload cycle. This kind of behavior is very similar to the behavior of unstable chaotic modulators.

Example 5.1 An example of a long transient is shown in Fig. 5.1 and Fig. 5.2 . The parameters are $m = 0.7$ and $v = 0.09$ cf. Eq. (5.1). A constant input of 0.4 was used. The initial conditions for this example were selected from a large number of simulations with random initial conditions □

The existence of long transients which suddenly blows up could lead to the assumption that the system has an exponential decay of survivors similar to unstable chaotic modulators. The next example will show that this is not true and that non-chaotic modulators in fact are very complex to describe with regards to stability.

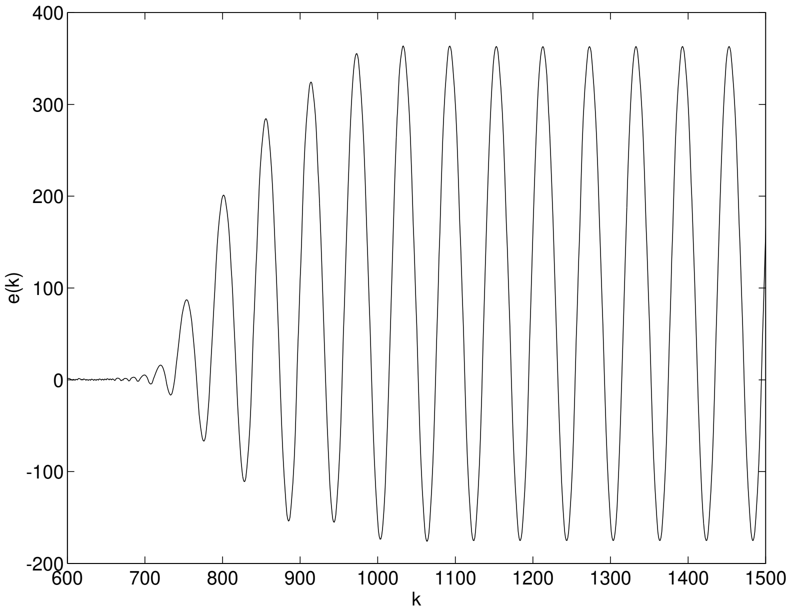

Figure 5.2: The onset of an overload cycle for the modulator cf. Eq. (5.1) with $m = 0.7$, $v = 0.09$ and constant input 0.4. The plot is the same as in Fig. 5.1 but with different axes.

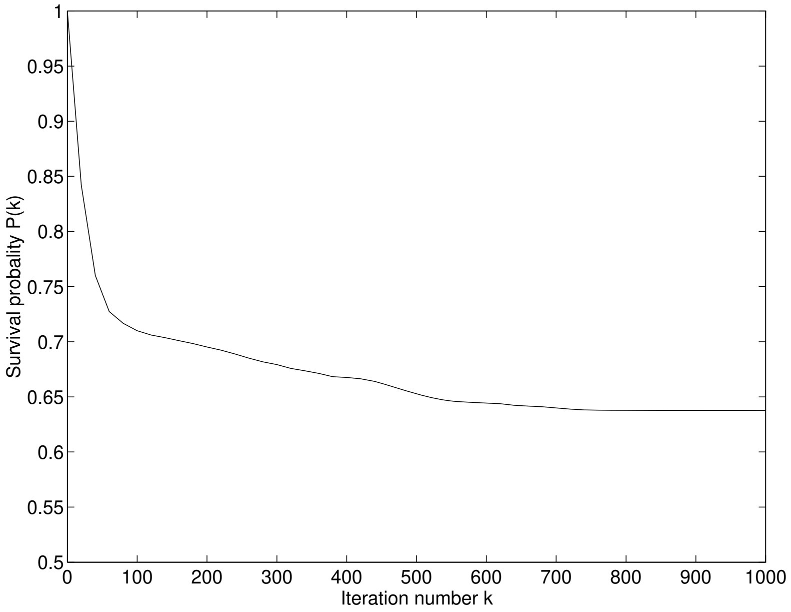

Figure 5.3: Survival plot for the modulator cf. Eq. (5.1) with $m = 0.7$, $v = 0.09$ and constant input 0.4.

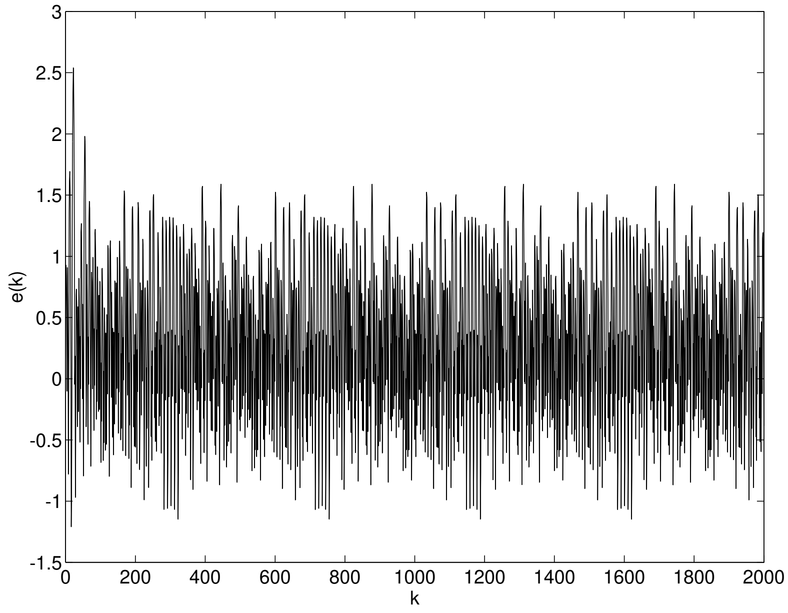

Figure 5.4: Example of a granular limit cycle.

Example 5.2 The modulator of Eq. (5.1) with $m = 0.7$, $v = 0.09$ and input 0.4 was simulated with a large number of random initial conditions. The fraction of orbits $P(k)$ staying below a suitable bound is plotted versus the iteration number k in Fig. 5.3 with linear scales. Asymptotically, a large fraction (approx. 64%) of the initial conditions are attracted to granular limit cycles with low amplitudes. An example of such limit cycle with period approximately equal to 500 is shown in Fig. 5.4.

The existence of long transients causes the $P(k)$ curve of Fig. 5.3 to approach its asymptotic value very slowly. Transients leading to overload cycles as long as 800 iterations can be found.

This example is quite strange: most of the initial conditions leads to 'stable' limit cycles while the rest causes the system to blow up after transients of varying length.

The system is very sensitive to the input: for constant input 0.5 every initial state is asymptotically attracted to overload cycles. The survival plot of Fig. 5.5 shows the existence of very long transients. The $P(k)$ curve seems to approach zero almost linearly with the iteration number k. This fact indicates that the transients are uniformly distributed in length up to a certain maximum length which is approx. 1300 for this example. Furthermore, the 'stable' basin of attraction B_S has shrinked to zero volume indicating that only overload cycles exist for constant input 0.5 □

Σ-Δ modulators with feedback filter poles close to $z = 1$ have a very high dc-gain. This means that the mean value of the modulator input must be very close to the mean value of the output code. Consequently, a given limit cycle will only exist for a small interval of constant inputs. For the extreme case where the feedback filter has one or more poles at

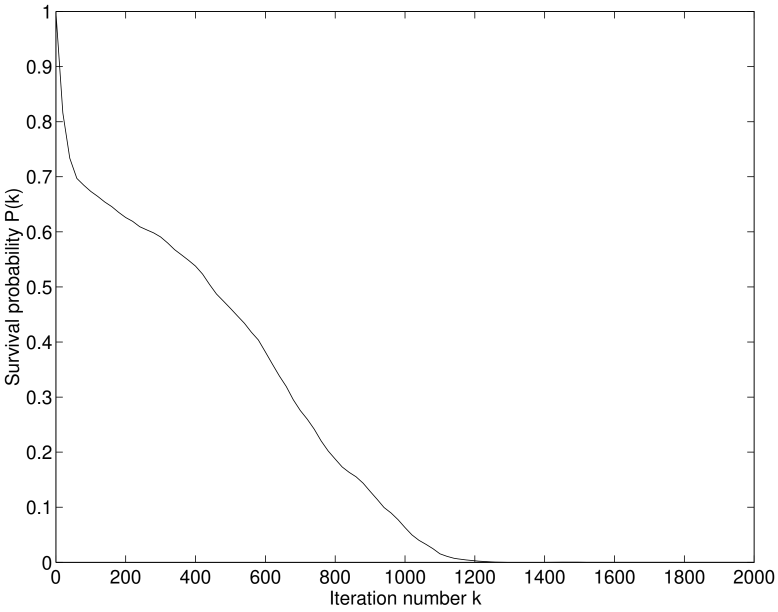

Figure 5.5: Survival plot for the modulator cf. Eq. (5.1) with $m = 0.7$, $v = 0.09$ and constant input 0.5.

$z = 1$, the dc-gain is infinite and a periodic binary sequence can only exist as a limit cycle for a constant input equal to the mean value of the binary sequence.

A high dc-gain can thus explain the sudden change in stability for some critical values of constant modulator input: some constant input values yield only few or no granular 'stable' limit cycles. The overload cycles are generally more robust due to their much higher quantizer input amplitudes; a dc-shift of the quantizer input is less likely to change the output code for overload cycles.

Example 5.3 The third order modulator used in the two preceding examples (i.e., cf. Eq. (5.1) with $m = 0.7$ and $v = 0.09$) was simulated with 512 constant input values equally spaced between 0 and 0.6. For each input value 1000 simulations was performed with random initial conditions; a maximum simulation length of 2000 was used. Fig. 5.6 shows the length of the longest unstable transient found for each input value and Fig. 5.7 shows the fraction of initial conditions leading to granular cycles, i.e. the fraction of survivors. The plots show that the long unstable transients only exist for small intervals of input. For most other input values the longest unstable transient has a length of approx. 75 time steps and this background level is observable up to high input values. The first window with long transients is found for constant input near 0.0528. Later for inputs around 0.11 a whole cluster of windows with long transients is found. A finer resolution of constant inputs would probably reveal more details and even longer transients.

There is a clear correlation between longer transients and smaller fractions of survivors. However, long unstable transients means generally not that the

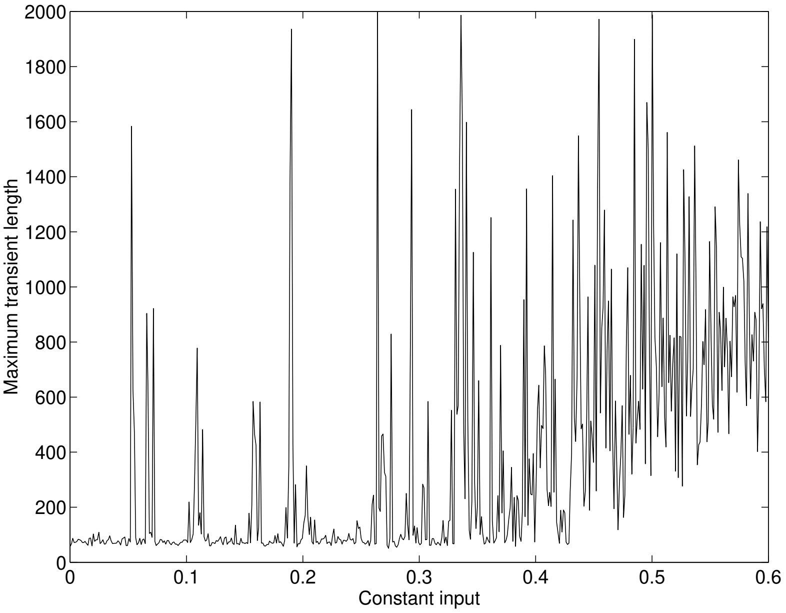

Figure 5.6: Maximum length of unstable transients versus constant input for the third order modulator cf. Eq. (5.1) with $m = 0.7$ and $v = 0.09$.

fraction of survivors become zero. This is only the case for some discrete input values □

The modulator from the previous example has indeed a very complex behavior that only reveals itself when when the constant input space is sampled with high resolution. The presence of very long unstable transients indicates that the structure of the two basins of attraction B_S and B_U are very complex and that these two sets might be 'mixed' down to small scales, i.e., points from B_U can be very close to the granular 'stable' limit cycles. This phenomenon is easily shown by simulation with constant input combined with a white noise signal: the system blows up even for small noise amplitudes when the constant input gives long transients for unperturbed input. small amplitude. An alternative explanation to this observation is that there might always be an input value for which no granular cycles exist in the immediate vicinity of input values showing long transients. The use of input noise will then eventually 'find' this critical input value which otherwise is hard to find when the entire input range is sampled by simulations.

The modulator from the previous example is therefore very unreliable due the many windows with long transients existing even for small constant inputs. Simulations with sinusoidal input confirms that instability arises even for small input amplitudes for this modulator.

The concept of unreliable modulators seems therefore also to make sense for non-chaotic modulators. The next example will show that a far more reliable system is obtained by perturbing the parameters from the previous example slightly.

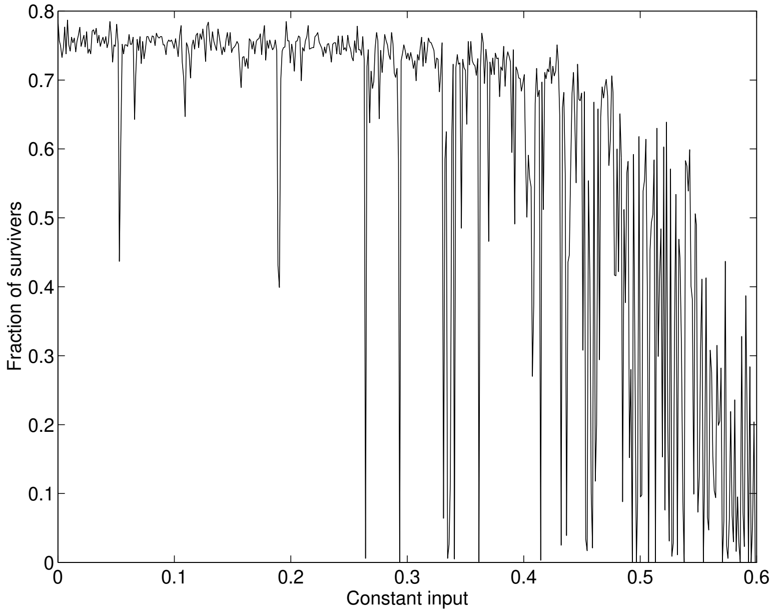

Figure 5.7: The fraction of survivors (i.e. initial conditions leading to granular cycles) for the modulator cf. Eq. (5.1) with $m = 0.7$, $v = 0.09$.

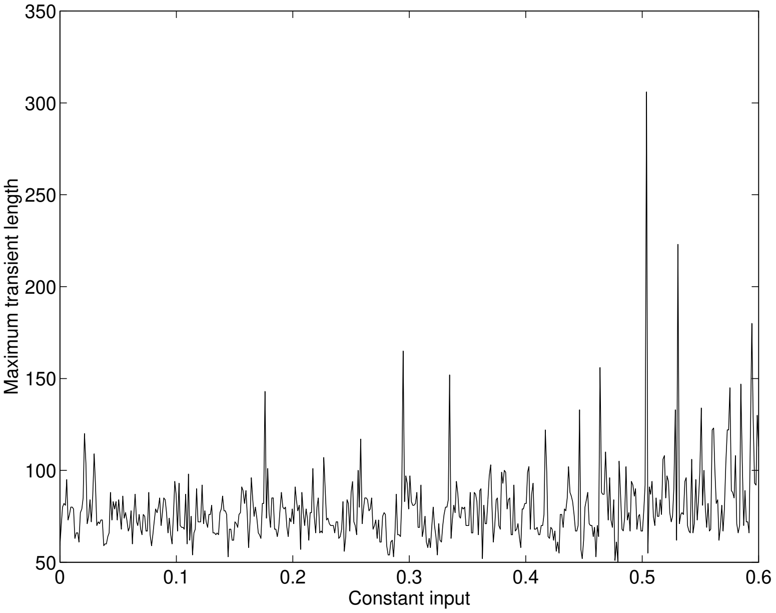

Figure 5.8: Maximum length of unstable transients versus constant input for the third order modulator cf. Eq. (5.1) with $m = 0.75$ and $v = 0.09$.

Example 5.4 The feedback filter zero moduli m as defined in Eq. (5.1) can be increased from 0.7 to 0.75 while the normalized zero frequency v is still equal to 0.09. This perturbation enhances the stability of the third order modulator significantly: Fig. 5.8 shows the maximum transient length versus constant input with the same simulation parameters as for the previous plots (i.e. 512 dc-inputs with 1000 simulations up to length 2000). The plot shows the same background level of maximum transients of approx. 75 iterations. The input range up to 0.5 is almost free from windows with long transients; the maximum transient is shorter than 200 time steps. This modulator is far more reliable and useful for practical purposes □

5.3 Comparison with Chaotic Modulators

The previous section clearly demonstrated that the absence of chaos does not mean that the dynamics of a Σ-Δ modulator becomes simpler to understand. The chaotic modulator shows in fact much more regularity: it is either unstable or stable. Furthermore, the escape rate γ quantifies the degree of instability for an unstable system and this parameter tells how often instability will occur in practice. The nonchaotic modulators in the previous section showed that parameter values exist for which the majority of initial conditions are attracted to granular low amplitude cycles while the rest are attracted to overload cycles after very long transients. For these parameter values even small perturbations of the input was enough to destabilize the system. Consequently, nonchaotic modulators cannot be classified simply as either stable or unstable for given parameters.

Nonchaotic modulators with many constant input windows yielding long transients are unreliable for practical purposes: it is very likely that a varying input signal eventually will activate an overload cycle. Simulations have proven that high-order chaotic modulators also may suffer from a similar kind of unreliability: the range of constant input values may have many narrow windows where the modulator is slightly unstable, i.e., the escape rate is small but positive.

Example 5.5 The feedback filter poles of Eq. (5.1) was mirrored outside the unit circle, i.e., the reciprocal pole locations were used in order to get a chaotic modulator. The maximum transient length versus constant input was found and shown in Fig. 5.9 using exactly the same procedure as used for Example 5.3 and Example 5.4, i.e., 512 dc-inputs with 1000 simulations up to length 2000. The parameter values $m = 0.9$ and $v = 0.11$ were used.

The plot shows that there again is a characteristic background level of transients with maximum length near 75. The first spike with long transients is found for constant input near 0.09 and later near 0.1. Several additional spikes are found in the input range up to 0.2 where the background level disappears and the curve approaches the maximum value of 2000.

This kind of maximum transient plot requires some interpretation in the unstable chaotic case: It is expected that the transient length has an exponential probability distribution with parameter γ when the initial conditions are picked randomly. The plot shows thus the maximum value of a random variable out of 1000 experiments with the same exponential distribution. The expected

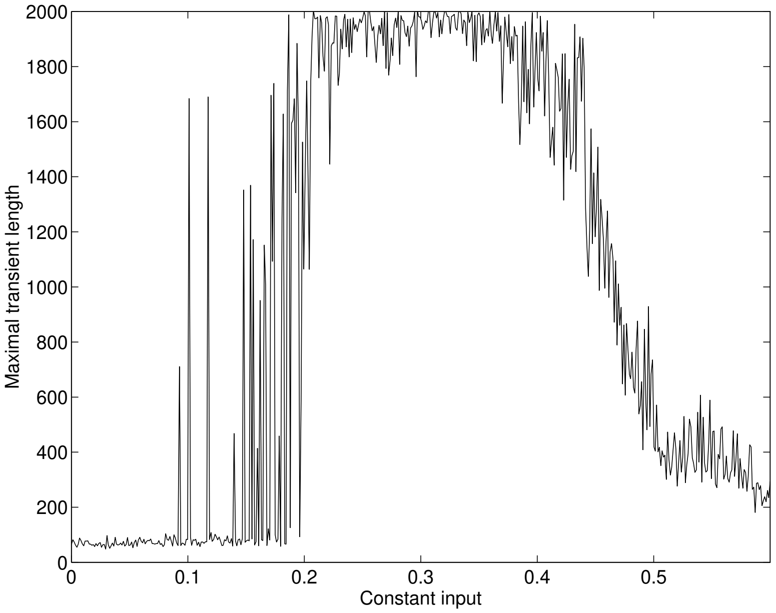

Figure 5.9: Maximum length of unstable transients versus constant input for the third order modulator cf. Eq. (5.1). The poles are mirrored outside the unit circle in order to get a chaotic system and the parameters $m = 0.9$ and $v = 0.11$ were used.

maximum value will be small both for very small and very large escape rates. The expected maximum value will thus have a maximum for medium escape rates. This fact explains why the plot in Fig. 5.9 is declining for inputs above 0.5 and this is not because the modulator becomes more stable □

5.4 Summary

Nonchaotic Σ-Δ modulators will always be asymptotically attracted to limit cycles. High-order modulators have both high amplitude overload cycles as well as low amplitude granular cycles. The state-space can thus naturally be partitioned into two basins of attraction corresponding to the two categories of limit cycles.

It was shown that nonchaotic modulators cannot be characterized simply as stable or unstable and that very long unstable transients may exist for certain parameter values. In addition, both chaotic and nonchaotic high-order modulators with a high dc feedback gain can be strongly unreliable: critical input values may exist for which the modulator is practically unstable. This kind of behavior reveals it self when the state-space and the constant input space is searched carefully for long unstable transients.

Part II

Modeling, Design and Optimization

Chapter 6

Quasilinear Modeling

6.1 Introduction

This chapter is devoted to the use of linearized models of Σ-Δ modulators. This is in fact the classical way to analyze such non-linear systems. The advantage of linearized analysis is that the modulator performance (e.g., the in-band noise power) can be predicted fairly accurately. The dynamical system description does not naturally provide such information. Traditionally, the linearized model has not been very useful for understanding instability properties. However, this chapter presents some extensions and new interpretations of the commonly used linearized models which leads to a practically useful stability criterion.

The linearized analysis also offers a very simple and yet efficient method for the design of feedback filters. This chapter reviews this method and uses the resulting filters as examples to demonstrate the accuracy of the linearized models.

6.2 Loop Analysis with Linearized Quantizer Model

The generic modulator of Fig. 2.4 contains a highly nonlinear circuit element, i.e., the one-bit quantizer or signum function. Fig. 6.1 shows a generic Σ-Δ modulator where the quantizer is replaced by a linear model, i.e., the one-bit quantizer is modeled as a gain by the factor K followed by the addition of a quantization noise source $q(k)$. The feedback filter $H(z)$ must have at least one sample delay in order to give a realizable system. The

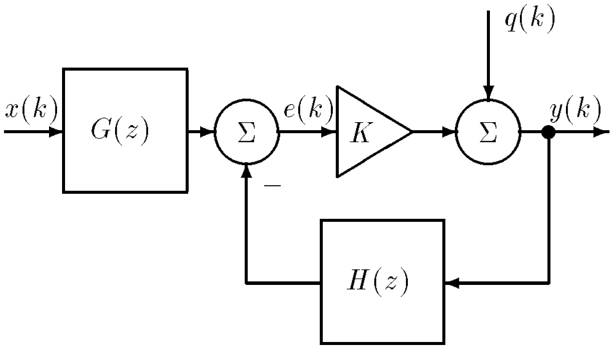

Figure 6.1: Generic Σ-Δ modulator with linearized quantizer model

linearized quantizer modeling turns the highly nonlinear modulator into a linear system with one output $y(k)$ and two inputs, i.e., the input signal $x(k)$ and the quantization noise $q(k)$ which accounts for the nonlinear effect of the one-bit quantizer. Note that this model is fully correct for any value of K — no approximations or assumptions have been made so far. The problem is now that the quantization noise $q(k)$ is unknown and depends on both the choice of K, the input signal and the initial conditions, i.e, it reflects the possibly chaotic non-linear dynamics of the system.

It is now possible to identify two transfer functions from the two inputs to the output, namely the *signal transfer function* $\mathrm{STF}_K(z)$ and the *noise transfer function* $\mathrm{NTF}_K(z)$ as defined by the z-domain equation:

$$Y(z) = \mathrm{STF}_K(z)X(z) + \mathrm{NTF}_K(z)Q(z) \tag{6.1}$$

where $Y(z)$, $X(z)$ and $Q(z)$ are the z-transforms of $y(k)$, $x(k)$ and $q(k)$, respectively. The K-indices indicate that the linearized transfer functions generally depends on K interpreted as a parameter.

It can be derived from from Fig. 6.1 that $\mathrm{STF}_K(z)$ and $\mathrm{NTF}_K(z)$ are given by:

$$\mathrm{STF}_K(z) = \frac{K \cdot G(z)}{1 + K \cdot H(z)} \tag{6.2}$$

$$\mathrm{NTF}_K(z) = \frac{1}{1 + K \cdot H(z)} \tag{6.3}$$

These two transfer functions show that the input signal and the quantization noise are shaped differently. Normally, the feedback filter $H(z)$ is a low-pass filter with large gain at low frequencies. Eq. (6.3) shows that $\mathrm{NTF}_K(z)$ consequently becomes a high-pass filter, i.e., the quantization noise is suppressed for low frequencies due to the high low-frequency loop gain. Eq. (6.2) and Eq. (6.3) show that $\mathrm{STF}_K(z)$ can generally be chosen independently of $\mathrm{NTF}_K(z)$ by selecting $G(z)$ appropriately. Recall from Sec. 2.5 that $G(z) = H(z)$ for the traditional modulator. Consequently, $\mathrm{STF}_K(z) \approx 1$ for low frequencies in this case.

In many papers describing linearized modulator analysis, the quantizer gain is fixed to unity [3, 8]. A consequence of this simplification is that the NTF changes as the feedback filter is scaled, whereas the real modulator is invariant to such scaling, cf. Sec. 2.5. The use of a quantizer model with a variable quantizer gain enables a feedback filter scaling to be compensated by a reciprocal scaling of the quantizer gain such that the overall invariance of the NTF is maintained. This fact is one of the most important motivations for using a quantizer model with variable gain. The use of such models was introduced in [1, 5, 38, 63, 69]

The noise transfer function has a fundamental property: the leading term ntf(0) of the associated impulse response is unity as a consequence of the one sample delay of the feedback filter. This implies that the average logarithm of the magnitude characteristic must be zero or positive according to the following theorem [20]:

Theorem 6.1 (Gerzon & Craven Noise Shaping Theorem) *A transfer function NTF(z) scaled such that ntf(0) = 1 satisfies the relationship:*

$$\int_0^1 \log |\mathrm{NTF}(e^{i\pi f})| df \geq 0 \tag{6.4}$$

Equality is attained if and only if NTF(z) *is minimum phase.*

The theorem states that for a minimum phase NTF magnitude characteristic plotted with linear frequency axis and logarithmic magnitude axis, the areas above and below the 0 dB line are equal.

The noise shaping theorem is very useful: for a desired NTF, the theorem establishes the necessary scaling of an optimal minimum phase filter which approximates the prescribed NTF. For example, a wide transition band from stop-band to pass-band is 'expensive' in terms of noise shaping, i.e., the gain in the pass-band must be increased in order to compensate for a wider transition region. This fact also explains why lower oversampling ratios requires a more narrow transition band higher which only can be achieved by a higher filter order.

From linear system theory, it is known that the poles of a stable system must be inside the unit circle. Eq. (6.2) and Eq. (6.3) show that the poles of $1/(1+K\cdot H(z))$ determine the stability, naturally provided that $G(z)$ is stable itself. Thus, the stability depends on the choice of K and $H(z)$. For a given $H(z)$ there will generally be a certain K-interval $[K_{min}, K_{max}]$ for which the linearized system is stable [49, 63]. The stable K-interval can easily be found using so-called Nyquist plots, i.e., the curve of $-H(e^{j\pi f})$ plotted in the complex plane. The critical points with zero imaginary part correspond to frequencies where the closed loop system has infinite gain. The associated K-value is the reciprocal of the real value of the critical point. Two such K-values will then form the end points K_{min} and K_{max} of the stable K-interval.

The stability (in the linear sense) of the linearized system is rather 'fictitious', since there is generally no link to the stability of the real non-linear system (see part I for a discussion of stability). For instance, a given unstable modulator can generally be modeled as a stable linearized system for a suitable K-value. Conversely, it is also possible to have an unstable linearized model of a stable modulator — it is just a matter of choosing K. The reason for this deficiency of the linearized model is that the quantization noise $q(k)$ is not an independent input — in fact it is generated by the system itself.

6.3 Quasilinear Modeling

In order to proceed with the system analysis it is necessary to make some simplifications and assumptions. First of all, the input signal $x(k)$ is restricted to constant signals. The motivation for this simplification is that the modulator input is usually heavily oversampled. Secondly, the quantization noise $q(k)$ is modeled as a stochastic noise source. Finally, the mean value $m_y = \mathrm{E}\{y(k)\}$ of the modulator output is used as a descriptive parameter, i.e., the modulator is supposed to operate with a constant input m_x which gives a prescribed mean output m_y. For the traditional modulator with $G(z) = H(z)$ and where $H(z)$ has very high DC-gain, the modulator loop ensures that $m_y \approx m_x$.

In order to model $q(k)$ as a zero mean noise source, it is necessary that the gain factor K only applies for AC-components, i.e., K is generally not the ratio between the mean values of the quantizer output and input. The linearized model of Fig. 6.1 will thus be treated as an AC-model, i.e., it only applies for the AC-components of the signals. However, the knowledge of the mean modulator output m_y affects the AC-model: the output signal $y(k)$ can only take the values $+1$ or -1, i.e., the total output power is fixed to unity. This means that the DC- and AC-components must fulfill the relation:

$$\mathrm{V}\{y(k)\} = \mathrm{E}\left\{y^2(k)\right\} - \mathrm{E}^2\{y(k)\} = 1 - m_y^2 \tag{6.5}$$

The variance of the output signal is due do the quantization noise filtered by $\mathrm{NTF}_K(z)$ and the mean output value is due to the constant input signal $x(k)$. Eq. (6.5) shows that as the DC-input signal power increases, the total output noise power must be reduced equivalently. This relationship gives the first hint of why stability is reduced for increasing DC-input.

Now the quantization noise $q(k)$ will be assumed to be white stochastic noise with zero mean and variance σ_q^2. Since the transfer function between $q(k)$ and $y(k)$ is known (i.e., $\mathrm{NTF}_K(z)$), it is possible to express the output variance as:

$$\mathrm{V}\{y(k)\} = \sigma_q^2 \int_0^1 |\mathrm{NTF}_K(e^{j\pi f})|^2 df = \sigma_q^2 \cdot \mathrm{A}(K) \tag{6.6}$$

where σ_q^2 is the variance of $q(k)$ and $\mathrm{A}(K)$ is the total noise power amplification factor using a linearized gain K [1].

Using Parseval's relationship, $\mathrm{A}(K)$ can also be found from the time domain equation:

$$\mathrm{A}(K) = \sum_{k=0}^{\infty} |\mathrm{ntf}_K(k)|^2 \triangleq \|\mathrm{ntf}_K\|_2^2 \tag{6.7}$$

where $\mathrm{ntf}_K(n)$ is the impulse response corresponding to $\mathrm{NTF}_K(z)$.

The noise amplification factor $\mathrm{A}(K)$ is thus the squared two-norm of $\mathrm{NTF}_K(z)$. Recall from Sec. 2.4 that the feedback filter $H(z)$ must have a delay of at least one sample (i.e., the impulse response satisfies $h(k) = 0$, for $k < 1$). This implies $\mathrm{ntf}_K(0) = 1$, consequently, the following inequality holds:

$$\mathrm{A}(K) \geq 1 \tag{6.8}$$

Combining Eq. (6.5) and Eq. (6.6), one obtains a very important relationship implied by the white quantization noise assumption:

$$\mathrm{A}(K) = \frac{1 - m_y^2}{\sigma_q^2} \tag{6.9}$$

The white noise assumption implies thus that the $\mathrm{A}(K)$-curve of the feedback filter puts a constraint on the quantization noise power σ_q^2 and the quantizer gain K. There is consequently only one degree of freedom left, e.g., the choice of K. Once K is determined, σ_q^2 can be found using Eq. (6.9).

It will appear that the $\mathrm{A}(K)$-curves of a feedback filter are very important for understanding the stability of a modulator. Therefore, different types of feedback filters will be investigated. Empirical studies indicate that three qualitatively different categories of $\mathrm{A}(K)$-curves exist. Table 6.1 describes the three categories of $\mathrm{A}(K)$-curves with corresponding examples of loop filters. The A_{min}-value is the global minimum of the curve. Fig. 6.2 shows the actual $\mathrm{A}(K)$-curves of corresponding to the examples of Table 6.1. The type III curves are convex with a global minimum somewhere in the middle of the stable K-interval and $\mathrm{A}(K)$ becomes infinite at the endpoints of the stable K-interval. The type III curve is found for every chaotic and every high-order ($N > 2$) modulator. The type I curve has the global minimum $\mathrm{A}_{min} = \mathrm{A}(K = 0) = 1$ and the curve is increasing. Both non-chaotic first- and second order modulators exhibit the type I curve. A special case is the type II curve which is only found for second order modulators with a double pole at

---

[1] the noise amplification factor for K fixed to unity was introduced as a design parameter in [3]

Different A(K)-curve types			
Type	Description	Loop filters	Example
I	A(K) is increasing and $A_{min} = A(0) = 1$,	non-chaotic first- and second-order with distinct poles	$H(z) = \frac{z^{-1}-0.5z^{-2}}{1-z^{-1}+z^{-2}}$
II	A(K) is increasing and $A_{min} = A(0) > 1$,	second-order with double poles at $z = 1$	$H(z) = \frac{z^{-1}-0.5z^{-2}}{1-2z^{-1}+z^{-2}}$
III	$\bigcup$-convex, $A(K) \to +\infty$ for both $K \to K_{min}$ and $K \to K_{max}$	chaotic and every high-order ($N > 2$)	$H(z) = \frac{z^{-1}-0.5z^{-2}}{1-2z^{-1}+1.2z^{-2}}$

Table 6.1: A(K)-curves

$z = 1$ (e.g., the traditional double-loop modulator). The type II curve resembles the type I curve except for the fact that $A_{min} > 1$.

It will be explained in Sec. 6.5 how the qualitative difference between the three types of curve reflects the stability properties of the associated modulators.

6.4 Estimating the Quantizer Parameters

The one-bit quantizer output $y(k)$ can naturally be split into three components, namely a DC-component, the amplified input AC-component and the quantization noise, as follows:

$$y(k) = m_y + K(e(k) - m_e) + q(k) \tag{6.10}$$

where m_e is the mean value of the quantizer input $e(k)$.

So far the quantizer gain K can be chosen arbitrarily. All other parameters in Eq. (6.10) are given when K is fixed. The question is now: what is a good choice of K ? The criterion to be chosen henceforth is that Eq. (6.10) should split $y(k)$ into three orthogonal components, i.e., the quantization noise should be uncorrelated with the quantizer input. The covariance between the quantizer input and output yields:

$$\begin{aligned} \mathrm{Cov}\{e(k), y(k)\} &= \mathrm{E}\left\{(e(k) - m_e)(y(k) - m_y)\right\} \qquad (6.11) \\ &= K \cdot \mathrm{E}\left\{(e(k) - m_e)^2\right\} + \mathrm{E}\left\{(e(k) - m_e)q(k)\right\} \end{aligned}$$

In order to fulfill the orthogonality, the last term, i.e., the covariance between $e(k)$ and $q(k)$, must be zero. Consequently, the quantizer gain can be derived:

$$K = \frac{\mathrm{Cov}\left\{e(k), y(k)\right\}}{\sigma_e^2} \tag{6.12}$$

where $\sigma_e^2 = \mathrm{V}\{e(k)\}$ is the quantizer input variance.

The quantizer parameter K thus depends on the statistics of the quantizer input signal. This constraint justifies the designation *quasilinear modeling*. The modeling framework was introduced in [5] and was inspired from techniques known from control theory. Unfortunately, the excellent work in [5] has not received much attention. In some papers, the

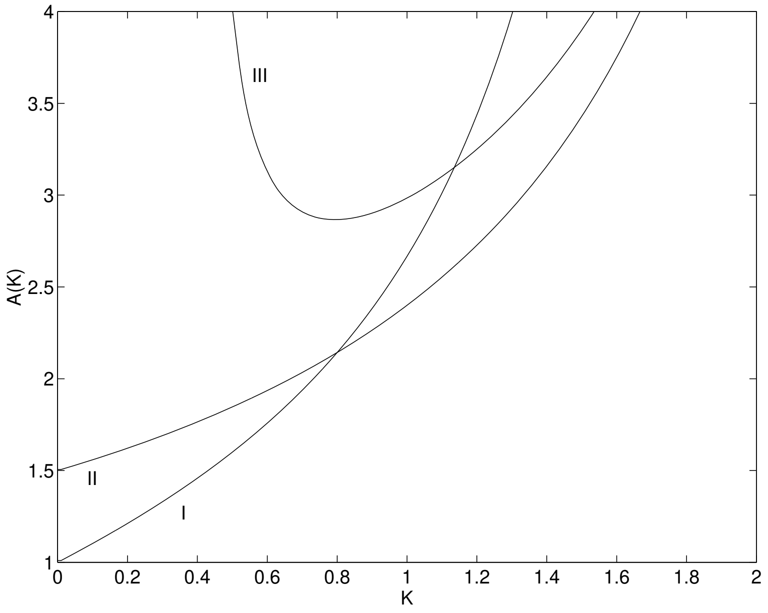

Figure 6.2: A(K)-curves for the type I, II and III loop filter examples in Table 6.1.

ratio between the mean values of the quantizer output and input is used as the quantizer gain [1, 69]. Another approach has been to use the mean value of $\frac{1}{|e(k)|}$ [56].

A direct consequence of the orthogonality criterion is that the quantization noise power σ_q^2 is minimized [5], i.e., as much as possible of the output variance is 'explained' by the linear model. Furthermore the orthogonality simplifies variance calculations: the output variance, cf. Eq. (6.5) and Eq. (6.10), is given as:

$$\mathrm{V}\{y(k)\} = 1 - m_y^2 = K^2\sigma_e^2 + \sigma_q^2 \tag{6.13}$$

The quantization noise power can be found by combining Eq. (6.13) and Eq. (6.12) while the orthogonality is used:

$$\begin{aligned} \sigma_q^2 &= 1 - m_y^2 - K^2\sigma_e^2 \\ &= 1 - m_y^2 - \frac{\mathrm{Cov}^2\{e(k), y(k)\}}{\sigma_e^2} \end{aligned} \tag{6.14}$$

Is fairly easy to calculate K and σ_q^2 for a known probability density function (pdf) of the quantizer input $e(k)$ [5, 49]. As stated in Sec. 2.5 a Σ-Δ modulator is invariant to a positive scaling of $e(k)$ due to the identity $\mathrm{sgn}(\alpha e(k)) = \mathrm{sgn}(e(k))$ for positive α. Such a scaling of $e(k)$ will naturally scale up $\mathrm{Cov}\{e(k), y(k)\}$ by α and σ_e^2 by α^2 and Eq. (6.12) shows that K scales inversely with α. Consequently, the total gain $K\alpha$ is invariant under such scaling. Furthermore, Eq. (6.14) shows that σ_q^2 is also invariant under a scaling of the quantizer input. This is a very important property of one-bit quantizers.

The quantization noise power σ_q^2 has been found for different types of quantizer input distributions with $m_y = \mathrm{E}\{\mathrm{sgn}(e(k))\} = \mathrm{Pr}\{e(k) > 0\} - \mathrm{Pr}\{e(k) < 0\}$ as parameter, i.e., the quantizer input is DC-shifted such that the mean output is m_y. The result for

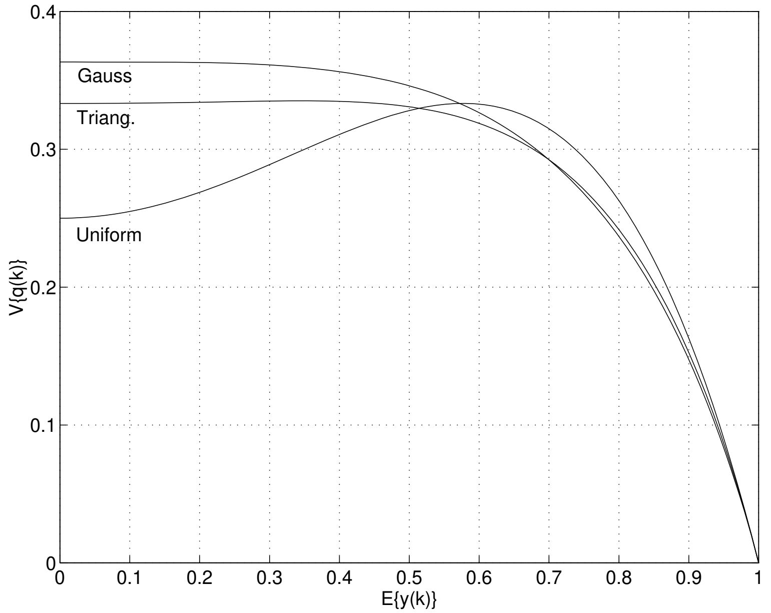

Figure 6.3: σ_q^2 versus $m_y > 0$ for three types of probability distributions, cf. Eq. (6.15), (6.16) and (6.17).

Gaussian pdf is [5]:

$$\sigma_q^2 = 1 - m_y^2 - \frac{2}{\pi} \exp\left(-2\left(\left(\mathrm{erf}^{-1}(m_y)\right)^2\right)\right) \tag{6.15}$$

where erf(·) is the error function.

Uniform pdf yields [49]:

$$\sigma_q^2 = 1 - m_y^2 - \frac{3}{4}\left(1 - m_y^2\right)^2 \tag{6.16}$$

Triangular pdf yields [49]:

$$\sigma_q^2 = 1 - m_y^2 - \frac{2}{3}\left(3(1 - m_y) - 2(1 - m_y)^{3/2}\right)^2 \tag{6.17}$$

Fig. 6.3 shows σ_q^2 versus m_y for these three types of probability distributions. The graph shows that the quantization noise power actually decreases to zero as m_y approaches unity. The intuitive explanations is that $+1$ and -1 are the only DC-values which the one-bit quantizer can approximate arbitrarily good. Traditionally, the quantization noise power of a uniform multibit quantizer in the non-overload region is assumed to be $\sigma_q^2 = \Delta^2/12$ where Δ is the quantizer step height. Consequently, the traditional assumption implies that $\sigma_q^2 = \frac{1}{3}$ for the one-bit quantizer with $\Delta = 2$. The triangular pdf is the only distribution which fulfills this (only for zero m_y). In [38, 63] the quantizer gain was found from an equation similar to Eq. (6.9) where σ_q^2 was assumed to be $\frac{1}{3}$ as given by the usual rule of thumb for multibit quantizer. In other words the noise amplification factor A was assumed to be 3 for zero input.

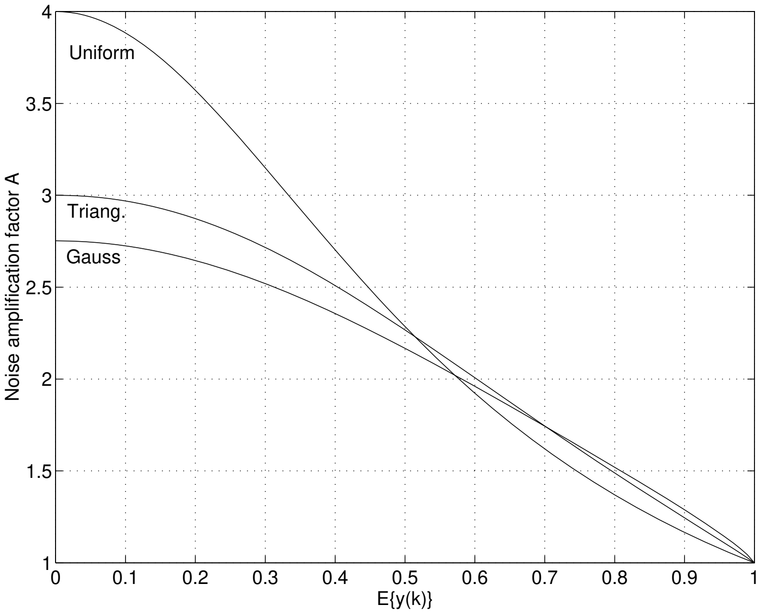

Figure 6.4: Noise amplification factor A versus m_y for these three types of probability distributions, cf. Eq. (6.15), (6.16) and (6.17).

Compared to triangular pdf, the Gaussian pdf gives higher σ_q^2 and uniform pdf gives lower σ_q^2 for low m_y values. As m_y approaches unity, the Gaussian pdf has the lowest σ_q^2.

A very interesting property of one-bit quantizers is that the noise amplification factor A as defined by Eq. (6.9) only depends on m_y when the quantizer input pdf type is known. The noise amplification factor for the quantizer is the ratio between the total output (AC) noise power and the quantization noise power. Fig. 6.4 shows A(m_y) obtained from Eq. (6.9), (6.15), (6.16) and (6.17). Note that the noise amplification factor curves are all decreasing functions of m_y.

6.5 Equilibrium and Stability

It was concluded in Sec. 6.3 that the feedback filter and the white noise assumption gives a relationship between the noise amplification factor A and the linearized quantizer gain K. The previous section showed that the orthogonality criterion links the noise amplification factor to the shape of the quantizer input pdf. The obvious question is now: for a real simulation, what will be the steady-state or equilibrium values of K and A where these parameters are determined using time-averages inserted in Eq. (6.12)?.

If the white noise assumption holds, the equilibrium point (K_{eq},A$_{eq}$) will naturally be on the A(K)-curve. The knowledge of the location on the A(K)-curve necessitates the knowledge of the actual quantizer input pdf.

If it is assumed that the quantizer input is, e.g., Gaussian, then the equilibrium can be found by solving the equation:

$$\mathrm{A}_{eq}(K_{eq}) = \frac{1-m_y^2}{\sigma_q^2} = \mathrm{A}(m_y) \tag{6.18}$$

where σ_q^2 is given by Eq. (6.15) and $\mathrm{A}_{eq}(K_{eq})$ is the equilibrium noise amplification factor given by Eq. (6.7).

The procedure is as follows: for a given m_y the associated noise amplification factor $\mathrm{A}(m_y)$ is found using Eq. (6.15). Subsequently, for a given feedback filter the K-value(s) giving the same A-value as the quantizer is found using the $\mathrm{A}(K)$-curve of the feedback filter, cf. Eq. (6.7).

The next topic to be addressed is the stability of the equilibrium. Recall that the loop parameters such as K_{eq} and A_{eq} are based on long term statistical averages. It is therefore interesting to investigate how the statistical equilibrium reacts to small perturbations reflecting randomness of the underlying stochastic signals. If K_{eq}, e.g., is increased slightly, the result will be a higher A_{eq}-value in the case where the $\mathrm{A}(K)$-curve is increasing around the equilibrium K. A higher A_{eq}-value means more circulating noise and, hence, more quantizer input noise and this will tend to decrease K_{eq} again. Such a mechanism forces the system back to equilibrium. Conversely, if the $\mathrm{A}(K)$-curve is decreasing around the equilibrium, even small perturbations destabilize the system.

The positive slope of the $\mathrm{A}(K)$-curve around the equilibrium has another implication: as m_y increases $\mathrm{A}(m_y)$ decreases (see Fig. 6.4) which gives a lower equilibrium K. The resulting reduced loop gain is detrimental to the noise suppression of the loop and the in-band noise power will consequently increase. This phenomenon is also observed from actual simulations [48].

The three types of $\mathrm{A}(K)$-curves according to Table 6.1 give rise to different stability considerations:

- **Type I:** There will always be a solution to Eq. (6.18) since $\mathrm{A}(K)$ covers the interval $]1,\infty[$ and the equilibrium is always stable since $\mathrm{A}(K)$ is increasing everywhere. Hence, the equilibrium is globally attracting.

- **Type II:** There is generally no solution for m_y close to unity since $\mathrm{A}_{min} > 1$. Due to the fact that the $\mathrm{A}(K)$-curve increases, the equilibrium is stable and globally attracting if it exists.

- **Type III:** There are generally two or no equilibrium points due to the convexity: for $\mathrm{A}(m_y) > \mathrm{A}_{min}$ there is a stable and an unstable equilibrium and the stable K_{eq}-value is higher than the unstable. As $\mathrm{A}(m_y)$ approaches A_{min} due to an increased m_y, the to solutions come closer and finally annihilate. At this point, the system conceptually escapes through the decreasing branch of the $\mathrm{A}(K)$-curve and the system is trapped in a state with a low quantizer gain synonymous with a high quantizer input magnitude. Note that the onset of instability occurs around the A_{min}-point which is in the middle of the stable K-interval, i.e., the stability of the linearized system does not describe the stability of the modulator.

It is seen that the quasilinear model predicts the unconditional stability of type I systems, i.e., non-chaotic first- and second-order modulators. Unconditional stability means that equilibrium is reached irrespectively of the initial conditions, i.e., the equilibrium is globally attracting. This conclusion is in agreement with observations for actual type I systems: these systems stay bounded and have no large amplitude overload limit cycles.

Type II systems have no equilibrium for high m_y values according to the quasilinear model because $\mathrm{A}_{min} > 1$ (e.g., the typical double-loop modulator with double pole at $z = 1$ has $\mathrm{A}_{min} \approx 1.5$ cf. Fig. 6.2). However, if the input amplitude is reduced the system

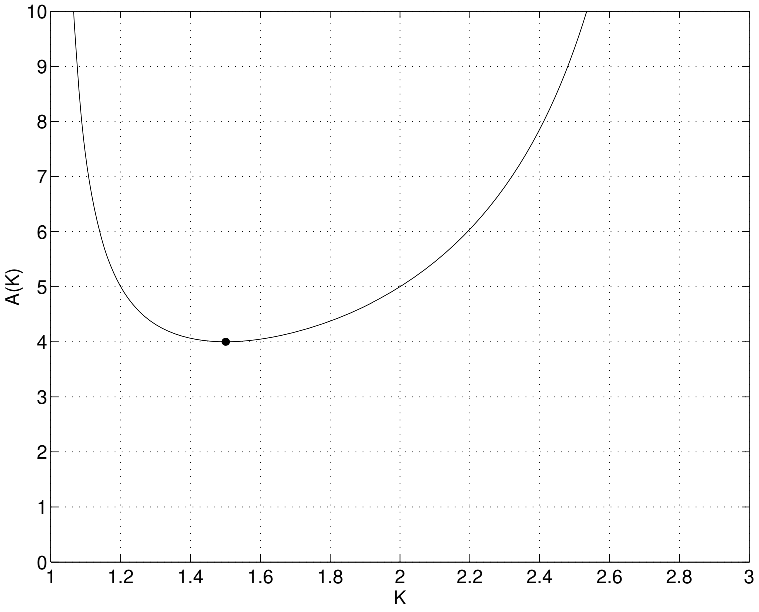

Figure 6.5: Noise amplification factor A versus K for the first order feedback filter $H(z) = z^{-1}/(1-2z^{-1})$. The A_{min}-point is shown with a dot.

reaches always equilibrium thanks to the increasing A(K)-curve. Real type II systems have in fact the property that $\max(|e(k)|) \rightarrow \infty$ as $m_y \rightarrow 1$, [68], i.e., type II systems become gradually unstable as the input amplitude approaches unity.

Type III systems have a characteristic abrupt onset of instability: the system 'blows up' as the A_{min}-point is reached, i.e., the quantizer input suddenly jumps to a high or even unbounded magnitude. The loss of stability is irreversible, that means, a decreased input magnitude does not reestablish stability unless the system is also reset, i.e., the state-variables are set to zero. These conclusions for type III systems made from the quasilinear model are qualitatively in good agreement with the fact that chaotic modulators loose the stability due to a boundary crisis (cf. Ch. 3) and that high-order non-chaotic systems have large scale overload limit cycles that can be activated by small perturbations (cf Ch. 5).

Example 6.1 Recall the chaotic first-order modulator discussed in Example 3.1. The feedback filter for this modulator class is given by:

$$H(z) = \frac{z^{-1}}{1-az^{-1}} \tag{6.19}$$

It was established in Example 3.1 that the modulator becomes unstable for zero input for $a > 2$. In this example, zero input and $a = 2$ will be used.

The closed loop linearized first-order system can only get unstable when the pole crosses the unit circle at $z = 1$ or $z = -1$. Insertion of these two real values into $H(z)$ shows that the stable K-interval is from $K_{min} = 1$ to K_{max}=3. Fig. 6.5 shows a characteristic type III A(K)-curve with A_{min} =

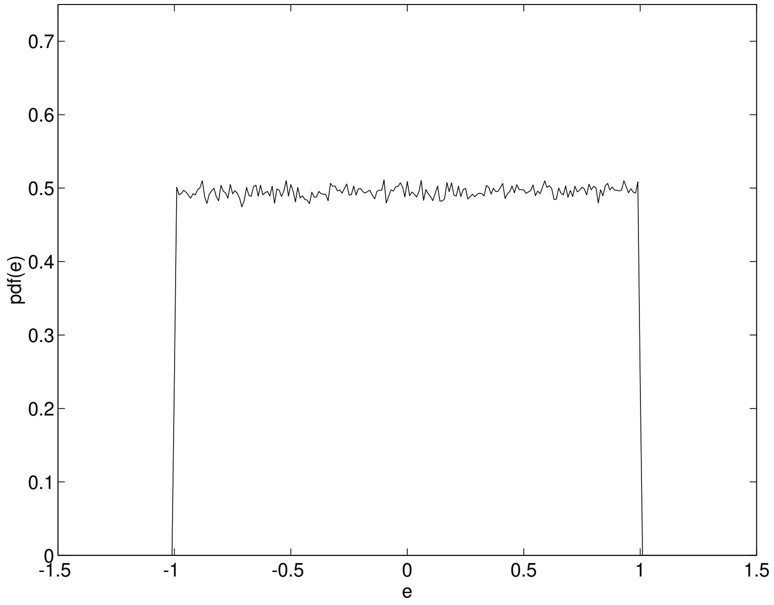

Figure 6.6: Estimated quantizer input pdf for the first order modulator with $H(z) = z^{-1}/(1-2z^{-1})$. Obtained using 200 uniformly spaced bins from -1 to 1 and 1 mill. samples.

A(1.5) = 4. The system should consequently be unstable when the quantizer input is assumed to be Gaussian since this type of pdf has $\mathrm{A}_{max} = 2.75$ for $m_y = 0$, cf. Fig. 6.4. A simulation of the system was performed and the quantizer input pdf was estimated and plotted in Fig. 6.6 which shows that the pdf seems to be uniform from -1 to 1. The uniform quantizer input pdf has $\mathrm{A}_{max} = 4$ for $m_y = 0$ according to Eq. (6.16) and Eq. (6.18). Furthermore, using Eq. (6.12) a uniform pdf from -1 to 1 gives a linearized quantizer gain of 1.5 [49]. It can therefore be concluded that the quasilinear model also predicts that the system is only marginally stable with zero input, i.e., it operates on the A_{min}-point. A slight parameter perturbation causes the system to blow up. The quasilinear explanation is the escape along the unstable (left-hand) branch of the A(K)-curve and the explanation from system dynamics theory is the collision between the limit set and the associated basin of attraction □

6.6 The Gaussian Stability Criterion

The assumption of a particular type of quantizer input pdf and the A_{min}-value for a type III system defines a maximum stable amplitude (MSA) which is the highest m_y which gives a stable equilibrium point. Empirical results have shown that the MSA derived on the assumption of Gaussian pdf is very accurate. In this case MSA can be found using:

$$\mathrm{A}_{\mathrm{Gauss}}(\mathrm{MSA}) = \mathrm{A}_{min} \tag{6.20}$$

where the noise amplification factor on the left hand side is found using Eq. (6.15) and Eq. (6.18).

It is difficult to justify this stability criterion since the quantizer inputs of real modulators are often very far from being Gaussian (as shown in Example 6.1 where the Gaussian criterion is rather pessimistic). However, investigations presented in Sec. 6.7 indicate that the Gaussian criterion becomes fairly accurate for high-order modulators.

6.7 Loop Filter Design Using NFT-prototypes

It is common practice to design feedback filters of Σ-Δ modulators from a desired noise transfer function, i.e., an *NTF-prototype* [1, 10, 49, 58]. This can be done using 'reverse engineering' on Eq. (6.3): Using a rational z-domain prototype $\mathrm{NTF}(z) = A(z)/B(z)$ one arrives at:

$$H(z) = \frac{1}{K}\frac{B(z) - A(z)}{A(z)} \tag{6.21}$$

The first problem is that the quantizer gain is generally unknown or can be considered as a free parameter. However, the equation shows that K only scales $H(z)$ and K can thus arbitrarily be set to unity since the modulator is invariant to feedback filter scaling.

The next problem is that the feedback filter must not be delay free in order to ensure implementability (see Sec. 2.4). Eq. (6.21) shows that this property is obtained when the NTF-prototype is scaled such that $A(z)$ and $B(z)$ have the same highest order z-term or, equivalently, scaled such that the NTF-prototype impulse response has the property $\mathrm{ntf}(0) = 1$. The term *the necessary scaling* will be used for such scaling of an NTF-prototype.

The NTF-prototype is usually designed as a high-pass filter with the base-band as stop-band in order to obtain a high base-band noise suppression. A fundamental property of high-pass filters is that as the NTF-prototype achieves better base-band rejection, the necessary scaling increases the pass-band gain. This effect is a direct consequence of Th. 6.1 and this has an important implication: the minimum noise amplification factor A_{min} increases due to the higher pass-band gain. The latter effect decreases the maximum stable amplitude (MSA) according to the Gaussian criterion presented in Sec. 6.6. Consequently, there is a fundamental trade-off between noise suppression and stable amplitude range.

Example 6.2 A class of good NTF-prototypes is the Chebychev II (i.e., inverse Chebychev) high-pass filters. The filters have unit circle zeros distributed in the stop-band for equiripple magnitude characteristic. Such filters are easily obtained using standard filter design packages, e.g., the MATLAB signal processing toolbox command:

```
[A,B]=cheby2(N,Rs,fb,'high');
```

produces an `N`th order unity pass-band gain Chebychev II prototype with minimum stop-band rejection `Rs` (in dB) in the stop-band up to `fb` relative to half the sample rate. The necessary scaling implies that the numerator polynomial `A` must be scaled such that the first element `A(1)` is unity. This scaling reduces the stop-band rejection and increases the pass-band gain. The feedback filter is then obtained using Eq. (6.21) with K set to unity.

Filter	Rs	A_{min}	Gauss. MSA	Sim. MSA
A	85.5 dB	1.76	0.69	0.69
B	105.5 dB	2.48	0.33	0.32
C	110 dB	2.73	0.08	0.13

Table 6.2: Parameters for three feedback filter examples

Fig. 6.7 shows the scaling factor (i.e., $1/\mathsf{A}(1)$) and A_{min} versus Rs for fifth order Chebychev II filters with fb $= 1/64$. Notice the dB-dB scale and that the noise amplification factor is a ratio between power levels, i.e., the plot shows $10\log_{10}(A_{min})$. The maximum noise amplification factor for Gaussian quantizer input is $1/(1-2/\pi) = 2.75$ corresponding to approx. 4.4 dB. Hence, filters with Rs greater than 111 dB are expected to be unstable even for zero input. Notice also that the scaling grows faster than A_{min}.

The maximum stable input amplitude was investigated by simulations for this class of fifth order Chebychev II modulators. A ramp type input, increasing linearly from zero to unity during one million time steps, was used for 150 feedback filters with Rs ranging from 60 dB to 115 dB. Fig. 6.8 shows the simulated maximum stable amplitude (MSA) versus Rs. The smooth curve is the Gaussian A_{min}-criterion of Eq. (6.20) and this theoretical curve follows the simulation results quite well. It has been claimed (wrongly) that the MSA is inversely proportional to the modulator order [55, 7]. However, the plot in Fig. 6.8 is typical for a wide range of modulator orders. In general, any desired MSA can be achieved by using a suitably low A_{min}.

The Gaussian criterion is somewhat pessimistic when the stable amplitude range is low. The simulated curve cuts off abruptly to zero input instability for Rs greater than approx. 113 dB (the Gaussian criterion predicts this point to at 111 dB). For very low Rs there is a high stable range but the slope in this end is very low, e.g., a 10 dB reduction in Rs gives only a marginal relative increase in stable range. The SNR will thus increase with Rs for low Rs because the in-band noise rejection improves and the stable range is only reduced slightly. Eventually, the stable input range curve becomes to steep, i.e., the increased in-band noise rejection is counterbalanced by the decreased stable input range. At this point the SNR has a maximum. A higher Rs allows too little signal power. The SNR maximum is typically found for modulators with maximum stable amplitude around 0.35; however, this value depends on the modulator order and other design criteria □

This example showed that the Gaussian stability criterion is fairly accurate for a certain class of high-order modulators. Example 6.1 showed that the A_{min}-point could describe the marginal stability of a certain first order modulator. However, the Gaussian criterion was grossly pessimistic. This can be explained by the fact that the real quantizer input had a uniform pdf. An interesting question is how accurate is the quasilinear model in general for high-order modulators? This question is addressed by Example 6.3.

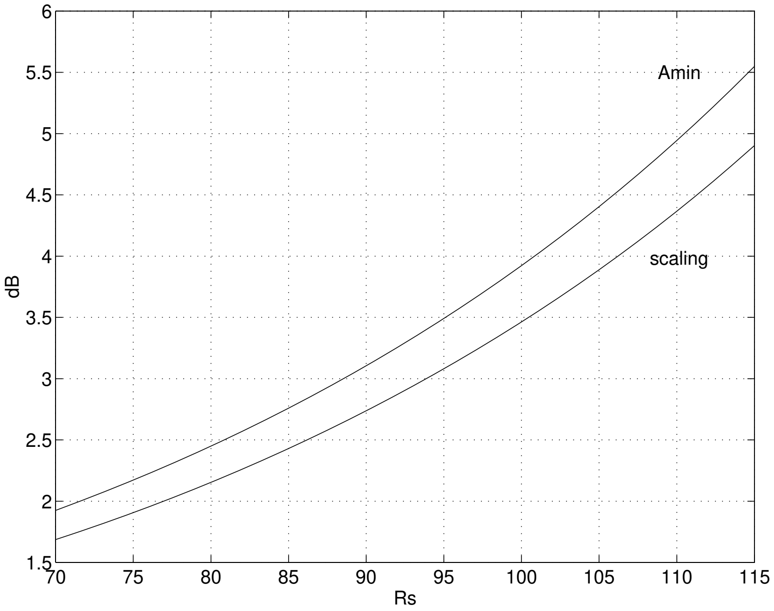

Figure 6.7: Scaling and A_{min} versus **Rs** (in dB) for fifth order Chebychev II filters with **fb** $= 1/64$.

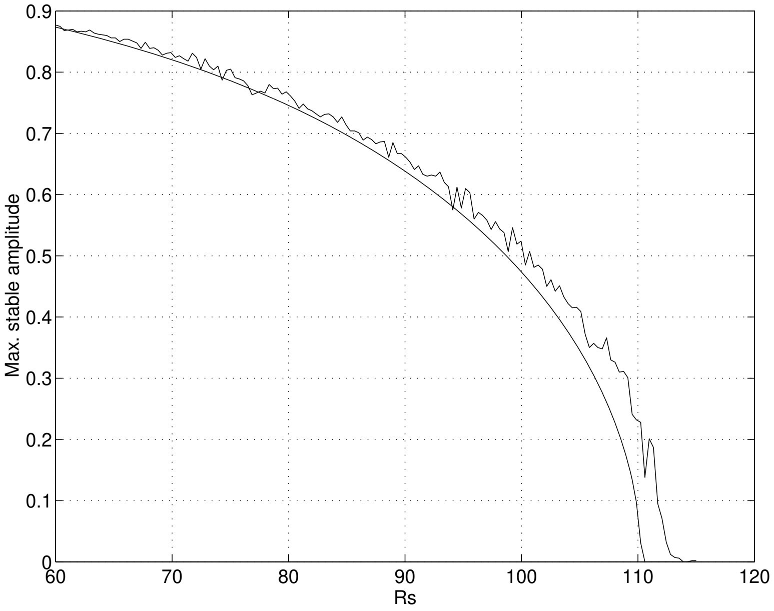

Figure 6.8: Simulated maximum stable input amplitude versus **Rs** for fifth order Chebychev II filters with **fb** $= 1/64$. The smooth curve is the Gaussian A_{min}-criterion of Eq. (6.20).

Example 6.3 Three fifth-order Chebychev II feedback filters with parameters shown in Table 6.2 (see also Example 6.2) will be investigated by simulations. The simulations are performed in segments of 10^6 time steps at a time with fixed dc-inputs from zero to unity in increments of 0.01. For each segment the quantizer parameters (K, σ_q^2) are found using time averages in Eq. (6.12) and Eq. (6.14). The maximum stable amplitude (MSA) is shown in Table 6.2 including the MSA based on the Gaussian criterion. The slower rise in input value gives a somewhat lower simulated MSA than for Fig. 6.8. Fig. 6.9 shows the noise amplification factor A plotted versus the quantizer gain K. The thick lines are the simulated results (i.e., using time averages within each 10^6 sample segment) and the dashed lines are the theoretical A(K)-curves based on the filter coefficients. It is clear that the simulated results are in good agreement with the theoretical. The three modulators operate as expected on the increasing (stable) branches of the A(K)-curves. However, the A_{min}-points are far from being reached: the instability begins for higher A- and K-values. Notice also that the A_{min}-points are found for K slightly below unity for this class of NTF-prototypes, i.e., the actual noise transfer functions do not match the NTF-prototypes due to the equilibrium K_{eq} being higher than unity.

Thus, it is obvious that the real quantizer input cannot have a Gaussian pdf. This is confirmed by Fig. 6.10 which shows the noise amplification factor A plotted versus the dc-input (almost equal to the output mean value m_y due to the very high dc loop gain). It is seen that the simulations give higher A-values corresponding to less quantization noise than for Gaussian quantizer input pdf. For low dc-input the simulated A-values are lower than the A-curve for uniform pdf. However, for higher dc-inputs, the simulations outperform even the curve for uniform pdf. The curves for filter B and C follow almost the same path corresponding to an almost constant quantization noise power $\sigma_q^2 = 0.285$.

Fig. 6.11 shows the simulated quantizer gain K versus the dc-input. The curve for filter A starts with a very high K-value and the curves for filter B and C start with lower K-values. This is because all modulators start with almost the same A-value (see Fig. 6.9). All curves show that K declines with increasing dc-input □

The previous example showed that a feedback filter designed from an NTF-prototype using $K = 1$ in Eq. (6.21) will operate with a somewhat higher K in practice. The resulting noise transfer function will consequently differ from the NTF-prototype. This fact is one of the major drawbacks of feedback filter design using NTF-prototypes. Due to the invariance to feedback filter scaling, there is an entire equivalence class of NTF-prototypes giving the same output spectrum corresponding to the same normalized feedback filter.

Example 6.4 A modulator with filter B (cf. Table 6.2) was simulated for 10 million time steps with a dc-input of 1/256. The quantizer parameters were estimated from the appropriate time-averages. Quantization noise was found using Eq. (6.10):

$$q(k) = y(k) - m_y - K(e(k) - m_e) \tag{6.22}$$

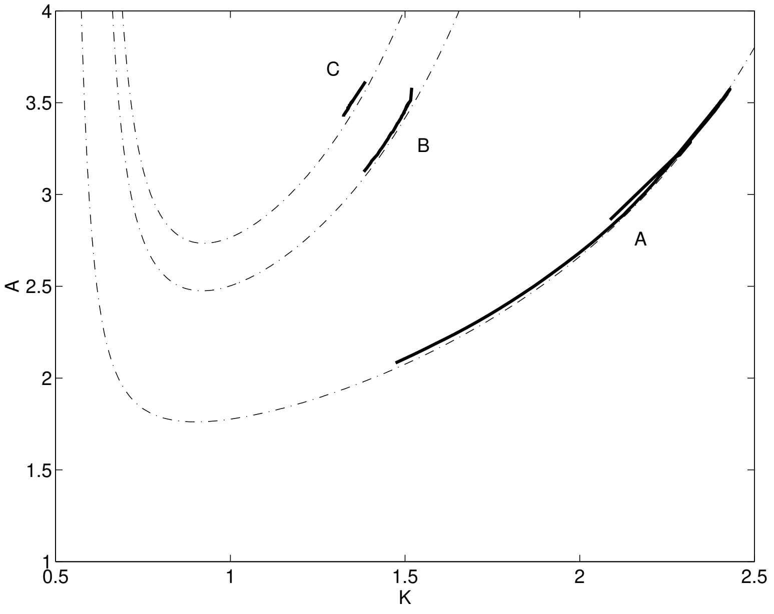

Figure 6.9: Simulated and theoretical (dashed lines) A(K)-curves for filter A, B and C, cf. Table 6.2.

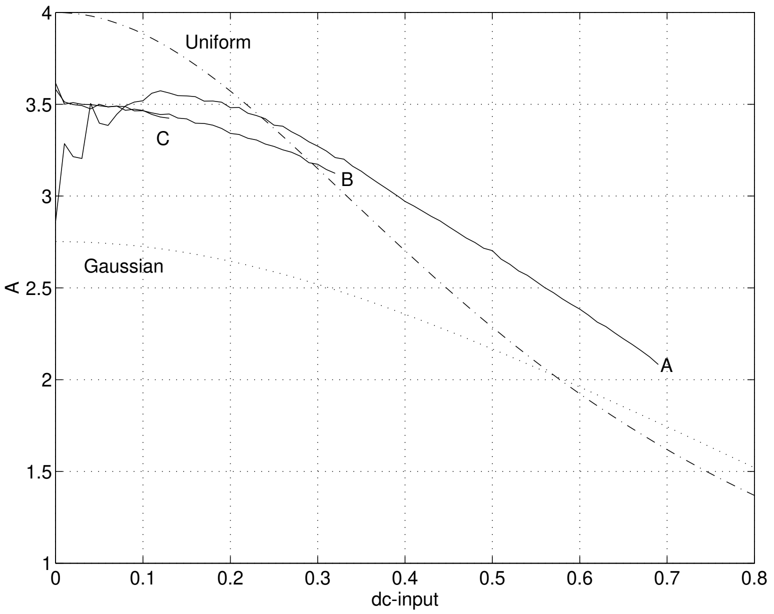

Figure 6.10: Noise amplification factor A versus dc-input found by simulations of filter A, B and C, cf. Table 6.2. Theoretical curves for Gaussian and uniform quantizer input pdf are shown with dashed lines.

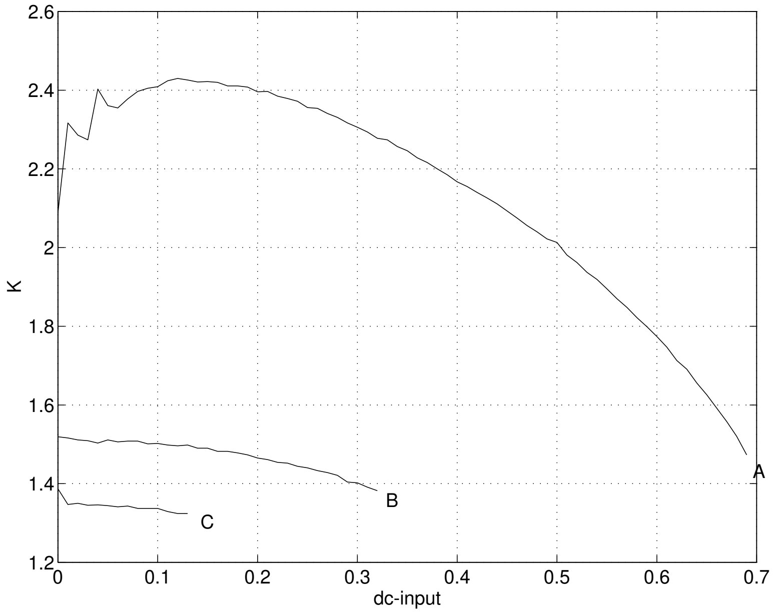

Figure 6.11: Quantizer gain K versus dc-input found by simulations of filter A, B and C, cf. Table 6.2.

> Fig. 6.12 shows power spectrum estimates for both the modulator output $y(k)$ and the quantization noise $q(k)$. The spectra are obtained using averaged 8k Kaiser-Bessel windowed FFT power spectra (Welch' method). It is observed that the spectrum of the quantization noise has a fairly flat spectrum, i.e., the white noise assumption is justified. Furthermore, the spectrum of the output does not follow the magnitude characteristic of the NTF-prototype (the passband is not flat as for the Chebychev II prototype). The quasilinear models predicts this fact by estimating a quantizer gain K higher than unity (see Fig. 6.11).
>
> The difference between the two spectra represents the noise transfer function. Consequently, since the NTF is minimum phase, the areas of the two regions between the two curves must be equal according to the noise shaping theorem (Theorem 6.1).
>
> Both spectra shows a strong tone near half the sample rate. This tone is very characteristic to all Σ-Δ modulators. The exact frequency of this tone is $(1-m_y)f_s/2$. A very weak intermodulation tone at the frequency $m_y \cdot f_s$ can be seen on the quantization noise spectrum. The suppression of these tones is the subject of chapter 8 □

The examples in this section have clearly demonstrated the fairly high accuracy of the quasilinear model. The NTF-prototype method combined with the Gaussian criterion allows efficient modulators to be designed with a prescribed maximum stable amplitude range. The remaining problem is that the actual equilibrium point is hard to predict and that the resulting noise transfer function is not identical to the prototype. However, most modulators seem to follow approximately the same A(m_y)-curve (see Fig. 6.10) and this

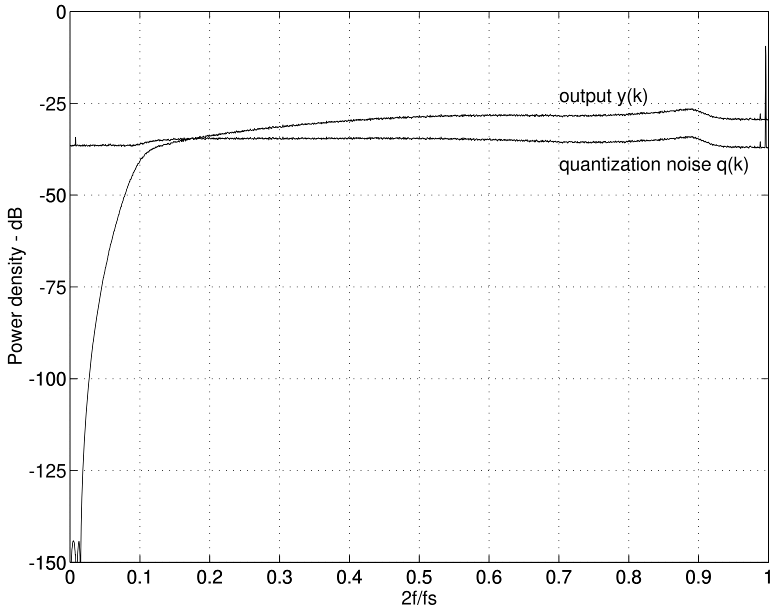

Figure 6.12: Power spectra of the modulator output $y(k)$ and the quantization noise $q(k)$ for a 10 million time step simulation of filter B, cf. Table 6.2 with a dc-input of 1/256. Note the frequency scale normalized in respect to half the sample rate.

enables fairly accurate equilibrium predictions to be made. The topic of the next section will be how to achieve even better equilibrium predictions.

6.8 Predicting the Quantizer Input pdf

As stated previously, the determination of the equilibrium point (i.e., the quantizer gain and noise amplification factor) requires the knowledge of the probability density function (pdf) of the quantizer input. This topic has been addressed in [64] for Gaussian modulator input. The purpose of this section is to introduce an algorithm which can predict the quantizer input pdf and thereby the important quantizer parameters such as the quantizer gain K and the quantization noise power σ_q^2. With these parameters in hand, it is easy to calculate other performance measures such as the in-band noise power.

A new assumption will is introduced: the quantization noise $q(k)$ is assumed to be an independent and identically distributed (i.i.d.) sequence. It was shown in the previous Section that the quantization noise typically is only close to being white. The independence assumption is therefore far from being fulfilled; however, the independence assumption makes calculations of probability distributions possible.

The first problem is to establish the connection between the pdf of the quantizer input and the quantization noise. When the quantizer parameters and the mean values are known (i.e., K, m_e and m_y) the time-domain relationship is given by the function (cf. Eq. (6.22)):

$$\mathrm{q}(e) = \mathrm{sgn}(e) - m_y - K(e - m_e) \tag{6.23}$$

This function is showed schematically in Fig. 6.13. Note that q(e) is non-invertible in the

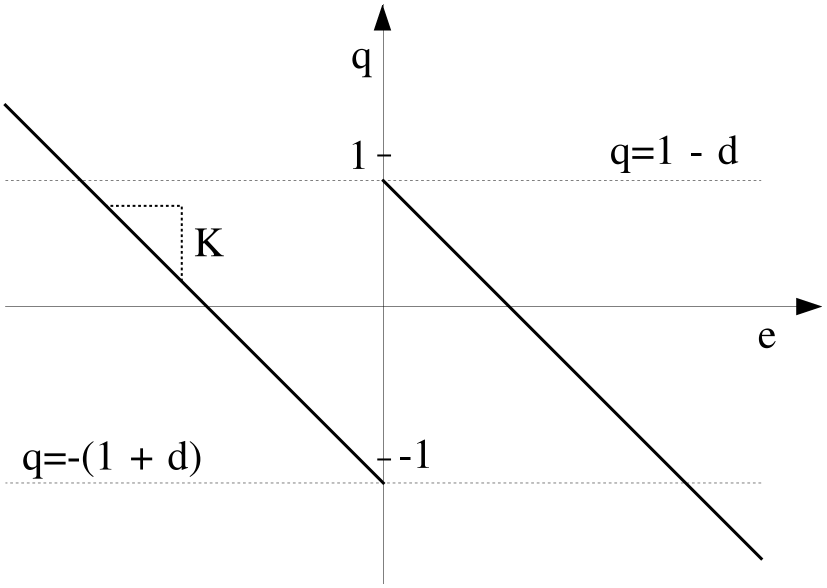

Figure 6.13: Quantization noise q versus quantizer input e for the quasilinear quantizer model, cf. Eq. (6.23) with displacement factor d> 0.

interval $[-1-d, 1-d]$ where the displacement d is defined as $d = m_y - Km_e$. For η in this interval the equation $\mathrm{q}(\xi) = \eta$ has two solutions ξ_1 and ξ_2.

For a given quantizer input probability density function $\mathrm{pdf}_e(\xi)$ the quantization noise probability density function $\mathrm{pdf}_q(\eta)$ is given by:

$$\mathrm{pdf}_q(\eta) = \sum_{\mathrm{q}(\xi)=\eta} \frac{\mathrm{pdf}_e(\xi)}{|\mathrm{q}'(\xi)|} = \frac{1}{K} \sum_{\mathrm{q}(\xi)=\eta} \mathrm{pdf}_e(\xi) \tag{6.24}$$

where the summations are over the maximum two solutions of the equation $\mathrm{q}(\xi) = \eta$ and $\mathrm{q}'(\xi)$ is the derivative of $\mathrm{q}(\xi)$.

The next step is to determine the pdf of the quantizer input by including the effects of the feedback filter. Recall that a generic modulator with a linearized quantizer model is characterized by two transfer functions, namely the signal transfer function $\mathrm{STF}(z)$ and the noise transfer function $\mathrm{NTF}(z)$. It is now convenient to consider a new transfer function: the *error transfer function*, $\mathrm{ETF}(z)$ between the noise source $q(k)$ and the quantizer input $e(k)$. It is seen from Fig. 6.1 that the relationship between $\mathrm{ETF}(z)$ and $\mathrm{NTF}(z)$ is given by:

$$\mathrm{NTF}_K(z) = 1 + K \cdot \mathrm{ETF}_K(z) \tag{6.25}$$

This implies that:

$$\mathrm{ETF}_K(z) = (\mathrm{NTF}_K(z) - 1)/K = \frac{-H(z)}{1 + K \cdot H(z)} \tag{6.26}$$

The quantizer input $e(k)$ is the quantization noise $q(k)$ filtered by $\mathrm{ETF}(z)$. Hence, in the time-domain $e(k)$ is given by the convolution:

$$e(k) = \sum_n q(k-n)\mathrm{etf}_K(n) \tag{6.27}$$

where $\mathrm{ntf}_K(n)$ is the impulse response of $\mathrm{ETF}_K(z)$.

Since $q(k)$ is considered to be an i.i.d. sequence, the pdf of $e(k)$ is the convolution of the pdf's of $q(k)$ scaled with each element of the impulse response $\mathrm{etf}_K(n)$, as shown by:

$$\mathrm{pdf}_e = \mathrm{pdf}_{q \cdot \mathrm{etf_K(1)}} * \mathrm{pdf}_{q \cdot \mathrm{etf_K(2)}} * \mathrm{pdf}_{q \cdot \mathrm{etf_K(3)}} * \cdots \tag{6.28}$$

Note that $\mathrm{etf}_K(k)$=0 for $k < 1$.

It is now possible to specify an algorithm which hopefully produces an invariant pdf of the quantizer input, such that both Eq. (6.24) and Eq. (6.28) are true: the quantizer input pdf should be invariant to the quasilinear model, i.e., the pdf should 'generate itself' using the quasilinear model.

The algorithm is as follows:

1. Select an initial pdf_e and select m_e as a fixed parameter.
2. Compute the quantizer gain $K = \frac{\mathrm{Cov}\{e,y\}}{\sigma_e^2}$.
3. Compute pdf_q using Eq. (6.24).
4. Compute $\mathrm{etf}_K(k)$ for $k < k_{max}$ where k_{max} is a suitable upper bound.
5. Compute $\overline{\mathrm{pdf}}_e$ using the n_{max} first terms of the convolution of Eq. (6.28).
6. Let $\mathrm{pdf}_e = \mu\overline{\mathrm{pdf}}_e + (1-\mu)\mathrm{pdf}_e$, where $0 < \mu \leq 1$ is a suitable step-size.
7. If not converged goto 2.
8. Stop and compute m_y.

The algorithm is stopped when a suitable steady-state solution is achieved. The convergence properties of the algorithm can be adjusted by choosing the step-size μ. The probability density functions must be sampled in order to facilitate computations on a digital computer. The convolutions can effectively be accomplished using the FFT-algorithm. The algorithm operates with m_e as a parameter, i.e., the quantizer input is always shifted such that it has mean value m_e. When steady state is reached, the resulting mean output value m_y can be computed.

Example 6.5 A modulator with feedback filter B from Table 6.2 was used to investigate the accuracy of the pdf algorithm. The algorithm used pdf's sampled 512 times over the interval $[-2.56, 2.56]$ and FFT's for the convolutions. A linear interpolation was used for computation of the scaled pdf's of the quantization noise and the first 30 terms of $\mathrm{etf}_K(n)$ was used.

Fig. 6.14 shows the predicted quantizer input pdf for $m_e = 0$ together with a simulated pdf estimate for the modulator with zero input. The two graphs are in relatively good agreement. Furthermore, the quantizer parameters match closely: the quantizer gain K is 1.536 for the prediction and 1.512 for the simulation. The noise amplification factor A is 3.53 for the prediction and 3.51 for the simulation.

The m_e parameter was increased to 0.2 and the algorithm run again. The resulting pdf is shown in Fig. 6.15. The mean output value for this pdf is

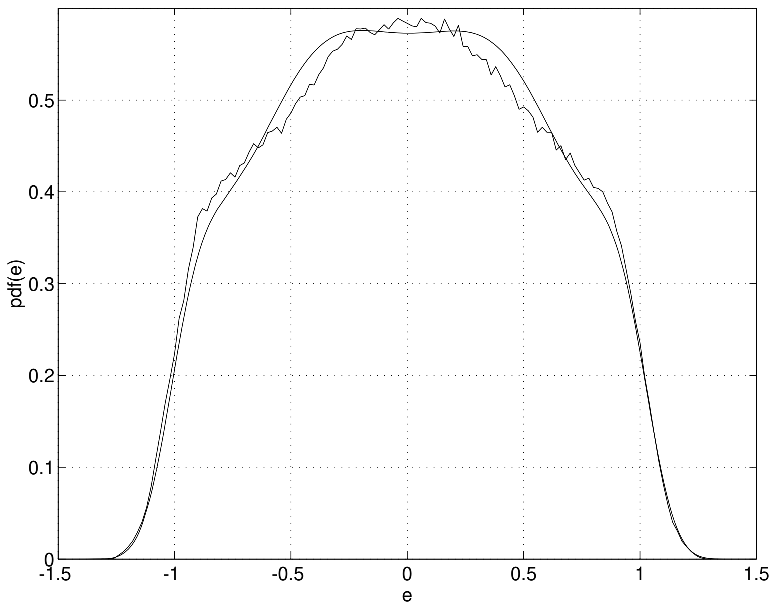

Figure 6.14: Predicted (smooth curve) and simulated quantizer input pdf for a modulator with filter B, cf. Table 6.2 and zero input.

> m_y=0.2443 and this constant value was used as input for a simulation. The resulting quantizer input pdf is also shown in Fig. 6.15. Both pdf's have a asymmetry which is characteristic when the constant input is non-zero. The asymmetry enhances the quantizer gain over symmetric pfd's. The predicted K is 1.434 and the simulated is 1.442. The simulations indicate an A-value of 3.28 and the prediction shows 3.18. The two pdf's do not match perfectly but they generally agree on the cut-off for high and low e-values □

The prediction method relies on the quantization noise i.i.d. assumption which is far from being fulfilled in practise. Even for white quantization noise, the quantizer input is generally non-white and the quantization noise will inherit some of the inter-sample dependence.

Nevertheless, the example showed that the equilibrium parameter predictions (i.e., K and A) are very precise for a certain high-order modulator. In some cases the pdf predictions obtained from the presented algorithm can be grossly wrong and in other cases, the algorithm gets unstable.

> **Example 6.6** The first order modulator from Example 6.1 with feedback filter $H(z) = z^{-1}/(1-2z^{-1})$ operates with zero input at an equilibrium K-value of 1.5 and a uniform quantizer input pdf over the interval $[-1,1]$. The ETF impulse response $\mathrm{etf}_K(k)$ is plotted for $K=1.5$ in Fig. 6.16. It is seen that $\mathrm{etf}_K(1)=1$ and that the succeeding values decay exponentially. The one-norm is found to be:
>
> $$\|\mathrm{etf}_{1.5}\|_1 = \sum_k |\mathrm{etf}_{1.5}(k)| = 2 \tag{6.29}$$

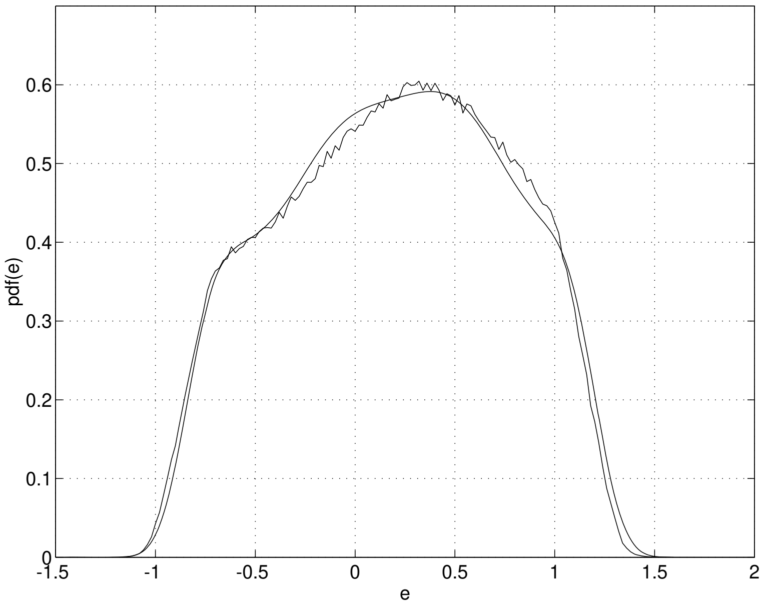

Figure 6.15: Predicted (smooth curve) and simulated quantizer input pdf for a modulator with feedback filter B, cf. Table 6.2 with $m_e = 0.2$ corresponding to $m_y = 0.2443$.

> The i.i.d. quantization noise assumption indicates that the maximum quantizer input should have a magnitude of 2 since the maximum quantization noise magnitude is at least unity. This is obviously not true since $e(k)$ is uniformly distributed over the interval $[-1, 1]$. In fact, the pdf algorithm gets unstable when this first order feedback filter is used: the quantizer input variance grows resulting in a lower K which gives a higher one-norm of $\mathrm{etf}_K(n)$ and this increases the quantizer input variance further $\square$

The pdf algorithm is not a very efficient tool for equilibrium and performance prediction of Σ-Δ modulators due to the time consuming numerical convolutions. The most straight forward way is in fact to simulate the system and extract the quantizer parameters directly. Using these parameters the in-band noise power can be estimated quickly. The best performance evaluation is of course to actually measure the in-band noise power. However, this requires time consuming numerical FFT or filtering operations and rather long simulations. Useful quantizer parameter estimates are typically obtained from rather short simulations. The main importance of the pdf algorithm is that it provides useful theoretical insight in the interactions between the feedback filter and the quantizer input pdf.

6.9 Summary

The use of linearized models for Σ-Δ modulators was addressed in this chapter. The nonlinear effects of the one-bit quantizer were described by an additive quantization noise signal. The linearized model has one degree of freedom, namely the choice of the linearized quantizer gain. When the quantization noise is assumed to be white and stochastic noise,

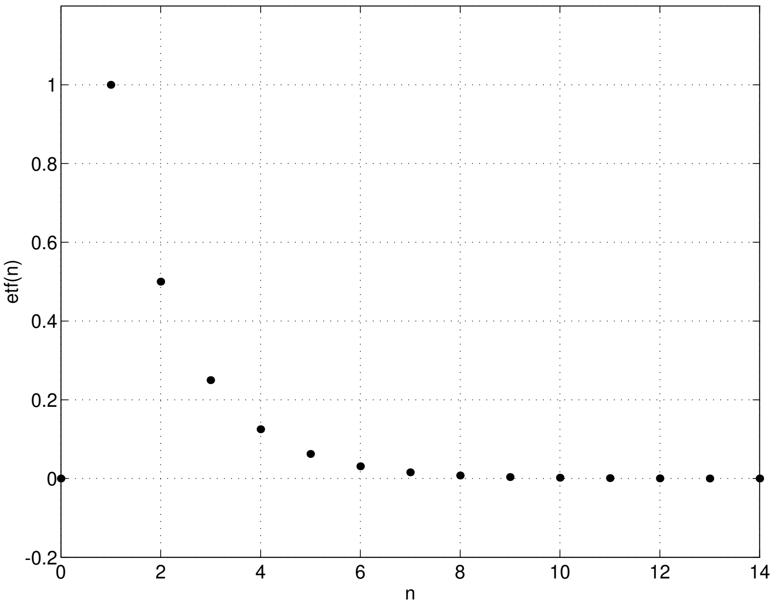

Figure 6.16: The impulse response $\mathrm{etf}_K(n)$ for the feedback filter $H(z) = z^{-1}/(1 - 2z^{-1})$ and $K = 1.5$.

the feedback filter and the unity output power constraint defines a noise amplification curve which relates the quantizer gain to the quantization noise power.

The quantizer gain was determined using an orthogonality criterion, i.e., the quantization noise should be uncorrelated with the quantizer input. This also ensures that the quantization noise has minimum power.

The quantization noise power of different quantizer input probability density functions were investigated and this defines a noise amplification factor versus the quantizer dc-output. Using the noise amplification curve of the feedback filter, the equilibrium quantizer gain and noise power can be predicted for any particular quantizer input pdf. The global minimum of the noise amplification curve of the feedback filter introduces a class of stability criteria. In particular, the Gaussian criterion was found to be fairly accurate. Furthermore, three different types of noise amplification curves have been identified and these three curves explain the qualitative difference in stability properties of the traditional first-, second- and high-order modulators. In addition, chaotic first- and second-order modulators behave like high-order modulators, i.e., the stability is lost abruptly and irreversibly. This interpretation of the loss of stability is not linked to the stability of the linearized system (i.e., the poles being inside the unit circle). In fact, for a high-order modulator, the stability is lost at a point where the linearized system is absolutely stable.

The presented modeling framework was verified on a class of modulators designed using standard Chebychev II filters as prototypes for the noise transfer function.

Finally, an algorithm was presented which can predict the probability density function of the quantizer input. This algorithm assumes the quantization noise to be an independent and identically distributed (i.i.d.) sequence.

Chapter 7

Optimizing Feedback Filters

7.1 Introduction

This chapter is devoted to the very complex subject of optimization of feedback filters. Sec. 6.7 reviewed a very simple but quite efficient design method, namely the use of NTF-prototypes obtained from standard filter design packages. Especially the use of Chebychev II prototypes seemed very promising. Once the width of the base-band has been decided, the only free parameter is the base-band attenuation `Rs`. When `Rs` is increased, the in-band noise power is lowered at the expense of a deteriorated stability. The stability can be expressed in terms of the maximum stable amplitude (MSA) and it can be predicted with some uncertainty using the Gaussian criterion of Eq. (6.20).

When an Nth order modulator is to be designed there are many degrees of freedom: The choice of N poles and $N-1$ zeros. The feedback filter poles becomes the zeros of the closed loop noise transfer function NTF(z). The best choice of feedback filter poles is to distribute them close to the unit circle in the base-band. The actual pole locations within the narrow base-band have very little affect on the stability properties. Consequently, the pole locations can be optimized separately. This has been done in [58] where analytical expressions for optimum (least in-band power) locations are listed up to $N = 5$ accompanied by numerical results for $N = 6, 7, 8$. The performance gain compared to N coincident poles at $z = 1$ is as high as 34 dB for $N = 8$.

The remaining degrees of freedom are the choice of $N-1$ feedback filter zeros. Recall that realizability considerations requires that the feedback filter has at least one sample delay and this gives only $N-1$ zeros. In addition, modulators are invariant to feedback filter scaling, so the actual number of free parameters is $N-1$. The zeros both affect the noise suppression performance and the stability. The design task can thus be described as an $N-1$ dimensional search for the 'best' trade-off between stability and SNR. The simple NTF-prototype method using Chebychev II filters is just a one dimensional curve in the $N-1$ dimensional parameter space and the contours for constant minimum noise amplification A_{min} are generally $N-2$ dimensional (hyper) surfaces.

An exhaustive search of the entire parameter space including extensive simulations for each visited parameter combination becomes rather utopian as the modulator order increases. This necessitates that the optimization must rely on performance and stability measures which are easily computable. The purpose of such optimization is to design modulators suited for practical use, i.e., the modulators should operate reliably with band-limited inputs within a certain amplitude range and the SNR should be optimal in respect to the give stability/reliability constraints. Recall from Sec. 3.8 that reliability basically means that the transition from stable to unstable operation is 'sharp', i.e., there is no extended amplitude range where the modulator is only weakly unstable with a very low escape rate. A low escape rate means that the mean survival time is long but finite, i.e., the modulator is seemingly stable for most short simulations. It is obvious that a modulator intended for a general purpose data converter cannot be tested for all possible input signals by simulations. However, if the modulator is absolutely stable for every constant input in the actual input amplitude range, there is a good chance that the modulator also can handle time varying signals with a suitable band-limitation. Even an exhaustive simulation test with constant inputs requires a lot of computational power.

The optimization methods presented become actually more and more necessary as the dimension increases further because the NTF-prototype method results in more and more unreliable modulators.

7.2 The Choice of Parameter Space

The coordinates for the $(N-1)$-dimensional parameter space of an Nth order modulator with fixed feedback filter poles can be chosen in many ways. The actual z-domain polynomial coefficients are a bad choice due to the extreme sensitivity to the zero locations. A better approach is to describe the zero locations directly. This can be accomplished by a suitable transform. E.g., for a fourth order modulator, the three zero locations z_1, z_2 and z_3 can be found using the transform:

$$z_1 = t_1 + \sqrt{t_2}\,, \quad z_2 = t_1 - \sqrt{t_2}\,, \quad z_3 = t_3 \tag{7.1}$$

where t_1, t_2 and t_3 are real parameters.

The third zero z_3 is always real and equal to t_3. The zero pair z_1 and z_2 are complex conjugates for negative t_2-values and real valued for positive t_2. There is a one-to-one correspondence between the three real valued parameters and the three zero locations of which two might be complex. Furthermore, there is a 'smooth' transition from complex to real zero locations.

The parameter to zero location transform of Eq. (7.1) can easily be extended to any dimension.

7.3 BIBO stability

The Gaussian stability criterion presented in Sec. 6.5 was related to the noise power amplification factor $\mathrm{A}(K)$ which is the squared two-norm of the noise transfer function of the linearized system. This stability criterion was derived by assuming the signals in the modulator to be stochastic signals. Hence, the Gaussian criterion is based on mean (squared) signal values. As stated previously, such criterion is only approximate: it is not sufficient and nor necessary. The purpose of this Section is to discuss a *bounded input bounded output — BIBO* criterion which is based on a worst case assumption.

For conceptual reasons it is convenient to transform the usual Σ-Δ modulator to a new equivalent circuit which is shown in Fig. 7.1. The circuit consists of an input prefilter $F(z)$ modifying the input to the signal $\hat{x}(k)$ which is fed to a so-called Noise-Shaper [38, 41, 62, 63]. The quantization noise $q(k)$ is found by subtracting the quantizer output from the quantizer input $e(k)$ and the quantization noise is fed back through a filter $F(z)$

means that the transition from stable to unstable operation is 'sharp', i.e., there is no extended amplitude range where the modulator is only weakly unstable with a very low escape rate. A low escape rate means that the mean survival time is long but finite, i.e., the modulator is seemingly stable for most short simulations. It is obvious that a modulator intended for a general purpose data converter cannot be tested for all possible input signals by simulations. However, if the modulator is absolutely stable for every constant input in the actual input amplitude range, there is a good chance that the modulator also can handle time varying signals with a suitable band-limitation. Even an exhaustive simulation test with constant inputs requires a lot of computational power.

The optimization methods presented become actually more and more necessary as the dimension increases further because the NTF-prototype method results in more and more unreliable modulators.

7.2 The Choice of Parameter Space

The coordinates for the $(N-1)$-dimensional parameter space of an Nth order modulator with fixed feedback filter poles can be chosen in many ways. The actual z-domain polynomial coefficients are a bad choice due to the extreme sensitivity to the zero locations. A better approach is to describe the zero locations directly. This can be accomplished by a suitable transform. E.g., for a fourth order modulator, the three zero locations z_1, z_2 and z_3 can be found using the transform:

$$z_1 = t_1 + \sqrt{t_2}\,, \quad z_2 = t_1 - \sqrt{t_2}\,, \quad z_3 = t_3 \tag{7.1}$$

where t_1, t_2 and t_3 are real parameters.

The third zero z_3 is always real and equal to t_3. The zero pair z_1 and z_2 are complex conjugates for negative t_2-values and real valued for positive t_2. There is a one-to-one correspondence between the three real valued parameters and the three zero locations of which two might be complex. Furthermore, there is a 'smooth' transition from complex to real zero locations.

The parameter to zero location transform of Eq. (7.1) can easily be extended to any dimension.

7.3 BIBO stability

The Gaussian stability criterion presented in Sec. 6.5 was related to the noise power amplification factor $\mathrm{A}(K)$ which is the squared two-norm of the noise transfer function of the linearized system. This stability criterion was derived by assuming the signals in the modulator to be stochastic signals. Hence, the Gaussian criterion is based on mean (squared) signal values. As stated previously, such criterion is only approximate: it is not sufficient and nor necessary. The purpose of this Section is to discuss a *bounded input bounded output — BIBO* criterion which is based on a worst case assumption.

For conceptual reasons it is convenient to transform the usual Σ-Δ modulator to a new equivalent circuit which is shown in Fig. 7.1. The circuit consists of an input prefilter $F(z)$ modifying the input to the signal $\hat{x}(k)$ which is fed to a so-called Noise-Shaper [38, 41, 62, 63]. The quantization noise $q(k)$ is found by subtracting the quantizer output from the quantizer input $e(k)$ and the quantization noise is fed back through a filter $F(z)$

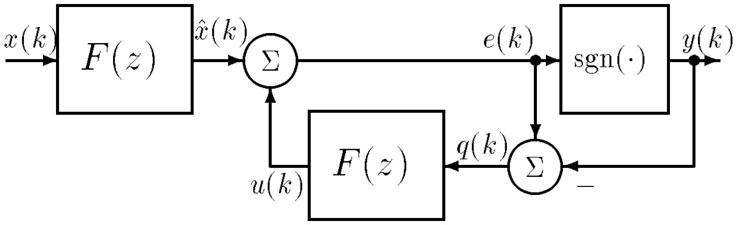

Figure 7.1: A Noise Shaper with a linear prefilter. The system is equivalent to a usual Σ-Δ modulator.

and added to $\hat{x}(k)$ in order to form $e(k)$. A circuit analysis reveals that the entire system is equivalent to a usual Σ-Δ modulator having a feedback filter:

$$H(z) = \frac{F(z)}{1 - F(z)} \tag{7.2}$$

Having $H(z)$ as starting point, the $F(z)$ filter can be found as:

$$F(z) = \frac{H(z)}{1 + H(z)} \tag{7.3}$$

Since $H(z)$ usually has a very high low frequency gain, it is seen that $F(z = 1) \approx 1$. This means that $F(z)$, as prefilter regarded, is quite harmless at low frequencies.

The first step for the BIBO criterion is to assume that the maximum magnitude of the quantization noise is below unity, i.e., $\max(|q(k)|) = \|q(k)\|_\infty \leq 1$. This assumption is only justified as long as the quantizer input magnitude $e(k)$ is below 2. Using a worst case limit of the quantizer input magnitude the BIBO criterion becomes:

$$\|f(k)\|_1 \cdot \|q(k)\|_\infty + \|\hat{x}(k)\|_\infty \leq 2 \tag{7.4}$$

where $\|f(k)\|_1$ is the one-norm of the impulse response of the $F(z)$ filter, i.e., the sum of the magnitude of the impulse response:

$$\|f(k)\|_1 = \sum_k |f(k)| \tag{7.5}$$

When Eq. (7.4) is fulfilled, the maximum quantization noise magnitude is unity:

$$\|\hat{x}(k)\|_\infty \leq 2 - \|f(k)\|_1 \tag{7.6}$$

This gives an upper bound on the stable input amplitude range, i.e., the one-norm of $F(z)$ must be as low as possible. The criterion of Eq. (7.6) is sufficient but rather conservative in most cases [58, 59]. This one-norm based stability criterion was proposed in [4] and further improved in [60].

The BIBO criterion can be improved (i.e., made less conservative) by using the modulator invariance to positive scalings of the feedback filter $H(z)$ [59]. Hence, $H(z)$ can be scaled by any positive number K in Eq. (7.3) in order to get different $F(z)$ filters resulting in exactly the same modulator with exactly the same stability properties:

$$F_K(z) = \frac{K \cdot H(z)}{1 + K \cdot H(z)} \tag{7.7}$$

Comparing to Eq. (6.3) it is seen that:

$$F_K(z) = \frac{1 + K \cdot H(z)}{1 + K \cdot H(z)} - \frac{1}{1 + K \cdot H(z)} = 1 - \text{NTF}_K(z) \tag{7.8}$$

Notice that the filter $F_K(z)$ is equal to the signal transfer function $\text{STF}_K(z)$, cf. Eq. (6.2) for quantizer gain K and $G(z) = H(z)$; however in this context K is just an arbitrary scaling factor.

The one-norm of $\text{NTF}_K(z)$ is defined as $S(K)$:

$$S(K) \triangleq \|\text{ntf}_K\|_1 = \sum_k |\text{nft}_K(k)| \tag{7.9}$$

Since $\text{ntf}_K(0) = 1$, it is noted that:

$$\|f_K(k)\|_1 = S(K) - 1 \tag{7.10}$$

This relationship leads to an improved BIBO criterion:

$$\|\hat{x}(k)\|_\infty \leq 3 - \text{S}_{min} \tag{7.11}$$

where S_{min} is the global minimum of $S(K)$.

The improved BIBO criterion is quite analogous to the Gaussian criterion: the BIBO criterion is based on the minimum one-norm of the NTF and the Gaussian criterion is based on the minimum two-norm of the NTF. A guaranteed zero-input stable modulator has $\text{S}_{min} \leq 3$ and corollary: $\text{A}_{min} \leq 3^2$. Since the leading ntf_K term is unity, it is concluded that $\text{A}_{min} \leq (1 + 2^2)$. The Gaussian criterion suggests that $\text{A}_{min} < 2.75$ for a zero input stable modulator. The difference between the norms is that the two-norm gives more weight to the big elements of the impulse response due to the squaring. The one-norm gives equal weights to all elements. This difference is reflected in the two criteria: the Gaussian two-norm based criterion works fine in many cases but it gives no guarantee what soever. The one-norm based BIBO criterion is ironclad but often too conservative because it is based on the worst case extremal quantization noise sequence. It is often very unlikely that this worst case sequence will occur in practise. The BIBO criterion can be further improved by including exact knowledge of the dependence between adjacent quantizer input samples. However, this approach is difficult and not very general. Such methods try to identify a trapping region in the state-space, i.e., a orbits starting inside the trapping region never leave it. Specialized stability tests based on these principles have been developed for second order modulators [67, 59]; see also Example 3.4.

Example 7.1 This example demonstrates the BIBO criterion for some simple first- and second-order modulators: the chaotic first order modulator with $H_1(z) = z^{-1}/(1 - 2z^{-1})$ and the usual second order modulator with $H_2(z) = (z^{-1} - 0.5z^{-2})/(1 - 2z^{-1} + z^{-2})$. The $S(K)$-curves for these two filters are shown

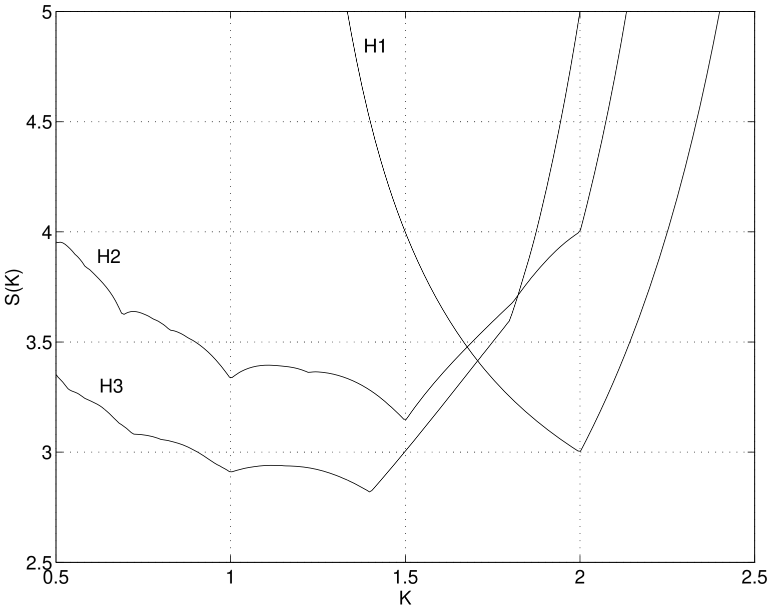

Figure 7.2: $S(K)$-curves for three feedback filters: $H_1(z) = z^{-1}/(1-2z^{-1})$, $H_2(z) = (z^{-1}-0.5z^{-2})/(1-2z^{-1}+z^{-2})$ and $H_3(z) = (z^{-1}-0.6z^{-2})/(1-2z^{-1}+z^{-2})$

on Fig. 7.2. The first order modulator $H_1(z)$ is marginally stable since $S(2) = \mathrm{S}_{min} = 3$. In this special case the Gaussian criterion is more conservative that the BIBO criterion.

The BIBO criterion cannot guarantee the stability of the popular second-order modulator $H_2(z)$ since $\mathrm{S}_{min} > 3$. Both $H_1(z)$ and $H_2(z)$ have noise transfer functions that become FIR filters for a suitable K-value, i.e., $K = 2$ in this case. There is a general class of feedback filters having this property. Every second-order modulator with this FIR-property conforms to Eq. (3.5). Every first order modulator has the FIR property, and furthermore, the S_{min}-value is taken when the noise transfer function is FIR. The second-order modulator $H_2(z)$ has $S(2) = 4 > \mathrm{S}_{min}$ in the FIR-case, i.e., the scaling giving FIR NTF is not the optimal scaling.

The zero of $H_2(z)$ can be modified slightly such that the second-order modulator is guaranteed stable for low input amplitudes, e.g., the $S(K)$ curve of $H_3(z) = (z^{-1}-0.6z^{-2})/(1-2z^{-1}+z^{-2})$ is also plotted on Fig. 7.2. This modulator has $\mathrm{S}_{min} \approx 2.82$ giving a guaranteed stable maximum input amplitude of 0.18. This stability enhancement takes place at the expense of a deteriorated in-band noise suppression due to a deteriorated low-frequency loop-gain □

7.4 Reliability

A practical constraint in feedback filter optimization is that the modulator is suitably reliable, i.e., the onset of instability occurs abruptly when the input amplitude is increased. The reliability ensures that the MSA values found by simulations using slowly

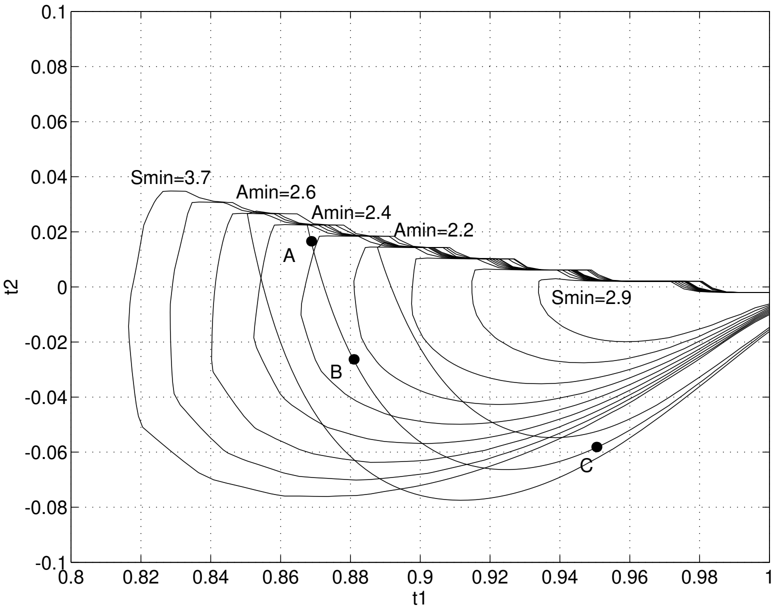

Figure 7.3: Contours of equal S_{min} and A_{min} for $t_3 = 0.7$. The parameters t_1, t_2 and t_3 determine the feedback filter zeros, cf. Eq. (7.1). The pole locations are given by Eq. (7.12). The S_{min} contours are shown for increments of 0.1 from 2.9 to 3.7. The A_{min}-contours are for 2.2, 2.4 and 2.6. Three points (A,B and C) are marked on the $\mathrm{A}_{min} = 2.4$ contour.

increasing ramp input will hold in practise when the modulator operates at full speed (e.g., 3 MHz sample rate) with input signals from 'the real world'. A reliable modulator gives a simulated MSA which is fairly independent of the dc sweep rate. Conversely, an unreliable modulator can exhibit an acceptable MSA for high sweep rates corresponding to short simulations — but longer simulations may reveal instability for much lower input amplitudes.

The question is now: how can feedback filters of reliable and unreliable modulators be distinguished from each other? Numerous simulations have indicated that the two-norm based Gaussian criterion predicts the onset of severe instability fairly well, i.e., the Gaussian criterion is consistent for high sweep rates. However, different filters with the same A_{min}-value may behave differently with respect to reliability.

The existence of extended input ranges with low escape rates is one of the characteristics of unreliable modulators. A low escape rate means that the system only escapes when the orbit enters certain rarely visited regions in the state-space, i.e., it is a worst case situation which occurs rarely. This situation might arise when the minimum one norm S_{min} is composed of many small contributions, i.e., when the NTF impulse response has slowly decaying (high-Q) oscillations. The worst case quantizer input can be very large but it is, on the other hand, very seldom that values close to the extremal occur. A more reliable modulator is expected to exhibit a suitably low S_{min}-value dominated by the first few terms of the NTF impulse response.

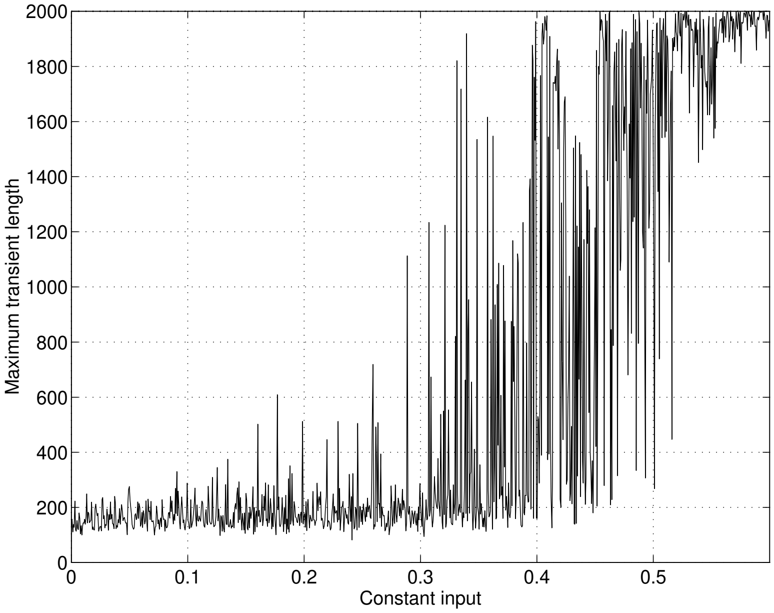

Figure 7.4: Plot of maximum transient length versus constant input for the modulator A, cf. Fig. 7.3. For simulation parameters see Example 7.2.

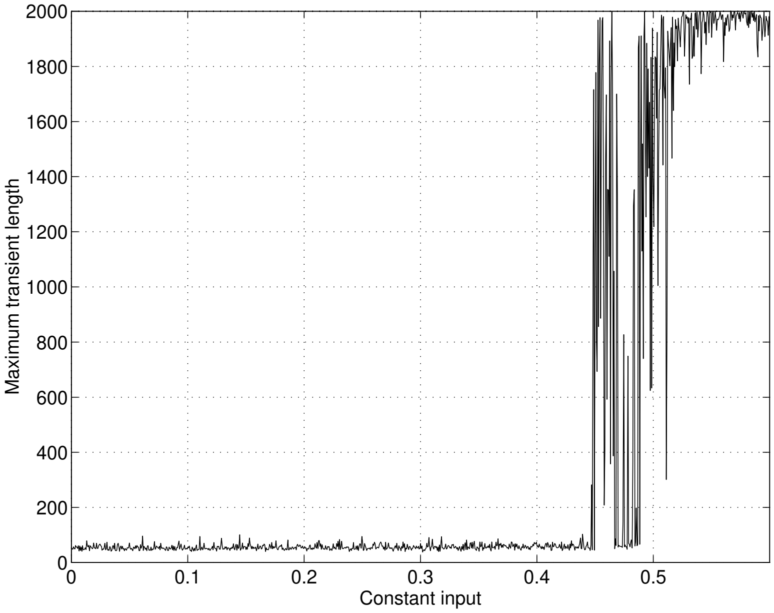

Figure 7.5: Plot of maximum transient length versus constant input for the modulator B, cf. Fig. 7.3. For simulation parameters see Example 7.2.

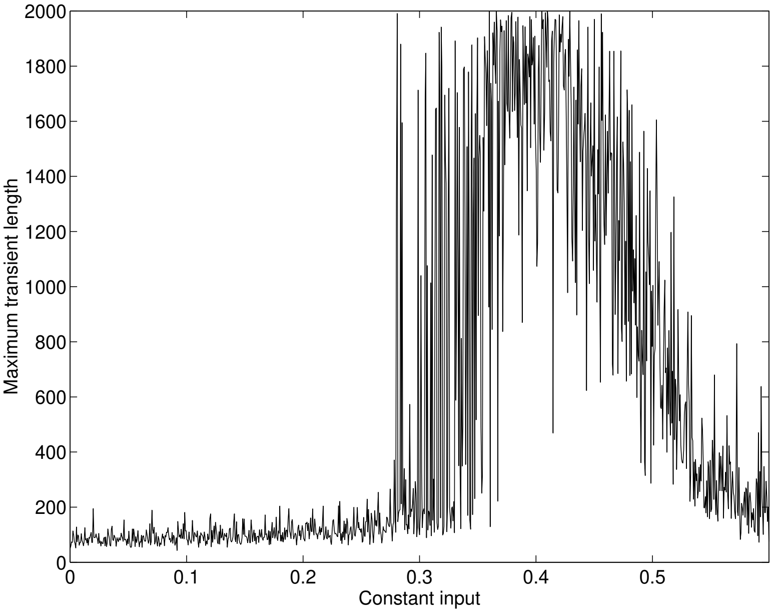

Figure 7.6: Plot of maximum transient length versus constant input for the modulator C, cf. Fig. 7.3. For simulation parameters see Example 7.2

Example 7.2 This example demonstrates the behavior of three fourth order modulators with $\mathrm{A}_{min} = 2.4$. The feedback filter poles are suited for 64 times oversampling, i.e., base-band $f_b = 20\,\mathrm{kHz}$ and sample rate $f_s = 64 \cdot 48\,\mathrm{kHz}$. The four optimized unit circle feedback filter poles are located at the frequencies [58]:

$$f_i \in \left\{ \pm\sqrt{\frac{3}{7} \pm \sqrt{\left(\frac{3}{7}\right)^2 - \frac{3}{35}}} \cdot f_b \right\} \tag{7.12}$$

These pole frequencies should minimize the unweighted total base-band noise power.

The (t_1, t_2, t_3) parameter space was investigated with the optimized poles for $f_b = 1/64 \cdot 20/24$ relative to half the sample rate. Fig. 7.3 shows two-dimensional contours for equal S_{min} with t_3 fixed to 0.7. The graph includes the contour for $\mathrm{A}_{min} = 2.2$, 2.4 and 2.6. There is a lot of similarity between the types of contours: the gradients are in almost the same directions, i.e., both the A_{min} and S_{min} measures will in most cases agree on whether a given parameter change enhances the stability or not. However, the two stability measures are still quite different in nature. This will be examined by following one of the A_{min}-contours. Three points A, B and C on the 2.4 contour are marked. Point B has minimum $\mathrm{S}_{min} = 3.24$ for $t_3 = 0.7$ and point A and C have somewhat higher values, namely 3.31 and 4.7, respectively. The modulators of the three points are not guaranteed stable for zero input according to the BIBO criterion but they should all have an MSA of approx. 0.38 according to the Gaussian criterion of Eq. (6.20).

The stability of the modulators was investigated in two ways: simulations with slowly increasing input ramp and an exhaustive search for long transients. The latter method was introduced in Sec. 5.2. The modulator to be tested is simulated for a large number of random initial conditions and randomly selected constant inputs for a maximum number of time-steps. If the quantizer input magnitude exceeds an upper bound of 10, the current number of time steps simulated is regarded as a transient length. The maximum transient length is found within each of a number of constant input bins. A plot of the maximum transient length versus the constant input magnitude gives a kind of fingerprint of the stability and reliability of the modulator. Such plots are shown on Fig. 7.4, 7.5 and 7.6 for the three modulators corresponding to point A, B and C. Each of the plots are obtained from 5 mill. simulations with random constant input distributed uniformly over the interval $[0, 0.6]$. The initial conditions were randomly selected using a uniform distributions in the quantizer input delay-space (i.e., the coordinates $\boldsymbol{e}_0 = [e(-1), ...e(-4)]$) such that $\|\boldsymbol{e}_0\|_\infty = 0.5$. The simulations were stopped either when the quantizer input exceeded a magnitude of 10 or after a maximum of 2000 time-steps. The maximum transient length was found within 1000 constant input bins.

The three maximum transient plots convey a lot of information: modulator B is very nice and is expected to operate reliably for input magnitudes up to approx. 0.45 where there is a transition region between 0.45 and 0.5. For magnitudes above 0.5 the modulator is definitely unstable. The maximum transient 'background level' is low and very constant for magnitudes less than 0.45. For this rather reliable modulator the Gaussian criterion is somewhat pessimistic.

The plots for modulator A and C have a quite different behavior: the graphs have a lot of spikes of different sizes for most of the entire input range. Furthermore, the background levels for both graphs are significantly higher than for modulator B. It is obviously difficult to judge the MSA for these very unreliable modulators. Simulations with ramp type input rising from zero to unity during 10^8 samples (corresponding to approx. 30 seconds real-time) have revealed that modulator B becomes unstable at input magnitudes as low as 0.33 and 0.26 for modulator C. Modulator B can handle input magnitudes up to 0.44 even for these very long ramp simulations. Such slowly rising ramp input is realistic for signals from, e.g., temperature and pressure transducers. The question is of course how these three modulators will behave generally in a real application. There is only one way to give an exhaustive answer: implement the circuits and test the devices for extremely long time. However, it seems likely that the more reliable modulator B is absolutely stable when operating with suitably band-limited signals having a maximum magnitude below 0.4.

Another observation from the maximum transient simulations was that the fraction of initial conditions surviving the maximum 2000 time steps vary from modulator to modulator. Fig. 7.7 shows the surviving fraction within each of the 1000 input bins for modulator B and C. It is seen that the surviving fraction for low input magnitude is almost 5 times higher for modulator B than for modulator C and this indicates that the volume of the basin of attraction for the unreliable modulator C is lower, i.e., it is more difficult to find stable initial conditions.

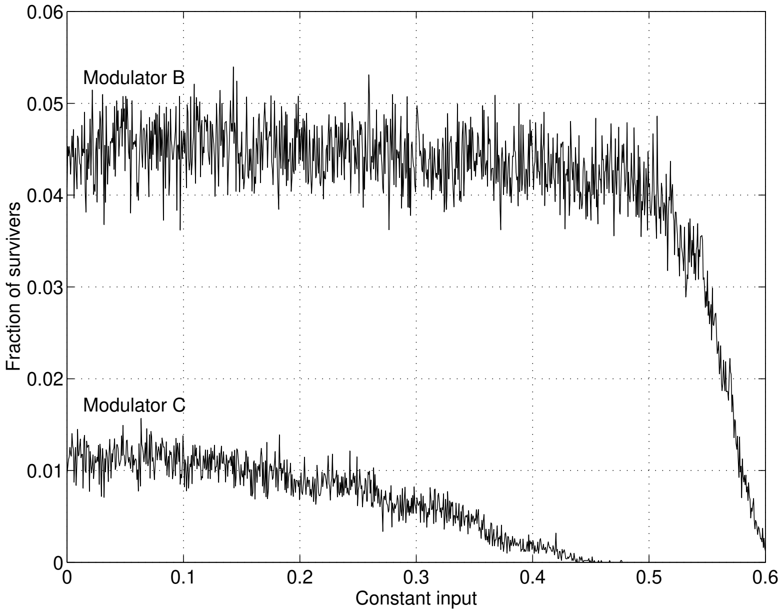

Figure 7.7: Fraction of initial conditions surviving 2000 time steps for the modulators B and C, cf. Fig. 7.3. For simulation parameters see Example 7.2

The surviving fraction for modulator B is fairly constant up magnitudes near 0.5. For higher magnitudes the fraction decreases fast to near zero for magnitudes around 0.6. This reflects a sharply increasing escape rate in this transition region. □

The previous example demonstrated the behavior of reliable and unreliable modulators. The very high S_{min}-value for modulator C can clearly be taken as an indication of unreliability. In fact, the most reliable modulator had the lowest S_{min}-value. More detailed information is probably conveyed by the entire NTF impulse response which corresponds to the S_{min}-value. In fact, the NTF impulse response fully describes the modulator and the question of stability is therefore 'just' a matter of proper interpretation.

Example 7.3 The $\mathrm{ntf}_{min}(k)$ impulse responses that gives the S_{min}-values for the modulators A, B and C of Fig. 7.3 are plotted on Fig. 7.8. By definition $\mathrm{ntf}_{min}(0) = 1$ due to the one sample feedback delay. Furthermore, $\mathrm{ntf}_{min}(1) = -Kc_1$ where c_1 is the first $H(z)$ numerator term which is normalized to unity in this example. Consequently, for the shown plots, $\mathrm{ntf}_{min}(1) = -K_{Smin}$ where K_{Smin} is the K-value that yields S_{min}. The K_{Smin}-values are all near 1.5 for the three modulators. Hence, the first two dominating terms make up a sum of approx. 2.5 and this shows that it is the 'tails' of the NTF impulse responses which are responsible for the instability and possible unreliability. First order modulators have FIR NTF, i.e., they have no 'tails' and the onset of instability is consequently very abrupt.

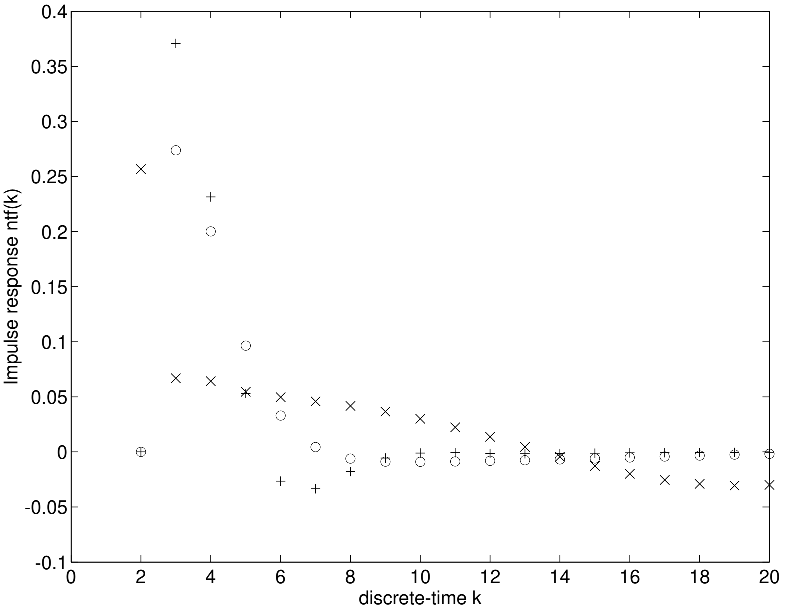

Figure 7.8: Plot of the $\mathrm{ntf}_K(k)$ impulse responses which achieves the S_{min}-values for modulator A(+), B(o) and C(×), cf.Fig. 7.3. The plot shows only the tails for $k \geq 2$.

For both modulator A and B, the third term (i.e., $\mathrm{ntf}_{min}(2)$) is zero and non-zero for modulator C. This is probably a consequence of the one-norm minimization.

The very high S_{min}-value for modulator C is due to a slowly decaying oscillation which probably causes the unreliability. It is more difficult to explain the cause of unreliability for modulator A. However, a comparison between the NTF impulse responses of modulator A and B reveals that the 'tail' of modulator B decays to zero without oscillation while the 'tail' of modulator A is more oscillating before it dies out. The worst case occurs when the past quantization noise values have signs which fit the signs of the NTF impulse response. This can cause the quantization noise to increase and blow up the modulator. Modulator B is probably more reliable and stable because of the almost unipolar tail, i.e., it is very unlikely, in normal stable operation, that a modulator produces long sequences of unipolar output codes. Conversely, oscillating code patterns are more likely to occur and this can produce a worst case situation for modulators with oscillating tails □

The simple one-norm S_{min} can explain some but not all of the differences in reliability. The observations done in the previous example suggests that the NTF impulse responses should be weighted suitably in order to provide a measure which indicates the degree of reliability. A possibility is to examine the NTF impulse response in the spectral domain. A necessary condition for BIBO stability is that the maximum magnitude of the filter $F_K(z)$ is below 2, i.e., when the one-norm of $f_K(k)$ is below 2 as required for BIBO stability, the magnitude response of $F_K(z)$ is likewise bounded by 2.

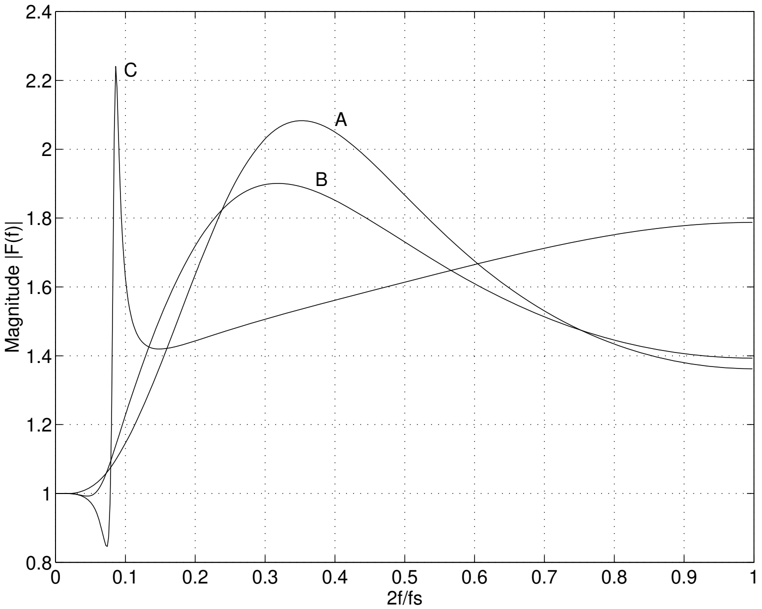

Figure 7.9: Magnitude responses of the $F_{min}(z)$ filters with minimum one-norm for modulator A, B and C, cf. Fig. 7.3.

> **Example 7.4** On Fig. 7.9 the magnitude responses of the $F_{min}(z)$ filters with minimum one-norm for the A,B and C modulators (se Fig. 7.3 are plotted. It is seen that modulator B is the only modulator with a magnitude maximum below 2. The oscillating behavior of modulator C is seen as a strong peak with a magnitude higher than 2. The oscillating behavior is also seen on Fig. 7.10 which shows the simulated output spectrum of modulator C. This spectrum has clearly a similar strong peak. The maximum magnitude for modulator B is well below 2 and this probably explains the better reliability. □

The previous example showed that the maximum magnitude of $F_K(z)$ is a good measure of reliability. This motivates the definition:

$$\mathrm{F}_{max} = \max\left\{\left|F_{min}(e^{j\pi f})\right|\right\} = \|F_{min}(z)\|_{\infty} \tag{7.13}$$

The F_{max} measure is thus the infinity norm of the $F_{min}(z)$ filter having minimum one-norm.

> **Example 7.5** Fig. 7.11 shows contours of constant F_{max} for $t_3 = 0.7$. The remaining parameters conform to Example 7.2 and Fig. 7.3 □

The three measures A_{min}, S_{min}, and F_{max} are based on the two, one and infinity norms, respectively. The combination of these three figures gives a quite good picture of the stability of a modulator. The unreliable modulators are typically characterized by the F_{max} and S_{min} values being too high compared to the A_{min} value.

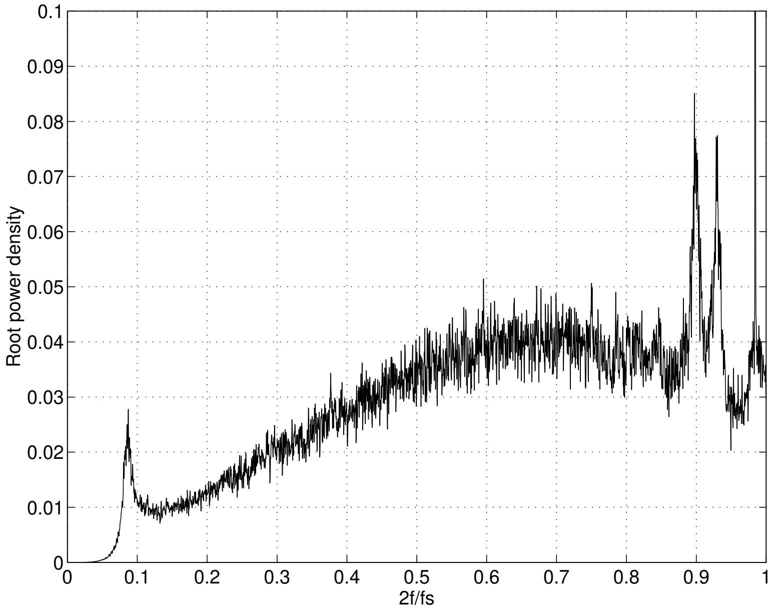

Figure 7.10: Power spectrum estimate for the simulated output of modulator C, cf. Fig. 7.3. A constant input of 1/64 was used. Notice the linear axes and the peak near $2f/f_s \approx 0.095$

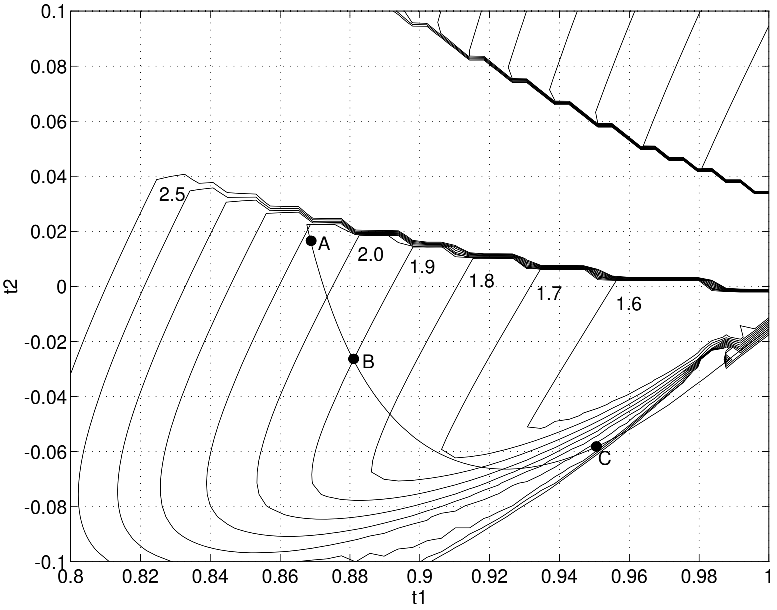

Figure 7.11: Contours of constant F_{max} for $t_3 = 0.7$. The $A_{min} = 2.4$ contour is plotted and the points A,B and C, cf. Fig. 7.3, are marked.

7.5 Performance Prediction

Optimization of loop filters will normally comprise a minimization of the in-band quantization noise power. A brute force method is to simulate the system and estimate the in-band noise power using a filter or an FFT spectral analysis. However, this approach can be too time consuming for use in connection with an automated optimization. Furthermore, for a finite length simulation, the obtained in-band noise power as a function of the filterparameters is not very 'smooth' due to the stochastic nature of the modulator. This effect can literally 'derail' an optimization routine which is based on numerically estimated gradients.

The use of linearized modeling can provide a computational shortcut to performance prediction. The linearized loop model discussed in Ch. 6 shows how the quantization noise is shaped and suppressed in the base-band. The apparent problem is that either the quantizer gain K or the feedback filter scaling can be chosen arbitrarily in such models. It has been suggested to use a fixed quantizer gain of 2 for a feedback filter scaling where the leading term of the feedback filter impulse response is unity [56].

The next problem is to predict the variance and spectrum of the quantization noise in order to compute the base-band noise power. The quasilinear modeling framework offers a fairly simple and yet very precise way to predict the noise spectrum of the modulator: once the equilibrium point (e.g., the quantizer gain K) is known, the quantization noise power as well as the noise transfer function is known and the in-band noise power can subsequently be found by integration. It was demonstrated in Sec. 6.8 that the equilibrium point can be predicted very well using an algorithm which predicts the quantizer input probability distribution function. However, this algorithm is far to complex and slow for use in a loop filter optimization procedure.

A simpler approach is to use the fact that most modulators approximately follows the same $\mathrm{A}(m_y)$-curve: for zero input, most modulators operate with equilibrium $\mathrm{A}_{eq} \approx 3.49$ (see Fig. 6.10). This rule of thumb applies to most 'good' and 'well behaved' modulators. The rule is typically less accurate for modulators having a very high stable amplitude range.

For a K_{eq} satisfying $\mathrm{A}(K_{eq}) = \mathrm{A}_{eq}$ the base-band quantization noise power σ_b^2 for zero input is given by:

$$\sigma_b^2 = \frac{1}{\mathrm{A}_{eq}} \int_0^{f_b} \left| \frac{1}{1 + K_{eq} \cdot H(e^{j\pi f})} \right|^2 df \tag{7.14}$$

where f_b is the upper base-band limit frequency relative to half the sample rate.

This gives an estimate of the in-band noise power for small amplitude input. The in-band noise power will increase somewhat for higher amplitudes.

Example 7.6 Fig. 7.12 shows contours of equal σ_b^2, cf. Eq. (7.14) with $\mathrm{A}_{eq} = 3.49$, in the (t_1, t_2) parameter plane with $t_3 = 0.7$. The feedback filter poles are given by Eq. (7.12) with $f_b = 1/64 \cdot 20/24$. The points A, B and C, cf. Fig. 7.3, are marked. The three modulators were simulated with a constant input of 1/64 and the output spectra was estimated using averaged and windowed 8k FFT's. The small dc-bias ensures that possible tones are outside the base-band. The base-band noise power was estimated by summing up the power of

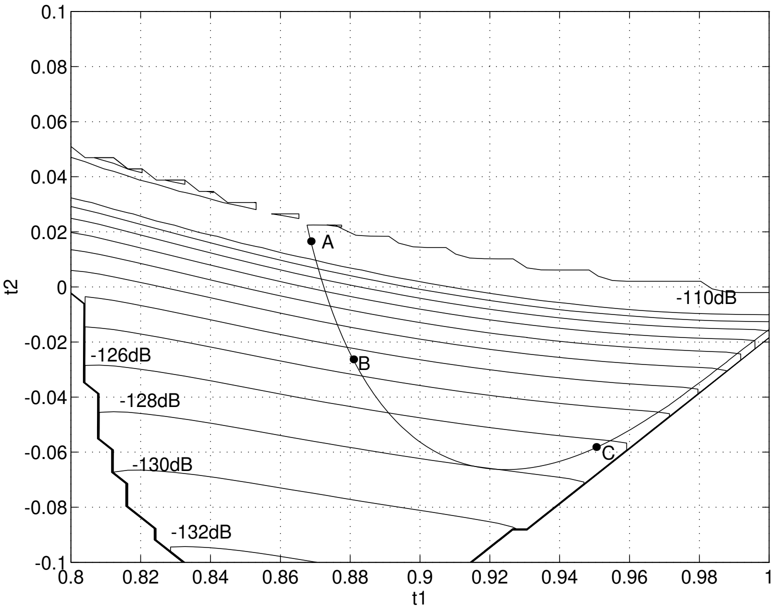

Figure 7.12: Contours of predicted σ_b^2 for $t_3 = 0.7$. The $\mathrm{A}_{min} = 2.4$ contour is also plotted and the points A,B and C, cf. Fig. 7.3, are marked.

the first $f_b \cdot 4096$ FFT-bins. The simulated and predicted values of σ_b^2 appears from the table:

Modulator	Predicted σ_b^2	Simulated σ_b^2
A	$-96.8\,\mathrm{dB}$	$-103.8\,\mathrm{dB}$
B	$-122.9\,\mathrm{dB}$	$-124.8\,\mathrm{dB}$
B	$-126.4\,\mathrm{dB}$	$-127.7\,\mathrm{dB}$

The predicted σ_b^2 is clearly to high in all cases. Especially for modulator A, the prediction is 7 dB to pessimistic. However, for modulators with reasonably high performance, the prediction is consistently only a few dB wrong.

It is seen that point A gives the highest in-band noise power and this is due to the feedback filter zeros being all real valued. The highest zero modulus is 0.998 and this almost cancels the effect of the dc-pole at $z = 1$. Furthermore, this modulator is also proven to be very unreliable. Point B and C with negative t_2 have a complex zero pair giving a higher low frequency loop gain and thereby a better in-band noise rejection. The difference between B and C is only a few dB. When the known unreliability and lower MSA of point C is taken into consideration, the modulator of point B must clearly be preferred.

Following the A = 2.4 contour on Fig. 7.12, is can be seen that a σ_b^2 minimum is found near the point $(0.92, -0.066)$. It appears from Fig. 7.3 and Fig. 7.11 that this point also gives acceptable stability and reliability properties □

Another useful approach for prediction of the equilibrium point, is that the K-value which minimizes the one-norm $\mathrm{S}(K)$ is often quite close to the equilibrium K found by simulations.

Another interesting observation is that the two-norm of the error transfer function $\mathrm{ETF}_K(z)$ (see Eq. (6.26)) has minimum close to the equilibrium K. Recall that the ETF given by Eq. (6.26) is the transfer function from the quantization noise source to the quantizer input, i.e., the modulator attempts to operate with minimum quantizer input variance. From the definition of the noise amplification factor A of Eq. (6.7) it can be derived that:

$$\|\mathrm{ETF}_K\|_2^2 = \frac{\mathrm{A}(K) - 1}{K^2} \tag{7.15}$$

The K-value which minimizes Eq. (7.15) can thus be used as K_{eq}. In some cases this approach gives very good performance predictions.

Example 7.7 The fifth order Chebychev II derived filter A from Table 6.2 has a very low zero input A-value. The minimum two-norm of the ETF_K is reached for $K = 2.05$ and this K gives A= 2.74. A simulation with zero input yields $K = 2.09$ and A= 2.86. However, Fig. 6.10 shows that for increasing input amplitude, the noise amplification factor is increasing towards a maximum above 3.5 for $m_y \approx 0.12$ □

7.6 Optimization Strategies

A loop filter optimization procedure should minimize the in-band noise power under suitable constraints. The constraints should first of all ensure that the modulator is stable and reliable in a prescribed amplitude range. A widely used constraint has been to impose an upper bound for the unity quantizer gain noise amplification factor $\mathrm{A}(K = 1)$ [3]. Other approaches have been to impose an upper bound on the maximum amplitude gain of the noise transfer function as used for NTF-prototypes in [10].

A better approach is to specify an upper bound for A_{min} determined by the Gaussian criterion and the desired MSA. This constraint takes the feedback filter scaling invariance into consideration. An optimization with this constraint alone will in the general case lead to an unreliable modulator with an MSA lower than prescribed by the Gaussian criterion. Consequently, suitable upper bounds for S_{min} and F_{max} must normally be specified in order to get a good result. In many cases, only some of the constraints are active at the same time, i.e., restricting the optimization. The rest of this section will demonstrate how a suitable mixture of one-, two- and infinity-norm based constraints leads to consistent results.

An optimization procedure employing the mentioned constraints was implemented in MATLAB using the Optimization Toolbox CONSTR routine [21] based on Sequential Quadratic Programming [1]. The approach is fairly robust and has been applied successfully to modulators with orders ranging from 3 to 8. A similar approach called CLANS (Closed Loop Analysis of Noise-Shaping coders) using CONSTR has been reported in [29] where a suitably weighted in-band noise power measure is minimized under a one-norm constraint. This approach was successfully used for optimized design of multibit noise-shapers. However, CLANS is not intended for one-bit modulators and the approach does not take the scaling invariance into consideration.

---

[1] The MATLAB tools are available from the author

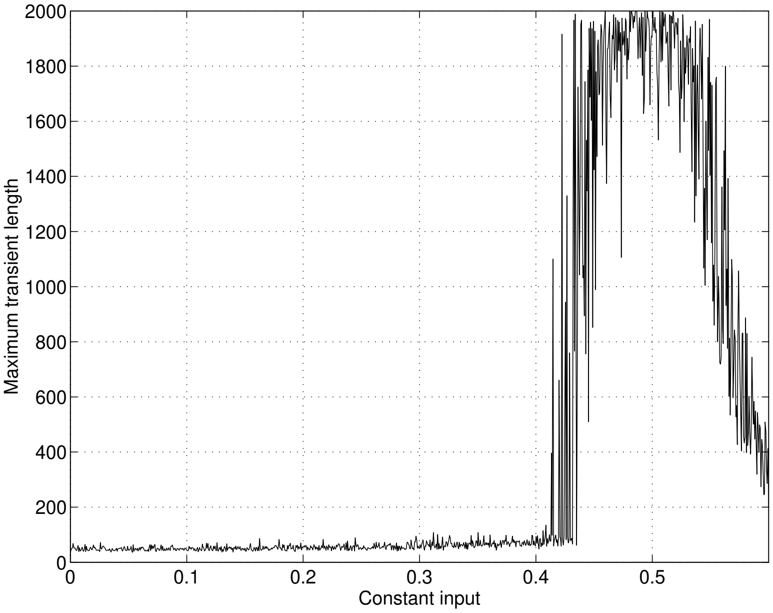

Figure 7.13: Maximum transient plot of an optimized modulator with the constraints $A_{min} < 2.4$, $F_{max} < 1.8$ and $S_{min} < 3.5$.

Example 7.8 A fourth order modulator was optimized for $f_b = 1/64 \cdot 20/24$. The feedback filter poles was found from Eq. (7.12) and the zeros were obtained from a constrained minimization of the predicted σ_b^2 using the constraints $A_{min} \leq 2.4$, $F_{max} \leq 1.8$ and $S_{min} \leq 3.5$. The base-band noise power was estimated using the assumption: $A_{eq} = 3.49$.

The resulting parameters are $(t_1, t_2, t_3) = (0.922, -0.0588, 0.6347)$. In this case, all of the three constraints were active (i.e., the upper bounds were reached). The found parameters satisfy consequently three equations with three variables and the parameters are hence given by the constraints only.

The predicted σ_b^2 was -128.7 dB and the simulated was -130.5 dB. Fig. 7.13 shows a maximum transient length plot for this modulator (same simulation parameters as for Example 7.2). The plot indicates that the optimized modulator has a reliable MSA of approx. 0.42 and that the onset of instability occurs abruptly. A simulation with a ramp increasing with a rate as low as 10^{-8}/sample confirms this MSA. The modulator achieves consequently a dynamic range of approx. 120 dB, or almost 20 bits of resolution. This performance is comparable to a recently reported fifth order modulator [17]. The simulated output spectrum of the optimized modulator for a constant input of 1/64 is seen on Fig. 7.14 □

In the previous example, all three constraints were active. When similar optimizations are carried out on third order modulators with the same constraints, it is typically the F_{max} limitation that is the dominating constraint, while the S_{min} constraint is inactive.

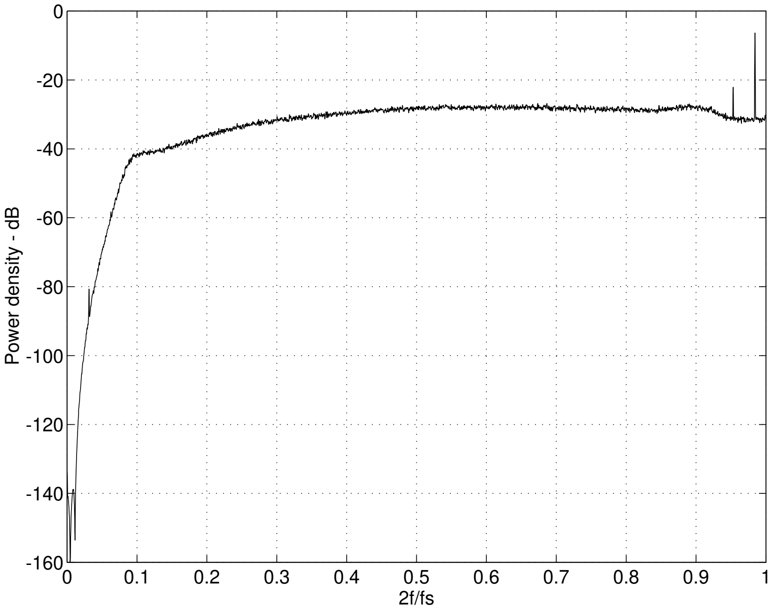

Figure 7.14: Estimated output power spectrum for the optimized modulator with the constraints $A_{min} \leq 2.4$, $F_{max} \leq 1.8$ and $S_{min} \leq 3.5$.

Conversely, for higher order optimizations ($N > 4$), the S_{min} constraint is dominating while the F_{max} constraint is typically inactive.

Example 7.9 An optimization was carried out for a third order modulator with the same f_b and constraints as in Example 7.8. The optimal feedback filter pole frequencies for third order are given by [58]:

$$f_i \in \left\{0,\ \pm\sqrt{\frac{3}{5}} \cdot f_b\right\} \tag{7.16}$$

The optimization gave the parameters $t_1 = 0.7752$ and $t_2 = -0.0663$. Only the $F_{max} \leq 1.8$ constraint was active: the optimization yielded $S_{min} = 3.22$ and $A_{min} = 2.32$. The predicted σ_b^2 was -96.7 dB and the simulated was -97.7 dB. A simulation with a ramp increasing with a rate of 10^{-8}/sample yielded an MSA of 0.41.

An optimized fifth order modulator was designed with the same constraints. The optimal pole frequencies are [58]:

$$f_i = 0,\ \pm\sqrt{\tfrac{5}{9} \pm \sqrt{\left(\tfrac{5}{9}\right)^2 - \tfrac{5}{21}}} \cdot f_b \tag{7.17}$$

The optimization gave the parameters:

$$(t_1, \cdots t_4) = 0.9697, -0.0328, 0.8001, -0.0452)$$

Both the $S_{min} \leq 3.5$ and $A_{min} \leq 2.4$ constraints were active while F_{max} was 1.787. The predicted σ_b^2 was -144 dB and the simulated was -144.7 dB. A simulation with a ramp increasing with a rate of 10^{-8}/sample yielded an MSA of 0.37. This corresponds an SNR of 133 dB for maximum amplitude sinusoidal input. A modulator for a commercial 64 times oversampling fifth-order A/D-converter was reported to yield a maximum simulated SNR of 119 dB for sinusoidal input with an amplitude of 0.4 [17]. The modulator has pole frequencies at 11kHz and 19kHz. However, the base-band used for the reported SNR figure was DC-22kHz with a sampling rate of $64 \cdot 48$ kHz and the SNR figure was actually measured with sinusoidal input. Even when these differences are compensated, the optimized modulator compares favorably to the reported commercial design.

For the sake of comparison, a fifth order modulator was designed using the NTF-prototype method described in Sec. 6.7. The feedback filter zeros were found using a Chebychev II prototype with Rs= 111.75 dB. The poles were given by Eq. (7.17). This modulator has $A_{min} = 2.4$, $S_{min} = 3.5$ and $F_{max} = 1.7$. The predicted σ_b^2 was -139.8 dB and the simulated was -140.97 dB. A simulation with a ramp increasing with a rate of 10^{-8}/sample yielded an MSA of 0.38. This modulator fulfills the constraints used for the optimization but the noise suppression is approx. 4 dB lower than for the optimized modulator $\square$

7.7 Considerations for Higher Order Designs

All of the four designs presented in Example 7.8 and Example 7.9 fulfill the same set of constraints and these four designs work reliably within nearly the same amplitude range in agreement with the Gaussian criterion. This shows that the used constraints are consistent for orders at least up to five.

It is generally true that a better performance is obtained by raising the order of the modulator for a given oversampling ratio (or base-band). The noise shaping theorem of Th. 6.1 gives a very simple explanation: For stability reasons, the high-frequency NTF gain must be limited (e.g., by bounding A_{min} or S_{min}) and the only way to improve the base-band rejection without violating the noise shaping theorem is to reduce the transition band from base-band to pass-band. An efficient way to reduce the transition band is to increase the filter order. As the transition band is reduced, both the decimation filter in A/D converters and the analog post-filters in D/A converters must be more steep and, hence, more complex.

Even with a 'brick-wall' NTF-characteristic, the noise shaping theorem and the need for suitable bounds on the NTF high-frequency gain limit the attainable performance. This is quite comforting, since it would be strange if it was possible to design a modulator yielding more bits or resolution than the oversampling factor (see Ch. 1). The rest of this section demonstrates the performance limit for modulators intended for 32 times oversampling.

Example 7.9 showed that the NTF-prototype method using Chebychev II filters gives good and reliable fifth order modulators. However, the optimization procedure obtained a 4 dB better result with almost the same stability. Unfortunately, it turns out that the Chebychev II prototype method gives more and more unreliable modulators for increasing order.

Example 7.10 A modulator order of eight is necessary for a CD-quality converter operating at only 32 times oversampling. An eight order Chebychev II prototype with fb= $1/32 \cdot 20/24$ and Rs= 101.5 dB achieves a A_{min} value of 2.4. Unfortunately, the S_{min} value is as high as 4.73 and this indicates that the modulator is rather unreliable. The F_{max} value is as low as 1.66. A ramp with a sweep rate of 10^{-6}/sample gave a MSA of 0.4 while a rate of 10^{-9}/sample yielded 0.21. The measured base-band noise power σ_b^2 for a constant input of 1/64 was measured to -128.9 dB. □

The previous example showed that loop filter optimization becomes mandatory for very high order modulators. The NTF-prototype method represents a one-dimensional curve in a very high dimensional parameter space. It is obvious that improvements can be obtained by exploring the entire parameter space.

Until now, the signal transfer function (STF) has received no attention. The STF as given by Eq. (6.2) will for the usual sigma-delta modulator normally be flat in the base-band due to the high base-band gain of the feedback filter. This is not the case for the popular multiple feedback topology (see Example 2.3). This topology is equivalent to a generic modulator (see Fig. 2.4) with $H(z) = C(z)/D(z)$ and $G(z) = 1/D(z)$ [49]. This means that a multiple feedback modulator needs a prefilter with a transfer function approximately equal to $C(z)$ in order to equalize the base-band STF. For low order designs with a high oversampling ratio, such equalization is not necessary because $C(z)$ is typically flat in the base-band. For higher order systems with lower oversampling ratios, the base-bandwidth is comparable to the transition bandwidth of the noise transfer function and there will generally be untolerable base-band ripple.

The necessary equalization can be accomplished digitally for both A/D and D/A converters. In an A/D-converter, the equalization is performed digitally on the one-bit stream. Consequently, if the STF is peaking significantly in some frequency bands, the MSA as seen from the modulator input will be reduced accordingly. In the cases where equalization is not possible, the feedback filter optimization must include a constraint on the maximum allowable STF base-band ripple. Such a constraint can decrease the attainable signal-to-noise ratio considerably. The use of an STF-constraint was also used in [29].

When $C(z)$ is significantly non-flat in the base-band, the base-band behavior of the NTF will also be affected. Hence, the optimization of the zeros given by $C(z)$ and the poles given by $D(z)$ can theoretically no longer be carried out separately due to this interaction. On the other hand, the increased dimensionality of the unified optimization problem can be too impractical.

Example 7.11 The 8th order modulator for 32 times oversampling of Example 7.10 was optimized with the usual constraints: $A_{min} \leq 2.4$, $S_{min} \leq 3.5$ and $F_{max} \leq 1.8$. The feedback filter poles were chosen according to [58]. The STF in a multiple feedback configuration is peaking 10.4 dB at the base-band edge. The base-band noise power was measured to -129.3 dB, i.e., better than for the NTF-prototype design in Example 7.10. The optimized modulator is very reliable and the MSA was estimated to 0.41 using a ramp with a rate of 10^{-9}/sample.

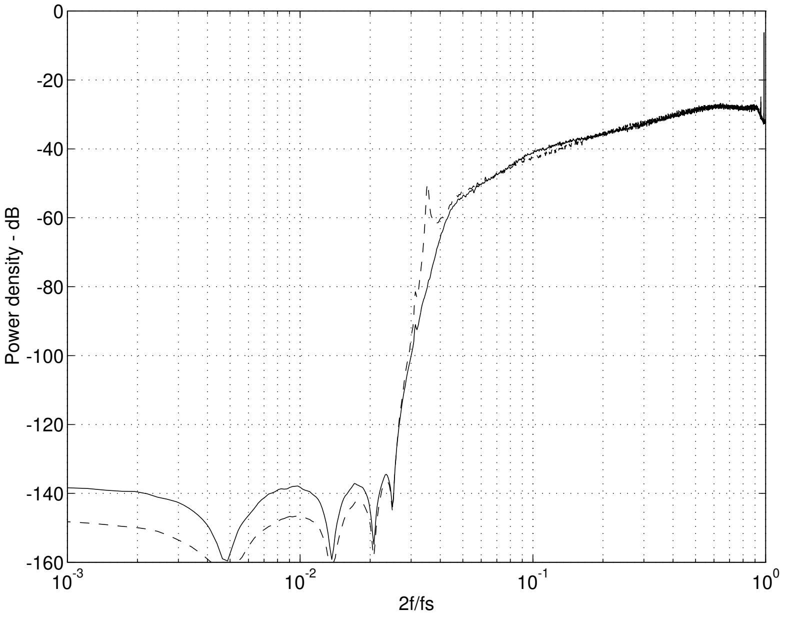

Figure 7.15: Estimated power spectra for the two optimized 8th order modulators in Example 7.11. The dotted line represents the modulator optimized without an STF magnitude constraint.

> The modulator was reoptimized using a ±0.25 dB constraint on the multiple feedback STF magnitude within the base-band. The resulting modulator yielded a σ_b^2 of -125.7 dB, i.e., the new STF constraint reduces the noise rejection. Using a ramp with a rate of 10^{-9}/sample, the MSA was estimated to 0.36.
>
> Fig. 7.15 shows the estimated power spectra of the two modulators with a constant input of 1/64. Note that the STF peaking can be observed on the base-band spectrum for the modulator optimized without an STF constraint. This modulator shows also a noise spectrum increasing faster above the base-band than for the modulator with the STF constraint. In a D/A converter, it is extremely difficult to suppress such out-of-band noise using analog filters of reasonable complexity. □

The previous example showed that the A_{min} and S_{min} constraints are fairly consistent even for eighth order modulators. However, the quadratic programming used by the CONSTR routine is very sensitive for high-dimensional systems, i.e., the optimization routine might be unstable for some initial guesses. It is strongly recommended to impose tight bounds on the parameters in order to confine the optimization process.

7.8 Optimization of Unusual Topologies

An extension of the multiple feedback modulator topology has been suggested in [43]. This topology includes an extra FIR filter in the feedback section, cf. Fig. 7.16. The hardware penalty for such an FIR filter is quite low, since the filter operates with a one-bit signal

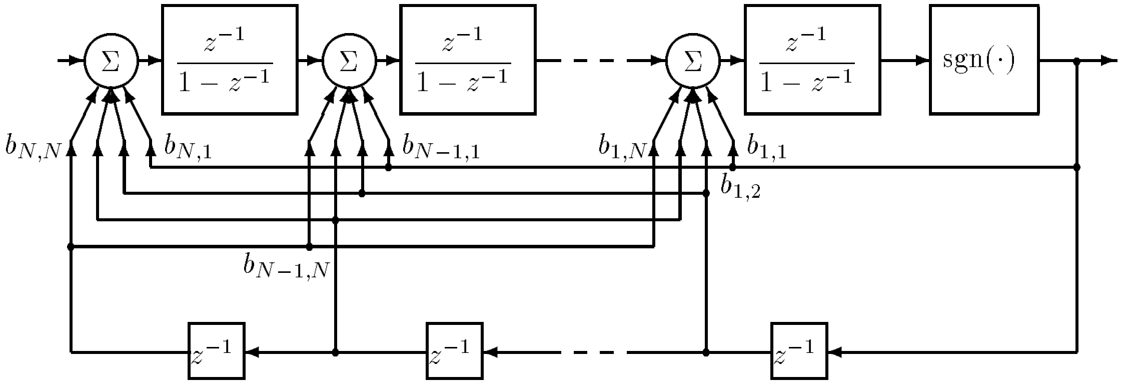

Figure 7.16: A proposed extension of the usual multiple feedback modulator (the figure is a modified version of a topology suggested in [43]). The topology allows an equivalent feedback filter with more zeros than poles.

as input. In a digital implementation, such FIR filter can be implemented using table look-up. In the analog domain, the structure on Fig. 7.16 can directly be used. The $b_{i,j}$-coefficients form a (N,N)-matrix which determine the $2N-1$ zeros of the feedback filter. This gives a redundancy which allows the use of simpler $b_{i,j}$ coefficients. A fourth order modulator designed with power-of-two coefficients was presented in [43].

The new topology allows an equivalent feedback filter with a numerator with a higher degree than the denominator, i.e., the number of design parameters increases. Such feedback filters can be split into a cascade of an FIR filter followed by a conventional IIR Nth order filter where N is the number of feedback filter poles. Feedback filters of this type can not be designed by the NTF-prototype method using conventional high-pass filters. Consequently, a direct design method like the proposed optimization framework must be used.

A number of experiments were carried out on a fourth order modulator, i.e., a modulator with four poles in the feedback filter. The modulator was optimized using 3, 4 and 5 feedback filter zeros. A number of 5 feedback zeros corresponds to an extra second order FIR filter. Only marginal improvements were obtained using these extra zeros: the base-band noise suppression was improved approx. 1 dB while the stability was preserved.

7.9 Summary

A framework for feedback filter optimization has been presented in this chapter. The method is based on a constrained optimization of a prediction of the base-band noise suppression. The base-band noise power prediction was based on the quasilinear modeling derived in Chapter 6. The constraints ensure that the modulator has a prescribed stable and reliable amplitude range. The modulator stability was analyzed in the BIBO sense using an equivalent noise-shaper topology. The scaling invariance of the feedback filter was utilized to improve the BIBO criterion. It was demonstrated that the BIBO-criterion can be based the minimum one-norm of the noise transfer function while the Gaussian criterion from Sec. 6.6is based on the minimum two-norm. A third stability measure was defined using the infinity norm of the signal transfer function.

Optimizations constrained by proper bounds on the three stability measures yielded good results. The constraints were consistent for a wide range of modulator orders, i.e., the same set of constraints yielded fairly consistent stability properties irrespectively of the modulator order.

Other constraints can easily be included. For instance, in some cases it is convenient to restrict the ripple of the base-band signal transfer function. Another possibility is to include the characteristic of a succeeding decimation filter into the optimization.

Chapter 8

Designing Tone Free Modulators

8.1 The Tone Problem

It is a well known problem that Σ-Δ modulators produce in-band as well as out-of-band tones in the presence of low amplitude constant input. In particular, there is often a very predominant tone near half the sample rate at the frequency $f = (1 - \mathrm{dc})f_s/2$ when the modulator is fed with constant input dc [2, 24, 32, 33]. This tone can be observed on all the spectra shown previously for modulators with constant input. The tone is actually frequency modulated by the input signal: sinusoidal input generates characteristic side-band tones (see the paper in Appendix C).

The persistence of this often very strong tone near half the sample rate is not easily explained. The classical first order modulator has the $\overline{10}$ limit cycle as the only steady-state solution for zero input [15]. This limit cycle gives indeed a strong tone at $f_s/2$. As a small dc-bias m_x is added to the input, the modulator will from time to time generate a code segment consisting of two identical codes. This happens because the error between modulator input and output accumulates in the integrator inside the loop. Consequently, the mean value of the modulator output must be equal to the dc-bias, i.e., $m_y = m_x$. The frequency of code repetitions must then be $m_x \cdot f_s$ and this also explains the tone at $(1 - m_x)f_s/2$. Higher order modulators seem to inherit this tone mechanism, though the output code becomes more complex. An Nth-order multiple feedback modulator can be perceived as a first order modulator embedded in a feedback loop containing $N - 1$ integrators. The extra integrators do only modify the low-frequency spectrum, i.e., more noise is removed from the base-band. The effect at higher frequencies is very little due to the low-pass filtering action of the extra integrators, and hence, the high-frequency tone caused by the innermost integrator is persisting [2].

The high-frequency tone is often accompanied by in-band tones at multiples of $f = \mathrm{dc}f_s$. This can be interpreted as a kind of intermodulation between the high-frequency tone and its image above half the sample rate. The high frequency tone is often very strong and can be quite close to full scale. This causes the random-like part of the quantization noise to be modulated, i.e., the total instantaneous output power is always unity and this must be shared by the random component and the tone. The result is that also the base-band noise power is modulated by the frequency $f = \mathrm{dc}f_s$, i.e., the random part of the quantization noise becomes non-stationary. The noise modulation itself cannot be seen on usual power spectrum estimates because time averaging is required. However it has been pointed out that periodic noise modulation can be perceived as a tone by the human ear [40]

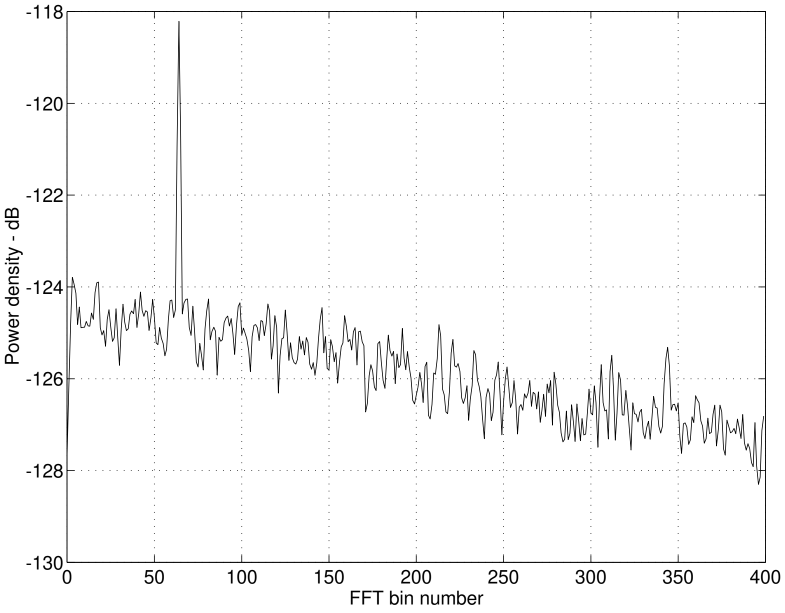

Figure 8.1: Estimated power spectrum of the decimated, high-pass filtered and squared modulator output based on 100.000 decimated samples simulated with a constant input of $1/(64 \cdot 128)$.

> **Example 8.1** The fourth order modulator of Example 7.8 was simulated with a constant input of $1/(64 \cdot 128)$ and the modulator output was decimated 64 times using a very selective filter. The resulting base-band noise signal contained a rather strong tone at the frequency $f = f_s/(64 \cdot 128)$ accompanied by harmonics. The decimated base-band signal was high-pass filtered in order to remove the dc-component, the tone and its harmonics. The filter used was an 8th order Chebychev II filter with Rs= 60 dB and stop-band edge at $0.1 \cdot f_s/2$. The filtered signal was subsequently squared and the power spectrum was estimated using the usual Welch method based on windowed 8k FFT's. The resulting spectrum is plotted in Fig. 8.1, and this spectrum shows a significant peak at bin number 64 which corresponds to the frequency $f = f_s/(64{\cdot}128)$. This peak is not caused by the tone itself due to the high-pass filtering. A possible interpretation is that the variance of the base-band signal is modulated periodically, i.e., the quantization noise is clearly non-stationary
> □

The simulations in Example 8.1 seem to confirm the periodic noise modulation is caused by the very strong high-frequency tone. There seems also to be a strong link between this tone and the base-band tones. Unfortunately, the out-of-band tones have not attracted very much attention in the literature so far.

Another observation from the time domain is that the out-of-band modulator output noise can be very 'impulsive' of nature, i.e., the signal power is concentrated in short bursts with high peak values [42].

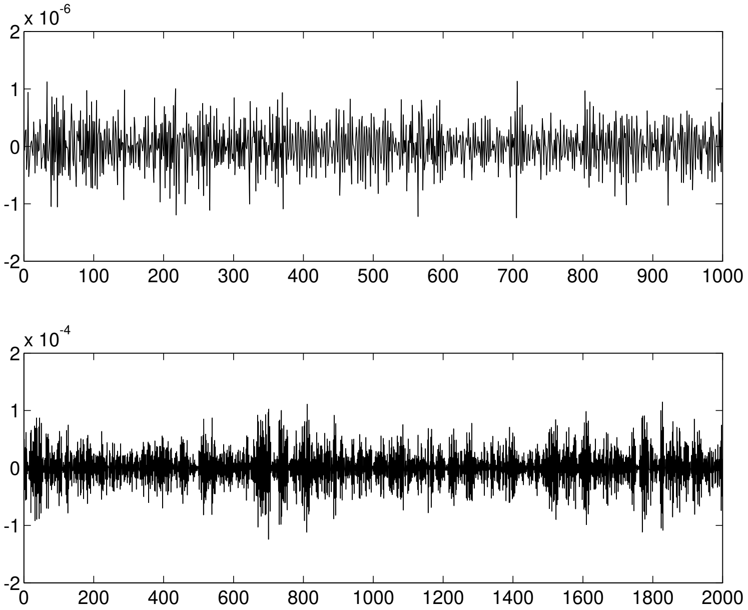

Figure 8.2: Decimated and high-pass filtered output signals of the fourth order modulator of Example 7.8 fed with a constant input of $1/(64 \cdot 512)$. The upper graph is decimated 64 times and the lower is decimated 32 times.

> **Example 8.2** Fig. 8.2 shows two signal sequences for the fourth order modulator from Example 7.8 fed with a constant input of $1/(64 \cdot 512)$. The upper graph shows the 64 times decimated and subsequently high-pass filtered modulator output. The plotted sequence corresponds to approx. two periods of the noise modulation which is only barely visible. The lower graph shows a similar sequence obtained by only 32 times decimation and subsequent high-pass filtering. Note the difference in magnitude scale due to the out-of-band noise. The length of the sequence matches the upper graph. The energy of the 32 times decimated sequence is, contrary to the 64 times decimated sequence, clearly concentrated in short bursts. This kind of non-stationary behavior is not revealed by usual spectral analysis □

The presence of tones and noise modulation are both unwanted effects of a modulation scheme. Even if the in-band tones are suppressed, the high frequency tone will still cause noise modulation. Furthermore, the presence of analog circuit non-linearities can still produce in-band tones and harmonic distortion due to intermodulation effects [2]. Consequently, the out-of-band tones must be taken into consideration when data converters are designed and evaluated.

8.2 Methods for Tone Suppression

The tone problem as far as in-band tones are concerned has received a lot of attention and several methods have been proposed for suppression of the tones [26, 37, 40, 57]. It

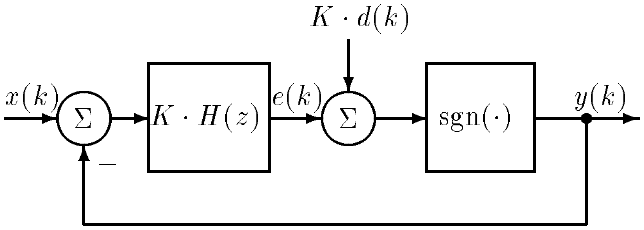

Figure 8.3: Σ-Δ modulator with scaled feedback filter $H(z)$ and dither source $d(k)$.

is generally believed that higher order modulators are less tonal [17]. This is only true when the in-band tone components are taken into consideration: the ratio between the power of then in-band tones and the in-band noise power gets typically lower for higher order modulation. However, the strength of the high-frequency tone changes only very little, i.e., the amplitude is generally in the range from -10 dBFs to -5 dBFs [1]. In many cases, the tone is stronger than the maximum amplitude of the sinusoidal modulator input. Consequently, high order modulators will still suffer from noise modulation and in-band tones caused by intermodulation due to circuit non-linearities.

A common approach is to add dither noise to the modulator loop [12, 40, 41, 42], e.g., accomplished by adding noise to the modulator input. However, this requires that the noise is heavily high-pass filtered in order not to sacrifice the signal-to-noise ratio. A more convenient approach is to add the dither to the quantizer input as shown in Fig. 8.3. This will automatically shape the spectral contribution of the dither according to the noise transfer function. The two approaches are theoretically equivalent when a suitable dither pre-filtering is applied in one of the cases.

The addition of a dither noise source has the effect that possible tones are 'dissolved', i.e., the signal power of the tones is partly transformed into random noise. At some dither level, the tones will be 'burried' in the random noise floor of a power spectrum estimate.

The effect of a dither source is 'discrete', i.e., the output code $y(k)$ will now and then change sign due to the dither signal. Theoretically, there is a large class of neutral dither signals which will not change the output code, e.g., as long as $d(k)$ has the same sign as $e(k)$, the operation of the modulator will be unchanged even for very large amplitude dither.

The presence of a dither source prior to the quantizer input will generally reduce the correlation between $e(k)$ and $y(k)$ and, hence, the quantizer gain K given by the quasilinear model will be reduced. This reduces the noise amplification factor A according to the A(K) curve of the feedback filter, i.e., the quantization noise power σ_q^2 of the quantizer is increased. Consequently, the reduced loop gain and the increased σ_q^2 will increase the base-band noise power σ_b^2. As the quantizer gain K is reduced, the equilibrium will move towards the A_{min}-point, i.e., the stability is deteriorated; hence, the stable amplitude range is reduced. The same conclusion can be drawn from the BIBO considerations in Sec. 7.3:

---

[1] dB relative to a full-scale sinusoid

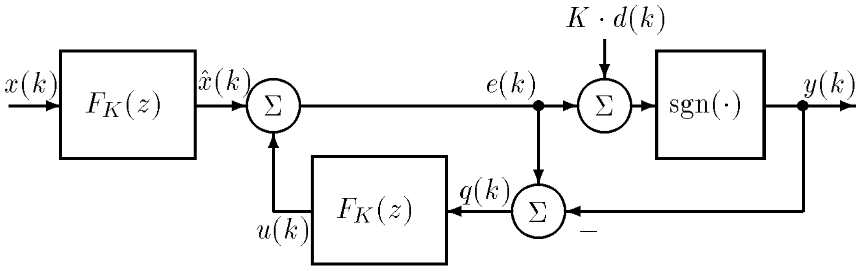

Figure 8.4: Equivalent noise-shaper topology of a Σ-Δ modulator with scaled feedback filter and dither source (see Fig. 8.3. The filter $F_K(z)$ is given by Eq. (7.7).

the magnitude of the quantization noise $q(k)$ can be as large as $1 + \|d(k)\|_\infty$ where $d(k)$ is the dither sequence (see the equivalent noise-shaper in Fig. 8.4). This maximum value occurs, e.g., when $e(k) = \|d(k)\|_\infty$ and $d(k) = -\|d(k)\|_\infty$. A BIBO stability criterion can now be derived:

$$\|f(k)\|_1(1 + \|d(k)\|_\infty) \leq 2 - \|\hat{x}(k)\|_\infty \tag{8.1}$$

where $\|f(k)\|$ is the one-norm of the impulse response $f(k)$ associated with $F(z)$.

A similar result was derived in [41].

The feedback filter scaling invariance applies only when the dither signal is scaled accordingly. Hence, a specification of the dither span without a specification of the scaling of the feedback filter is imprecise. Utilizing a feedback filter and dither scaling by a factor K, the BIBO criterion becomes:

$$(S(K) - 1)(1 + K\|d(k)\|_\infty) \leq 2 - \|\hat{x}(k)\|_\infty \tag{8.2}$$

where $S(K) = \|f_K(k)\|_1 + 1$ is the one-norm of the impulse response of the noise transfer function as a function of the quantizer gain K, cf. Eq. (7.9).

The best (i.e., the most optimistic) criterion is obtained for the K-value that minimizes the left hand side of Eq. (8.2). Eq. (8.2) shows that as the dither magnitude $\|d(k)\|_\infty$ is increased, the stable amplitude range is reduced.

Consequently, both the quailinear analysis and the BIBO criterion show that modulator feedback filters designed for use with a dither noise source must be designed very conservatively in order to preserve a prescribed stable amplitude range.

As an alternative to dither, it has been suggested that the use of chaotic modulators can suppress the tones [26, 37, 41, 57]. The output of a chaotic system is generally non-periodic and the spectrum must be continuous. However, this property does not exclude that tones might exist combined with non-periodic noise like components. As stated in Chapter 3, chaotic modulators are obtained when one or more of the poles of the feedback filter are located outside the unit circle. This means that the noise transfer function is mixed phase, i.e., a minimum phase transfer function with the same amplitude characteristic can be obtained by reflecting the zeros inside the unit circle to their reciprocal locations. The zero reflected minimum phase noise transfer function has a lower two-norm or noise amplification factor A than the mixed phase characteristic due to the ntf(0) = 1 scaling requirement. The reason is that any mixed phase filter can be composed of a minimum phase part and an all-pass part [44]. Due to the necessary scaling, the amplitude gain of the all-pass term is greater than unity. The zero reflected minimum phase filter is obtained when the all-pass term is removed. Consequently, for a given base-band noise suppression, a chaotic modulator will always have a larger A_{min}-value than for a non-chaotic modulator. According to the Gaussian criterion, this means that the stable amplitude range is reduced. A similar argument holds for the BIBO-criterion since the one-norm of the mixed phase noise transfer function is higher than for the minimum phase. Consequently, both the use of dither and chaos will deteriorate the loop stability.

The paper in Appendix A compares dithering and the use of chaos with respect to the suppression of the high-frequency tones. The comparisons are made for sixth order modulators intended for 64 times oversampled CD-standard digital audio reproduction. The conclusion is that both dithering and chaos can attenuate the high-frequency tones more than 30 dB. The price paid is a significant (20-30 dB) reduction of the attainable signal-to-noise ratio. It was shown that a feasible way to design tone free chaotic modulators is to extend a conventional NTF-prototype with a first-order all-pass term having a (real valued) zero α outside the unit circle. The resulting feedback filter will then have a pole at α. The best result is obtained when α is negative, i.e., closer to the tone frequencies near $z = -1$ which corresponds to half the sample rate. It was concluded that a tonefree and dithered sixth order modulator has a performance similar to a tonefree and chaotic seventh order modulator with a sixth order NTF magnitude characteristic. This shows that the use of chaos and dithering have very similar effects on the tones and the stability.

A totally different approach for tone suppression is used in [33]: a fine-scale 'trim bit' with, e.g., 8-bit lower significance is added to the coarse one-bit quantizer output and the result is a rather unusual 4-level quantizer with a large number of thresholds where the output code changes. It is demonstrated that such coarse/fine quantizer gives better low level resolution without in-band tones. However, the effect on the strong out-of-band tones is not clear.

8.3 Measures of Information Loss

Both the use of dithering and chaos introduce an element of uncertainty into the modulation. For a dithered modulator, the next output code will to some extend be uncertain. If the state of a chaotic system is known with some precision at time-step k, then the state of the system at time-step $k+1$ is known with a reduced precision due to the sensitivity to initial conditions. This property represents a loss of information for each time-step.

The product of the feedback filter pole moduli was introduced as the system map expansion factor Λ, cf. Eq. (4.7). The rate of information loss can be expressed in bits per time-step as the base two logarithm of Λ. As Λ approaches two, the loss of information becomes close to one bit per time-step. It was previously shown that a first order modulator becomes unstable when Λ is greater than two (see Example 3.1). Higher order systems become typically unstable for lower Λ values. Consequently, an information loss rate of one-bit per time-step is the upper limit for a one-bit modulator. The noise shaping theorem (Theorem 6.1) is in fact inspired by information capacity considerations. When the equality of Eq. (6.4) holds, the information capacity of 'the noise shaped transmission channel' is preserved [20]. This is only the case for minimum phase noise transfer functions, i.e., for non-chaotic modulators. Consequently, the the use of chaos reduces the information capacity.

Based on listening experience, it has been claimed in [31] that multiple feedback filter poles at $z = 1$ gives less correlated quantization noise than for complex unit circle poles. A likely explanation is that the number of periodic points increases as $n^{(N-1)}$ where n is the period length and N is the multiplicity of the $z = 1$ poles (i.e., the order), whereas the number of periodic points is asymptotically constant for the case of complex unit circle poles.

It is also possible to quantify the loss of information for a dithered modulator. Let $X = \text{sgn}(e(k))$ and $Y = \text{sgn}(e(k) + d(k))$ be stochastic variables. The conditional entropy $\text{H}(Y|X)$ (in bits) expresses the uncertainty about the current modulator output code Y given the output code X without dither. The undithered system yields $\text{H}(Y|X) = 0$, since there is no uncertainty about Y once X is known. As more and more dither is added, $\text{H}(Y|X)$ increases. The upper limit is of course one-bit in the extreme case where Y is statistically independent of X and the two outcomes of Y are equally likely.

It is now possible to compare the amounts of dither and chaos. The dithered sixth order example from the paper in Appendix A operates for zero input with an uncertainty $\text{H}(Y|X)$ of approx. 0.69 bit/time-step. The all-pass derived sixth order modulator has a 'chaotic' pole at $\alpha = -1.25$ corresponding to an information loss rate of 0.32 bits/time-step. Consequently, the chaotic modulator can suppress the high-frequency tone at a lower rate of information loss than for the dithered example. If the all-pass zero is flipped to $\alpha = 1.25$, the high-frequency tone is only reduced slightly while the pole modulo is preserved.

8.4 Limit Cycle Analysis of Dithered Modulators

The limit cycle analysis framework presented for chaotic modulators (Ch. 4) can be extended to include dithered modulators. The analysis uses the usual modulator topology of Fig. 8.3. To test if a given periodic code sequence exists as a limit cycle, the loop is cut open at the input of the quantizer and the periodic steady-state loop filter output $e(k)$ is found when the period code sequence $y(k)$ is fed back. If the quantizer will reproduce the output code sequence from $e(k)$, then the code sequence exists as a limit cycle. However, it should be emphasized that a modulator dithered by an independent, stochastic dither signal is not a deterministic autonomous system. Consequently, the concept of limit cycles does not exist for such dithered modulators. Only when the dither signal is a known periodic sequence, the limit cycle concept is meaningful.

The addition of a periodic dither signal $d(k)$ to $e(k)$ changes the composition of possible limit cycles, i.e., the dither can change the quantizer output code at suitable instants and other periodic sequences can be reproduced. This kind of limit cycle analysis provides information about the possible code sequences of a dithered modulator. A stochastic dither generator may theoretically produce a large class of periodic signals. Hence, the objective of the limit cycle analysis is to find every theoretically possible limit cycle for a given dither generator. In practise, the limit cycles are not observable for real simulations with random dither, since each of the limit cycles requires a particular class of periodic dither signal added in order to be sustained. This is exactly the same situation as for chaotic modulators: there is a large number of unstable limit cycles and, hence, it is very unlikely that a particular cycle is sustained for a longer period of time due the sensitivity to initial conditions. However, the knowledge of the composition of possible limit cycles describes the possible code sequences, i.e., the modulator will always be arbitrarily close to a limit cycle.

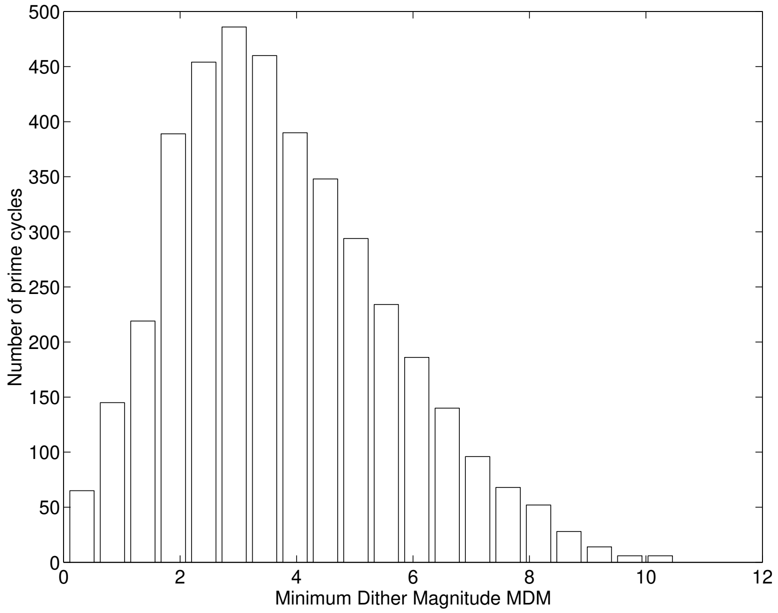

Figure 8.5: An Minimum dither magnitude (MDM) histogram for the set of period 16 prime cycles of the modulator with feedback filter cf. Eq. (8.4) with zero input. A total of 4080 prime cycles of length 16 exist cf. Table 4.1.

It was stated in Sec. 4.3 that the number of periodic points of a chaotic modulator must grow exponentially with the period length in order to maintain stability. To be more specific, the number of periodic points must grow like Λ^n where n is the period length and Λ is the expansion factor (i.e., the product of pole moduli greater than unity).

The limit cycle analysis can thus describe the similarity between the effects of dithering and the use of chaos. In both cases, the number of limit cycles increases and this tends to randomize the output code sequence.

It must be emphasized that the limit cycle analysis of a dithered modulator does not provide information about the stability, i.e., a modulator will always be less stable when random and independent dither is applied.

A practical way to analyze limit cycles for a dithered modulator is to find the steady state $e(k)$ signals for each of the prime cycles up to a given length. Subsequently, the minimum dither magnitude (MDM) necessary to reproduce the corresponding prime cycle is computed for each $e(k)$ sequence:

$$\mathrm{MDM}_i = -\min(e_i(k) \cdot y_i(k)) \tag{8.3}$$

where $y_i(k)$ and $e_i(k)$ are the code sequence and steady-state loop filter output corresponding to prime cycle i.

A negative MDM_i means that the prime cycle i exists without any dither, i.e., $y(k) = \mathrm{sgn}(e(k))$. If the dither signal $d(k)$ is a bounded random i.i.d. sequence then every prime cycle i satisfying $\mathrm{MDM}_i < \|d(k)\|_\infty$ will theoretically exist, i.e., there is a (small) possibility that the random dither signal will support the limit cycle for some time given the right initial conditions.

Example 8.3 This example investigates the second-order modulator feedback filter:

$$H(z) = \frac{z^{-1} - 0.5z^{-2}}{1 - 1.95z^{-1} + z^{-2}} \tag{8.4}$$

The feedback filter has complex unit circle poles, i.e., the number of periodic points is expected to be asymptotically constant without dither applied. The following analysis is accomplished for zero input.

Fig. 8.5 shows a histogram of the MDM-values of the set of period 16 prime cycles. There are no negative MDM-values, i.e., no period 16 prime cycles exist without dither. As the dither span is increased, the number of existing prime cycles increases. For a maximum dither magnitude of 10.5, every prime cycle of length 16 exists (a total of 4080, cf. Table 4.1).

Fig. 8.6 shows the number of periodic points versus the period length for different maximum dither magnitudes. The plot shows only values for even period length since there are generally fewer limit cycles with uneven length (these cycles have a non-zero mean value). For no dither, the number of period points levels out to a fairly constant number. For higher dither levels, the curves follows the same initial steep slope and at some point, the slope levels out. Unfortunately, the developed routines in MATLAB were not able to handle prime cycles of length greater than 16. However, from the plots on Fig. 8.6 it seems likely that the curves will continue to grow exponentially with exponents depending on the dither level, i.e., the same asymptotic behavior as for chaotic modulators □

The limit cycle analysis for a dithered modulators confirms the similarity between dithering an the use of chaos: In both cases, the number of periodic points increases (exponentially) with the period length.

A dithered non-chaotic modulator with poles inside the unit circle will typically display a dither threshold effect, i.e., for low dither levels, the output will be strictly periodic and at a certain dither level the output becomes more random. This phenomenon is easily explained: modulators with poles inside the unit circle have attracting limit cycles and if the dither level is insufficient to eventually change an output code, the limit cycle will trap the modulator. The limit cycle which can tolerate the highest dither levels will then dominate and asymptotically attract the system. This limit cycle has the lowest (i.e. most negative) MDM-value. Typically, the classical $\overline{10}$ limit cycle requires the most dither to knock out. This gives also a credible explanation on the persistence of the high-frequency tone.

Example 8.4 A modulator with a feedback filter given by Eq. (8.4) was modified such that the unit circle poles were scaled by a factor of 0.95. the steady state feedback filter output for the $\overline{10}$ limit cycle is $e(k) = (0.3995, -0.3995...)$, i.e., it requires a dither signal with a peak magnitude of at least 0.3995 to dissolve this attracting limit cycle. Simulations with uniformly distributed dither confirms this threshold effect □

For a non-chaotic modulator, it is meaningless to apply less dither than prescribed by the dither threshold.

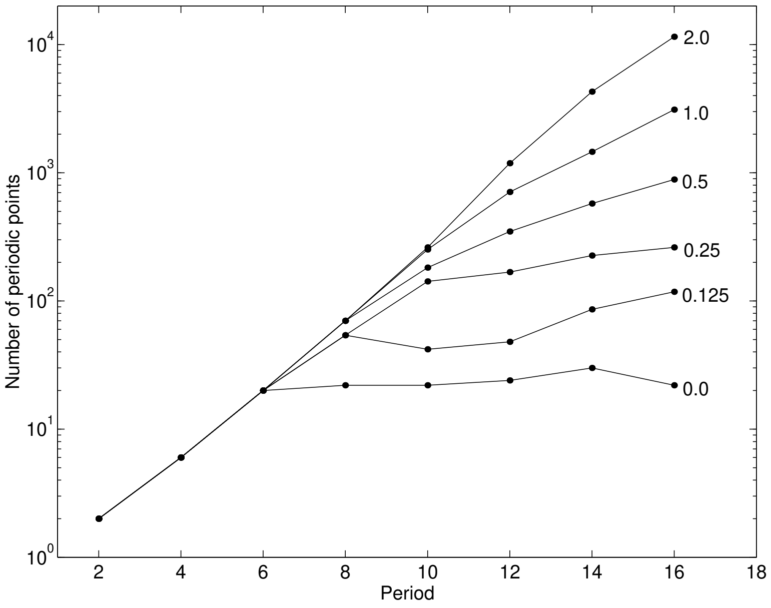

Figure 8.6: Number of periodic points versus period length for the modulator cf. Eq. (8.4) with zero input. The parameter on each plot is the peak dither magnitude $\|d(k)\|_\infty$.

8.5 Summary

This chapter reviewed different approaches for suppression of the strong tones which can be observed in the output spectra of modulators fed with small amplitude constant input. The paper in appendix A presented some practical design examples which demonstrated that both dithering and the use of chaos can suppress such tones. The price paid in both cases is a significantly reduced attainable signal-to-noise ratio due to a detrimental effect on the modulator stability.

The use of chaos and dithering was compared from a theoretical point of view. Both methods induce a degree of uncertainty into the modulation which can be measured as a rate of information loss. Another way of comparing was to investigate the number of periodic points versus the period length. Both dithering and chaos increases the number of possible limit cycles and this randomizes the output sequence.

The question of whether to use dithering or chaos is quite complex. In digital systems, pseudo random number generators can easily be made with a controllable probability distribution. On the other hand, it also takes only relatively little extra hardware to increase the order of the modulator by one as required by the first order all-pass extension.

In analog modulators, the design of dither noise generators with prescribed probability density functions can be difficult: noise generators based on thermal noise has Gaussian probability density and the variance is strongly temperature dependent. Furthermore, the use of thermal noise requires usually high amplification and this may lead to unintentional feedback and deteriorated power supply rejection. A better approach is to use a digital noise generator and a suitable D/A-converter. However, this approach is rather hardware consuming. Consequently, for A/D-converters, the use of chaos can be a very attractive.

Chapter 9

Σ-Δ Modulation using Vector Quantization

9.1 Introduction

In this chapter, a novel class of modulators is introduced. In the traditional Σ-Δ modulator, the output code is determined as the sign of the feedback filter output, i.e., the decision of which output code to generate is based on a one-dimensional projection (many-to-one mapping) of the general N-dimensional state-space of the feedback filter. From a theoretical point of view, such a one-dimensional projection does only convey very little information about the filter state. A more general way to make the output code decision is to use the whole state-space, i.e., the scalar one-bit quantizer is replaced by a one-bit *vector quantizer* (VQ) [19] which maps the entire state-space into the binary output alphabet. The class of modulators using vector quantization is naturally a superset of the traditional class of modulators with scalar quantization.

Each of the modulators using vector quantization can be transformed into an equivalent traditional modulator with a scalar quantizer operating on a scalar filter output $e(k)$ combined with a vector quantizer which can change the decision of the scalar quantizer. The negated transfer function from the modulator output to the scalar signal $e(k)$ is still defined as the feedback filter $H(z)$. Such transformation can naturally be accomplished in many ways. Fig. 9.1 shows the topology of the the proposed modulator with combined scalar and vector quantization.

Conceptually, the topology of Fig. 9.1 can also be perceived as a traditional modulator

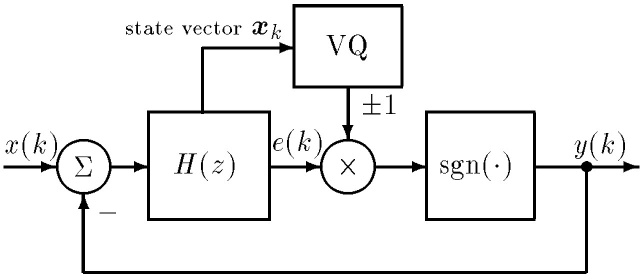

Figure 9.1: Σ-Δ modulator with vector quantization

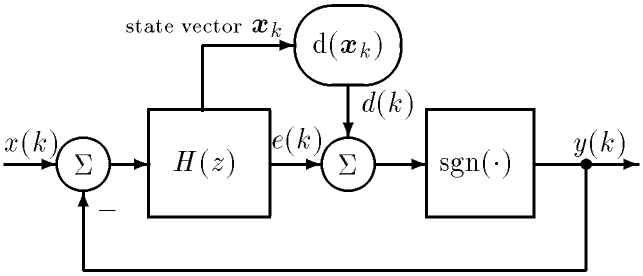
state vector $\boldsymbol{x}_k$
$\mathrm{d}(\boldsymbol{x}_k)$
$d(k)$
$x(k)$
Σ
−
$H(z)$
$e(k)$
Σ
sgn(·)
$y(k)$

Figure 9.2: Σ-Δ modulator with vector quantization implemented as a deterministic dither signal

with a 'dither signal' added which is a deterministic projection of the filter state. Such topology is shown on Fig. 9.2. The dither signal produced by the generally nonlinear mapping $\mathrm{d}(\boldsymbol{x}_k)$ causes the output code of the scalar quantizer to change at some instants. If the dither signal is generated as a linear projection of the filter state, the system is equivalent to a traditional modulator with a modified $H(z)$ transfer function. To be more specific, a linear state-space projection can only change the zeros of $H(z)$ since the poles are always the eigenvalues of the transition matrix of the filter. Hence, the only interesting situation is when the deterministic dither signal is a nonlinear projection of the filter state. Since the usual scalar filter output $e(k)$ and the artificial dither signal is added before the sum is quantized by the scalar one-bit quantizer, the proposed new class of modulators could also be characterized as a modulator having a nonlinear feedback filter output. The nonlinear mapping could be implemented using a suitable neural network. It has previously been suggested to use a nonlinear feedback filter [65]; however, the proposed topology is not equivalent to a modulator with a vector quantizer

The question is obviously: is it possible to improve the performance of a modulator by the introduction of a vector quantizer? It seems very likely that it is possible to improve the stability. The reason is that a vector quantizer can allow more limit cycles to exist and this can be done in a selective way, i.e., the VQ can be designed to generate more limit cycles belonging to the bounded limit set. Recall from Ch. 2 and Ch. 4 that a chaotic limit set is made up of a large number of limit cycles and that the stability of the limit sets depends on the 'density' of limit cycles. The VQ might thus keep the modulator operating in a bounded region while the usual noise shaping effect is preserved, i.e., the base-band noise rejection is primarily controlled by the filter poles which remain the same. Such effects can increase the overall modulator performance.

The possibility of designing the composition of limit cycles might also be used to change and hopefully reduce the amplitude of the possible tones (see Ch. 8 for a discussion on tones). An unwanted limit cycle can be removed by a slight modification of the VQ. Since the modulator state is always arbitrarily close a limit cycle, the spectral properties of the modulator output is also influenced when changing the set of limit cycles. The possibility of using VQ for tone suppression is an open field for future research.

9.2 Specification of the VQ

It is impossible to investigate the entire class of modulators using vector quantization. Furthermore, there is currently no way to accurately model these systems and this makes systematic optimization very difficult. However, a systematic method for the design of the vector quantizer is proposed in [52] reprinted in Appendix B. The method starts from a usual modulator with loop filter $H(z)$ given in a state-space description. Subsequently, a so-called *legal set* is specified. The purpose of the vector quantizer cf. Fig. 9.1 is to ensure that the modulator never leaves the legal set. This is accomplished by identification of a number of regions F_n in the state space where the VQ should change the code generated by the scalar quantizer. These regions have the property that states in F_n leaves the legal set after precisely n time steps, if the usual quantizer is used. In addition, if a time-step is taken with an inverted output code followed by $n-1$ time steps with normal quantization, the modulator stays within the legal set. A direct implication of these properties is that the F_n regions are mutually disjoint. A VQ which changes the output when the modulator state is in one of the F_n regions for n below some limit, can help to confine the modulator operation (i.e., the limit set) to be inside the specified legal set. If the unmodified modulator was unstable, then the VQ can in some cases stabilize the system.

The proposed method requires some tweaking, i.e., the choice of the legal set is very crucial. A simple way to define the legal set is to specify an upper and a lower bound for the scalar filter output $e(k)$. This choice of legal set simplifies also the structure of the F_n sets, i.e., these sets are *polytopes* (sets delimited by hyper-planes in the state-space [19]). This enables the vector quantizer to be implemented using a number of scalar quantizers operating on auxiliary scalar filter outputs.

9.3 The Back-Step Algorithm

For a given legal set, the method presented in Appendix B defines a vector quantizer and a numerical test is presented which can tell if the modulator output should be inverted. However, this test has a high numerical complexity. In some cases, it is more efficient to use the so-called *back-step* algorithm: the modulator is simulated in forward time until the system leaves the legal set. At such instants, the algorithm steps back one step at the time and tries to flip the output code. Each time a code is flipped, the system is simulated in forward time until the time instant where the first legal set violation was encountered. The output codes are flipped successively back in time until the system stays within the legal set. A convenient way to arrange the numerical data is to store the signal sequences in circular buffers.

1. Simulate in forward time until the legal set is exceeded and let $\boldsymbol{n}_c$ be the time-step where this situation occurred.
2. **backstep** $= 0$
3. **backstep** $=$ **backstep** $+ 1$, if **backstep** $> n_{max}$ then goto 1.
4. Flip the output code at time time $\boldsymbol{n}_c - \textbf{backstep}$ and simulate to time $\boldsymbol{n}_c$. If the system leaves the legal set then goto 3.
5. Goto 1

The back-step algorithm is faster when the vector quantizer only intervenes rarely, i.e., the algorithm stops only when leaving the legal set.

Example 9.1 An eighth-order modulator was designed using a Chebychev II prototype with the parameters Rs= 108 dB and fb= 1/32·20/24. The resulting feedback filter has a A_{min} of 2.63, and $S_{min} = 5.05$, i.e., the modulator is expected to be fairly unstable and unreliable (see Ch. 7).

The back-step algorithm was used on the eight-order feedback filter and a simulation was performed with a constant input of 1/256. An upper magnitude limit of $e(k)$ was set to 1.065. This resulted in a high frequency of output code changes. The output spectrum is seen on Fig. 9.3. The high vector quantization activity gives a significant suppression of the high frequency tone: the amplitude is as low as −17.7 dB. The base-band noise power is −131.1 dB. This shows that a reduction of the tones not necessarily sacrifices the in-band noise rejection.

The back-step algorithm was enable to stabilize the modulator for constant inputs up to 0.4, i.e., the vector quantization scheme can indeed extend the stable amplitude range considerably. In order to be stable, the $e(k)$ bounds were changed to -1.1 and 1.65. The algorithm had to go more steps back in time in order to limit the magnitude of $e(k)$. The spectrum showed again that the amplitude of high frequency tone was below −20 dB □

9.4 summary

A novel class of modulators employing one-bit vector quantization was introduced. This enables the output codes to be generated on the basis of the entire state-space information instead of on just a linear projection. Conceptually, such a modulator is equivalent to a modulator with a usual scalar quantizer which from time to time is overruled by a vector quantizer. Alternatively, this effect can also be achieved by a usual modulator with an added dither signal which is a deterministic (and nonlinear) function of the filter state. The operation of the feedback filter is unchanged, i.e., the filter has still the difference between the modulator input and output as input.

The introduction of a vector quantizer can change the stability and limit cycle composition of a modulator. A design method which aims at enhancing the stability was presented. A simulation example showed that the stability of an eighth-order modulator was enhanced significantly while the high frequency tone was reduced.

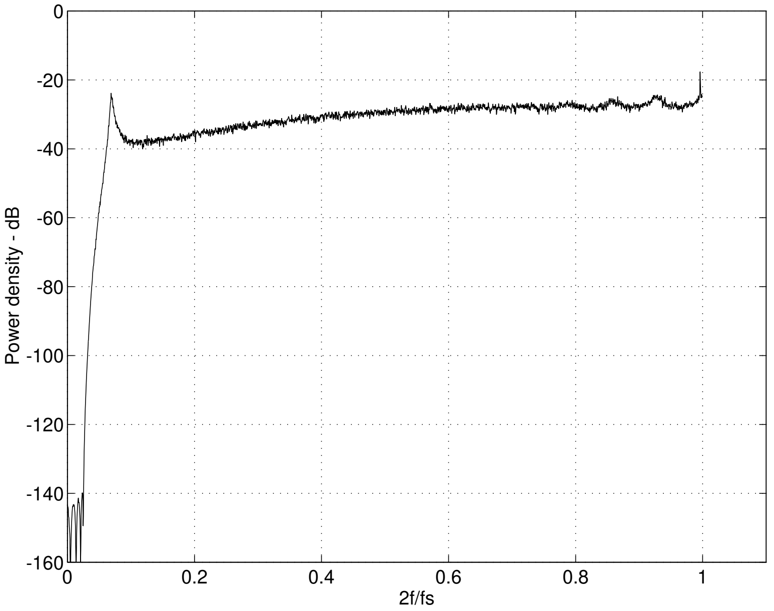

Figure 9.3: Simulated power spectrum estimate (500.000 samples) for an eighth-order modulator obtained from a Chebychev II prototype with the parameters **Rs**$=108\,\mathrm{dB}$ and **fb**$=1/32\cdot 20/24$. The modulator was simulated using the back-step algorithm with a $|e(k)|\leq 1.065$ bound and a constant input of $1/256$.

Chapter 10

Conclusions

The present work has mainly focused on the stability of Σ-Δ modulators. The stability issue is the most delicate problem when design of high-order modulators is addressed. The analysis from the nonlinear dynamics point of view, as presented in part I, explained the onset of instability in a chaotic modulator as a boundary crisis emerging when the limit set (attractor) of the system collides with the boundary of its basin of attraction. The degree of instability can be quantified using the escape rate.

Even for second-order systems, the basins of attraction revealed extreme diversity and complexity. This fact tells that a simple and general stability criterion is very difficult to find: Stability can be proven by showing the existence of a bounded trapping region which naturally must be inside the basin of attraction. In the simple first-order case, such trapping region is an interval. However, as the dimension increases, more and more complex geometric analysis has to be applied. If a given stability criterion proves the existence of a simple class (e.g., hyper cubes) of trapping regions, then the criterion will be very conservative, especially for high-order systems. This is in fact the case for the BIBO criterion based on the NTF one-norm, as presented in Sec. 7.3. Less conservative stability criteria are of higher complexity (i.e., proving more complicated sets as trapping regions) — in the limit of complexity, it is just as easy to just simulate the system to find the answer.

The Gaussian stability criterion as presented in Ch. 6 takes advantage of the regularity hidden in the complexity of high-order systems. The Gaussian criterion is based on considerations for the power of the signals within the modulator loop, i.e., the criterion is based on two-norms. Unfortunately, this criterion is only approximate, but on the other hand, it is generally fairly precise for reasonably 'well behaved' systems.

Empirical stability studies showed that some modulators have very unreliable behavior giving a very slow transition to the unstable regime. Unreliable modulators may also have discrete constant input values which causes instability. The overwhelming problem concerned with unreliable modulators is the existence of regions or discrete values of constant input where the escape rate can be extremely small. This means that simulations cannot be trusted since these modulators are seemingly stable for most inputs.

Ch. 7 combined the approximate Gaussian stability criterion with the rigorous one-norm based BIBO criterion and an infinity-norm based stability measure. The result was a set of constraints which consistently lead to very reliable modulators in connection with an optimization routine. The resulting modulators had a very sharp transition to instability and this means that the maximum stable amplitude limit found by simulations can be trusted, i.e., the modulators are expected to be very stable if the input is magnitude limited slightly below the found amplitude limit. The results were used to implement a highly stable eighth-order Σ-Δ DAC.

The topic of stability was also investigated from a more abstract mathematical point of view, i.e., the symbolic dynamics analysis presented in Ch. 4. The conclusion is that for stable system, the growth in the number of periodic points must outbalance the volume expansion of the system map of the modulator. The results are not suited for practical design, and only in a few extreme cases, it was possible to derive an analytical expression for the escape rate. However, the context of symbolic dynamics gives an enlightening new view on the stability issue.

The symbolic dynamics was also used to show the similarities between dithering and the use of chaos. In both cases, the number of periodic points increases with the period length.

In Ch. 9 a new class of modulators using vector quantization was presented. It was shown that a suitably designed vector quantizer can improve stability considerably. The symbolic dynamics can explain this: The vector quantizer might change the composition of limit cycles, and this might increase the growth of periodic points.

The existence of spurious tones in the presence of small dc input bias is another problem concerned with Σ-Δ modulators. The tone problem was treated empirically, i.e., no theoretical analysis was presented. The use of chaos and dithering was compared and the result was that both options are nearly equally efficient, but there is a severe SNR penalty. In Ch. 9 it was shown that heavy use of vector quantization might both improve the stability and reduce the amplitude of the tones without any SNR penalty. This fact should obviously be subject to future research.

It is still an open question why spurious tones might persist even high amounts of dither or chaos. Again, more theoretical analysis is needed.

Appendix A

Appendix for Chapter 8

This appendix is connected to chapter 8 and is a preliminary preprint of the paper: L. Risbo, “On the design of tone free Σ-Δ modulators”, to appear in *IEEE Transactions on Circuit and Systems — II.* The references are included in the bibliography section on page 173.

On the Design of Tone Free Σ-Δ Modulators

Lars Risbo

Abstract

Traditional one-bit sigma-delta modulators used for A/D and D/A conversion produce very predominant tones near half the sample rate which might intermodulate in the analog converter section and cause in-band tones. This paper demonstrates how the use of chaos can substitute dither as a means for extinguishing these tones. Especially, modulator feedback filters derived from noise transfer functions having an all-pass term seem very promising.

A.1 Introduction

Sigma-delta modulators are extensively used in the design of oversampling audio A/D and D/A data converters [1, 48].

One of the drawbacks of conventional sigma-delta modulators is the possible existence of tones in the output spectrum [40, 41, 57, 37]. A very common property is that when the modulator is fed with a dc-input, λ, the modulator spectrum contains a very strong out-of-band tone at the frequency $(1-\lambda)f_s/2$, where f_s is the sample rate [48, 40, 57]. Simulations have indicated that the tone close to $f_s/2$ causes the in-band noise power to vary periodically with the frequency λf_s (unpublished). Such periodic noise modulation is not observable on usual power spectra estimates but can be seen in the time domain and might be perceived as a tone by the human ear [40]. The frequency modulation of the tone near $f_s/2$ seems to be very general, i.e., sinusoidal input with frequency f produces characteristic side-band tones at the frequencies $f_s/2 - kf$ for integer k [48]. In addition, the high frequency tone and the side-bands can produce in-band intermodulation components produced by analog circuit nonlinearities [2, 24].

The conclusion from the observations mentioned above seems to be that precautions against the out-of-band tones must be taken if sigma-delta converters free from audible tonality are to be designed.

Enhanced suppression of both in-band noise and in-band tones can generally be achieved by increasing the order of the modulator, since this normally allows a higher attainable in-band loop gain. However, this has very little or no effect on the amplitude of the tones near $f_s/2$ since the out-of-band loop gain must remain nearly the same in order to preserve loop stability. Even eighth order modulators seem to suffer from these high frequency tones [48].

A different approach to tone suppression is the introduction of a degree of 'uncertainty' into the modulation. This is accomplished either by adding dither noise into the modulator loop [40, 41] or by the use of a chaotic modulator [41, 57, 37]. The purpose of this letter is to compare these two strategies with respect to the suppression of the tones near $f_s/2$. The advantage of a chaotic modulator is that no dither noise source is needed, thus eliminating the non-trivial task of designing a temperature independent noise source with a prescribed probability density function in analog implementations.

A.2 Feedback Filter Design from NTF-prototypes

Figure A.1 shows a typical sigma-delta modulator composed of a feedback filter (i.e, loop filter) $H(z)$ and a one-bit quantizer (i.e., signum function).

When the one-bit quantizer is crudely modeled as a unity gain with an additive noise source, the noise transfer function, $\mathrm{NTF}(z)$, from the noise source to the modulator output is given by:

$$\mathrm{NTF}(z) = \frac{1}{1 + H(z)} \tag{A.1}$$

The NTF describes how the quantization noise is spectrally shaped and suppressed in the base-band.

Sigma-delta modulators are typically designed by specification of a desired high-pass noise transfer function, i.e., an NTF-prototype [1, 49, 51, 58]. From an NTF-prototype given as a rational transfer function $\mathrm{NTF}(z) = A(z)/B(z)$ the feedback filter $H(z)$ can be derived using (A.1):

$$H(z) = \frac{1}{\mathrm{NTF}(z)} - 1 = \frac{B(z) - A(z)}{A(z)} \tag{A.2}$$

Since delay free loops around a quantizer are not implementable, $H(z)$ must have at least a one sample delay, i.e., the associated impulse response must have the property $h(k) = 0,\ k < 1$ [1, 41]. Equation (A.2) shows that this is achieved when the NTF-prototype is scaled such that $A(z)$ and $B(z)$ have the same highest order z-term. This necessary scaling causes a unity pass-band gain NTF-prototype to have a higher pass-band gain and a lower stop-band attenuation after scaling.

A good in-band noise suppression can be obtained by designing the NTF-prototype as a Chebychev II high-pass filter with unit circle zeros distributed in the base-band [1] (see also [58] for optimum unit circle zero locations). Such scaled prototype filters can easily be designed using the MATLAB Signal Processing Toolbox commands [49, 35]:

```
[A,B]=cheby2(N,Rs,fb,'high');
A=A/A(1)
```

where N is the filter order, Rs is the desired base-band attenuation in dB (prior to scaling) and fb is the upper base-band edge normalized with respect to $f_s/2$.

Increasing Rs will obviously improve the suppression of in-band noise and thereby the SNR; however, the necessary scaling causes the high frequency gain of the NTF-prototype to increase and, unfortunately, too much high-frequency noise destabilizes the modulator loop, i.e., the stable amplitude range of the modulator is reduced [1, 51, 58]. This phenomenon is caused by an overload of the simple one-bit quantizer, i.e., the quantizer output power is always unity and must be shared between the in-band signal and the circulating noise.

The feedback filter design is thus a trade-off between in-band noise suppression performance and loop stability. See [51, 58] for a discussion of design trade-offs and stability criteria for high-order modulators.

A.3 Chaotic Sigma-Delta Modulators

An Nth order sigma-delta modulator is a non-linear discrete-time dynamical system, which can be analyzed by observing trajectories in a corresponding N-dimensional state-space.

Generally, the state-space has a large number of periodic trajectories (i.e., limit cycles) with different periods [25].

It can be shown that if all poles of the feedback filter $H(z)$ are inside the unit circle, the limit cycles have attracting regions in the state-space which cause the modulator to lock asymptotically into strictly periodic modes [25].

If one or more of the $H(z)$ poles are outside the unit circle, the limit cycles will be unstable, i.e., even small perturbations in the state space will be amplified exponentially in time [57]. The sensitivity to initial conditions and a high density of limit cycles are important characteristics of chaos. Chaotic systems are unpredictable on long time scales and generate non-periodic outputs. Despite these properties, chaos does generally not guarantee that the quantization noise of sigma-delta modulators is free from tones. [57, 37].

The moduli of the poles outside the unit circle is a measure of the 'amount of chaos' in the modulator, i.e., how fast nearby points in state-space diverge per time step. The base two logarithms of the pole moduli are in fact equal to the so-called Lyapunov exponents used in chaos theory [25] and these exponents express the loss of state-space information in bits per time step.

A.4 NTF-prototypes with All-pass Terms

In order to obtain chaos, $H(z)$ must have at least one pole outside the unit circle. This is achieved when an NTF-prototype with zeros outside the unit circle is used, cf. (A.2).

An initial approach was to use a standard sigma-delta modulator and scale up the $H(z)$ unit circle poles with a certain factor; however this procedure reduces the in-band noise suppression significantly due to the resulting lower loop gain [57, 37].

Another way to obtain chaos is to use a minimum phase NTF-prototype and reflect one or more zeros to their reciprocal locations outside the unit circle, as demonstrated for FIR filters in [41]. The necessary scaling of the zero-reflected prototype scales up the magnitude characteristic by the product of the moduli of the reflected zeros, and this reduces the loop stability considerably. The stability was secured by adding more quantizer levels in [41]. The benefit of zero-reflection is that the in-band noise suppression is only reduced proportional to the necessary scaling.

An alternative to zero-reflection is to introduce an all-pass term into a standard IIR NTF-prototype with all zeros on the unit circle. A first order all-pass term must naturally have a (real valued) zero outside the unit circle and a pole inside the unit circle which is the reciprocal of the zero [44]. The advantages of an all-pass extended NTF-prototype are that the zeros are preserved on the unit circle for optimum in-band noise suppression and that the degree of chaos can be adjusted independently of the NTF magnitude characteristic by adjusting the all-pass zero location. A normalized all-pass term with a zero at $z = \alpha,\ |\alpha| > 1$ has the form:

$$\mathrm{NTF}_{ap}(z) = \frac{1 - \alpha z^{-1}}{1 - \frac{\alpha}{|\alpha|^2} z^{-1}} \tag{A.3}$$

This all-pass term may be multiplied on any normalized NTF-prototype thus giving a new normalized NTF-prototype with a filter order increased by one. Since the magnitude response of (A.3) is equal to $|\alpha|$, the loop stability deteriorates as $|\alpha|$ is increased. Consequently, 'more chaos' and thereby better tone suppression can be obtained at the expense of a lower stable amplitude range.

A.5 Design and Simulation Examples

In this section, three methods for suppressing the tones near $f_s/2$ are compared , i.e., NTF-prototype zero scaling, one-bit quantizer dithering, and finally, the use of an all-pass extended NTF-prototype.

For the sake of comparisons, the simulation examples will all be for sixth order modulators based on Chebychev II prototypes designed with MATLAB. All modulators are intended for 64 times oversampling, i.e., fb=$1/64 \cdot 20/22.05$ (suitable for e.g., 20 kHz baseband and $64 \cdot 44.1$ kHz sample rate). An additional design criteria was to ensure a stable input range up to 0.35 relative to full scale. The simulations use a DC-input of 1/256 corresponding to possible high frequency tones that might intermodulate and generate base-band tones at $f = 1/128 \cdot f_s/2$. The spectra shown in this section are obtained by averaging Kaiser-Bessel windowed 8k FFT power spectra for sequences of two mill. one-bit samples. The heavy averaging allows tones to be distinguished from random noise. In this context, the designation 'tone free' is used when possible tones are comparable to or below the noise floor provided by the given spectral resolution. An enhanced spectral resolution (e.g., larger FFT size) can naturally reveal more tones because the noise floor is lowered. For each spectrum the in-band noise power is computed by summing up the power of the appropriate FFT-bins.

The first modulator tested had a prototype designed with the parameters N = 6 and Rs = 120 dB. This modulator is stable for input amplitudes up to approx. 0.35 relative to full scale. Figure A.2 shows the simulated output power spectrum. The in-band noise power is approx. -142 dB (relative to full scale) corresponding to more than 21 bits of resolution. However, the tone seen near $f_s/2$ is no more than 6.7 dB below full scale — the tone is in fact stronger than any possible modulator input!

Experiments with scaling of the NTF-prototype zeros were then carried out. This resulted in a design based on a prototype with N = 6 and Rs = 90 dB. A zero scaling by a factor of 1.06 was necessary to suppress the high frequency tones below the noise floor. The resulting tone free spectrum is shown on Figure A.3. The in-band noise power is as high as -61 dB and this is a consequence of the NTF-zeros being removed from the unit circle. Furthermore, the experiments showed that this modulator is too unstable for any practical purposes.

The next approach was to add a dither noise source to the quantizer input. Obviously, this introduces more noise into the loop which consequently becomes more unstable. The NTF-prototype must therefore be designed with a somewhat lower Rs such that the same stable input range is preserved when dither is applied. The combination of N = 6, Rs = 95 dB and a spectral white dither source with a uniform amplitude density from -0.5 to 0.5 resulted in a tone free design with a stable input range up to approx. 0.35. The spectrum is showed on Figure A.4 and the in-band noise power is approx. -115 dB. A peak at -26.4 dB is seen at $(1 - 1/256)f_s/2$. When a lower dither span is applied, both a higher SNR and stable input range is obtained; however, the high frequency tone becomes stronger.

Finally, a sixth order modulator was obtained from a fifth order NTF-prototype extended by a first order all-pass term cf. (A.3). Experiments have shown that negative values of α are far more efficient than positive values in respect to suppression of high frequency tones — probably because an NTF-zero near $z = -1$ corresponds to $f = f_s/2$. The combination of N = 5, Rs = 80 dB and $\alpha = -1.25$ resulted in a tone free modulator with a stable input range up to approx. 0.35. The spectrum is showed on Figure A.5 and the in-band noise power is approx. -105 dB. A peak at -37.9 dB is seen at $(1-1/256)f_s/2$.

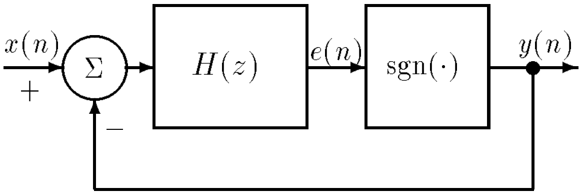

Figure A.1: Sigma-Delta modulator with feedback filter $H(z)$

The modulator has obviously a fifth order noise shaping amplitude characteristic and this is limiting the SNR. If the total order is increased to 7, it is possible to obtain a tone free design with a better SNR and the same stable amplitude range: the parameters N = 6, Rs = 92 dB and $\alpha = -1.2$ yields an in-band noise power of approx. −114 dB, i.e. almost the same performance as for the dithered sixth order example.

A.6 Conclusions

It has been demonstrated that the use of chaos effectively can substitute dither as a means for suppression of the predominant tones near half the sample rate in one-bit sigma-delta modulators. In particular, the use of a general all-pass term seems to be very advantageous. Investigations have shown that a first order all-pass term with a real valued negative zero outside the unit circle is a feasible choice. The price paid for the tone suppression is that both dithering and the use of chaos reduce the loop stability significantly. In order to restore stability, the in-band loop gain must be reduced resulting in a lower attainable SNR for a given modulator order. However, this can to some extend be remedied by increasing the modulator order.

The presented examples indicate that the dithered modulator obtains a 9 dB better SNR than a chaotic modulator for the same total order and stable amplitude range. However, this is to some extend counterbalanced by the high-frequency tone being approx. 11 dB weaker for the chaotic modulator example than for the dithered. Future research should therefore attempt to uncover the exact relations between tone strength, dither span and the all-pass zero modulus.

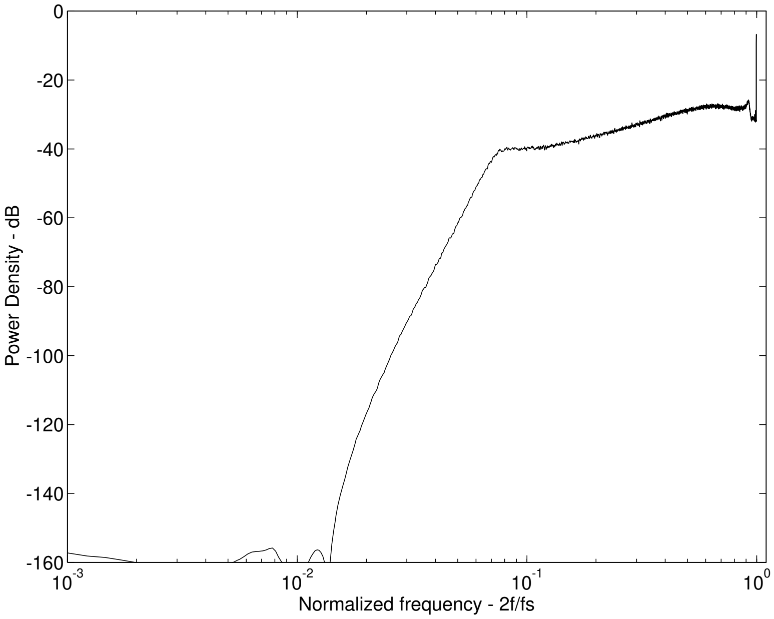

Figure A.2: Power spectrum from a simulation of a sixth order modulator with Rs = 120 dB.

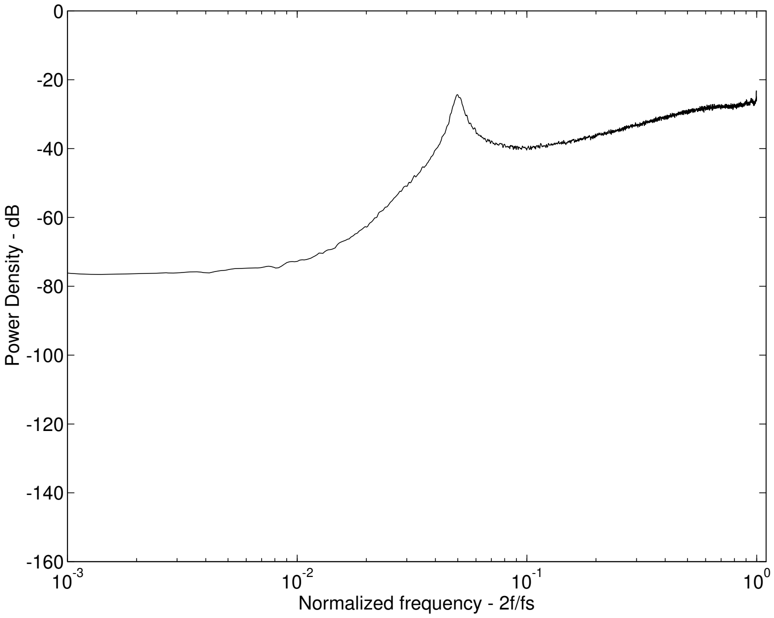

Figure A.3: Power spectrum from a simulation of a sixth order modulator with Rs = 90 dB and an NTF-zero scaling of 1.06.

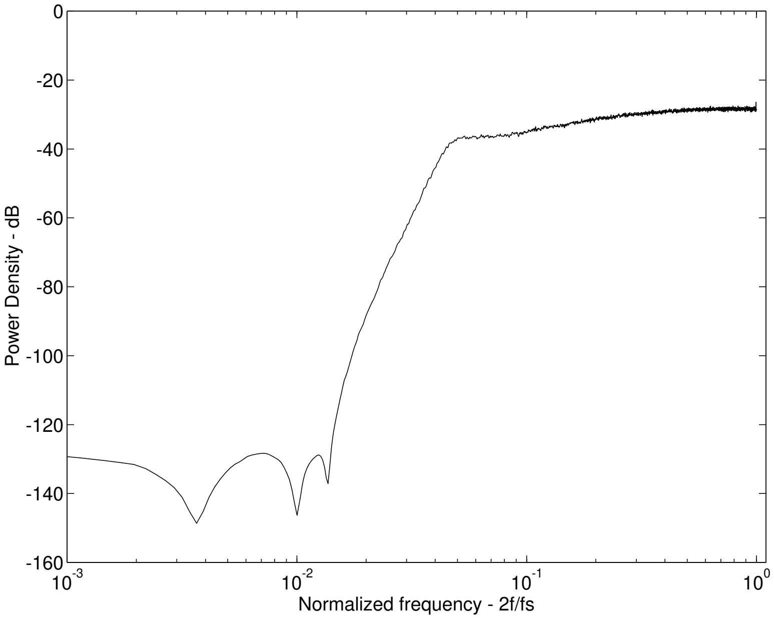

Figure A.4: Power spectrum from a simulation of a sixth order modulator with Rs = 95 dB using dither uniformly distributed over the interval [-0.5,0.5].

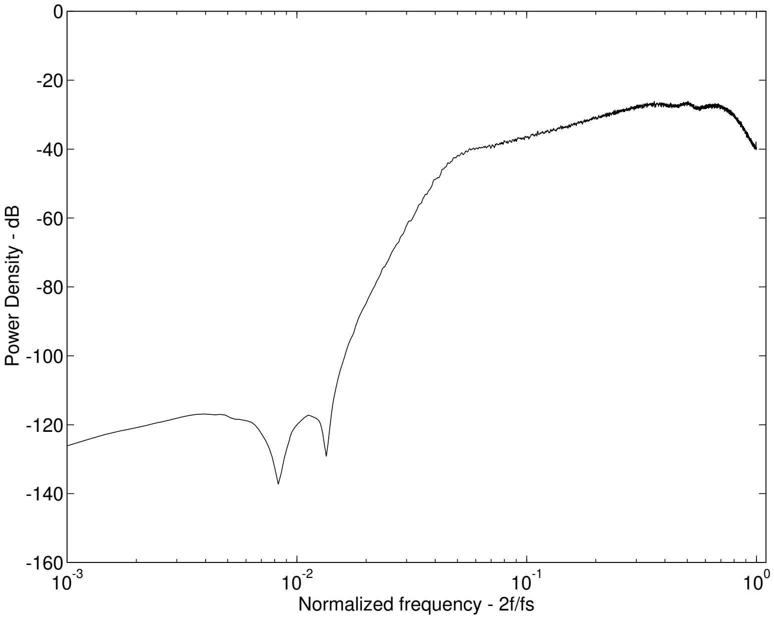

Figure A.5: Power spectrum from a simulation of a fifth order modulator with Rs = 80 dB extended to sixth order using an all-pass term with $\alpha = -1.25$.

Appendix B

Appendix for Chapter 9

This appendix is connected to chapter 9 and is a reprint of reference [52]: L. Risbo, "Improved stability and performance from Σ-Δ modulators using one-bit vector quantization," *IEEE Proc. ISCAS'93*, pp. 1361–1364, May 1993.

Appendix C

Implementation of an Audio DAC

This appendix is a reprint of the reference [48]: L. Risbo, "FPGA based 32 times oversampling 8th order sigma-delta audio DAC" presented at the 96th Audio Engineering Society Convention, Amsterdam, The Netherlands, 1994 February 26 — March 01, Preprint 3808.

The paper presents a Σ-Δ D/A-converter which was developed during the Ph.D. study. The references are included in the bibliography section on page 173.

Abstract

*A novel Sigma-Delta based audio DAC system is presented. The converter operates at only 32 times oversampling (*1.4 MHz*) and uses an 8th order Sigma-Delta modulator implemented in a single FPGA circuit. Two different DAC circuits have been implemented: One using switched capacitor charge pulses and one using non-return-to-zero pulses. A VCXO based PLL ensures sampling clock recovery free from deterministic jitter components.*

C.1 Introduction

In recent years, Sigma-Delta based oversampling A/D and D/A converters have become very popular for demanding audio applications [38, 1, 61, 23]. Especially, converters using two level signal encoding, i.e., one-bit single loop Sigma-Delta converters, have attractive properties: The almost perfect level linearity and the lack of differential non-linearity. Furthermore, these devices do not require high precision trimming of current sources as for the conventional multibit types.

A Sigma-Delta modulator (SDM) consists of a one-bit quantizer (i.e., a comparator) and a suitable loop filter – or feedback filter. The one-bit output signal of the quantizer is forced to reproduce the input signal due to the feedback loop. The resulting output signal will consequently consist of the modulator input signal and a quantization noise component which is spectrally shaped by the feedback loop. The use of heavy oversampling enables the quantization noise to be removed from the audio band, and therefore concentrated at high frequencies. When the resulting one-bit signal is low-pass filtered, the input signal is reproduced with a high signal-to-noise ratio (SNR).

A Sigma-Delta D/A-converter uses a digitally implemented Sigma-Delta modulator to resample the high resolution input signal (i.e., 16–20 bit) and produce a one-bit signal at a sample rate typically 64–1024 times higher than for the input. The one-bit signal is subsequently fed to a one-bit DAC, i.e., a switch, and the resulting two-level current or voltage is low-pass filtered in the analog domain.

Early Sigma-Delta based DACs used a low order feedback filter and a high oversampling ratio (e.g., second order modulator with 256 times oversampling [38]). A consequence is that the design of the analog converter section becomes very demanding due to the very high sample rate ($> 10\,\mathrm{MHz}$). Furthermore, the quantization noise of second order SDMs is generally not very random and is somewhat correlated with the input signal. However, this can be remedied by proper dithering.

The current trend is to use high order feedback filters. Recently, a 64 times oversampling 5th order DAC has been reported [61]. High order feedback filters allow a more frequency selective noise-shaping and consequently a lower oversampling ratio is needed. In addition, the resulting quantization noise becomes less correlated with the audio signal due to the higher complexity in the one-bit encoding process. These converters have properties which are comparable to those of analog amplifiers: The noise is almost independent of the input and is therefore not perceived as distortion of the input signal. Since a Sigma-Delta modulator of course is a deterministic system, the quantization noise is not random; however, high-order feedback makes the relationship between input signal and the noise inscrutable to the ear. Conventional multibit converters introduce, unlike the one-bit types, an error which is directly dependant on the signal amplitude or digital input code.

The objective of this paper is to present a new converter based on an 8th order modulator operating at only 32 times oversampling, i.e, approx. 1.4 Mhz for a usual CD-standard input. The entire converter, including standard SP/DIF digital audio interface and analog post filtering, is exclusively made of standard off-the-shelf components. The digital modulator section is implemented in a general purpose Field Programmable Gate Array (FPGA) circuit.

C.2 Converter Overview

The converter consists of six blocks shown on Figure C.1:

- Digital interface receiver
- Interpolation filter
- 8th order SDM
- One-bit DAC
- Analog post filter
- Clock jitter attenuator (VCXO PLL)

The digital interface receiver (Crystal CS8412) demodulates the standard SP/DIF or AES/EBU digital audio format and regenerates a clock and a serial 16-bit data signal. Subsequently, this signal is fed to an interpolation chip (NPC SM5803) which oversamples the signal eight times. The resulting 18-bit output signal is fed to the 8th order Sigma-Delta modulator, which oversamples the signal four times additionally and performs the one-bit encoding. The analog section of the converter consists of a one-bit DAC operating on the 32 times oversampled modulator output and a post filter which removes the out-of-band quantization noise. The last block is a voltage controlled crystal oscillator (VCXO) based Phase Locked Loop (PLL) for clock jitter attenuation. The PLL suppresses timing jitter on the clock recovered by the digital interface receiver which otherwise would cause a degradation of the DAC performance. In the following sections, the Sigma-Delta modulator and the analog section will be described in greater detail along with some background theory and design considerations.

C.3 Sigma Delta Modulator

C.3.1 Design and Analysis

The modulator topology used for the DAC is shown on Figure C.2 for even modulator order, N. This topology is known as the *multiple feedback* type [1, 55, 49]. The input signal is fed trough a cascade of delaying discrete time integrators to the one-bit quantizer which output is distributed back to each of the integrators in the chain using the weight factors b_i. Additional local feedback is applied around two adjacent integrators as indicated by the a_i-coefficients. This topology has a very regular structure which is a cascade of simple second order sections. Furthermore, this topology has very low sensitivity to coefficient errors [49].

The local feedback introduced in the integrator chain implies that each of the a_i-coefficients introduces the following complex open loop pole pairs:

$$d_i = 1 \pm j\sqrt{a_i} \tag{C.1}$$

Note that these poles are generally outside the unit circle implying that the integrator chain is totally unstable without the feedback. Furthermore, poles outside the unit circle will destabilize possible limit cycles making the modulator chaotic and non-periodic [49].

By linearized analysis, the modulator can be characterised by a Signal Transfer Function, $STF(z)$, and a Noise Transfer Function, $NTF(z)$ [1, 54]. The $STF(z)$ shows how the input signal is affected and $NTF(z)$ shows the noise-shaping characteristic for the quantization noise.

One advantage of the multiple feedback modulator is that $STF(z)$ becomes a low-pass filter [49]. This relaxes the requirements of the preceeding interpolation filter. Especially for a high modulator order (e.g., 8th order) the last stages of the interpolation filter can be replaced by a simple zero order hold circuit [49].

When a modulator is designed, the a_i coefficients are specified at first such that the open loop poles are distributed within the audio-band which ensures good noise suppression for $NTF(z)$. Subsequently, the b_i-coefficients are chosen for an acceptable trade off between SNR and loop stability. Furthermore, the choice of b_i must also ensure that the resulting $STF(z)$ is reasonably flat in the audio-band.

The implemented modulator uses coefficients obtained using tools described in [54, 51]. The 8th-order modulator was optimized for 32 times oversampling, a flat $STF(z)$ and stable operation up to output amplitudes of 0.4 relative to full scale. The a_i-coefficients used were rounded to power of two values allowing binary shift operations to be used instead of general multiplications.

The maximum stable amplitude range (MSAR) is a very important design parameter. Generally, any MSAR can be obtained by chosing the b_i-coefficients properly [49]. However, optimum SNR is obtained when MSAR is around $0.3 - 0.35$ relative to full scale [49]. If MSAR is lower, the SNR will deteriorate due to the lower maximum signal power. For a higher MSAR, the in-band noise power increases faster than the signal power, and consequently, the SNR decreases. Modulators with a high MSAR also exhibit higher intrinsic harmonic distortion and more variation of the in-band noise power when varying the signal power [49]. When other analog noise sources are taken into consideration, the optimum MSAR may be shifted towards higher values.

Figure C.3 shows power spectra estimates of the modulator output obtained by simulations using the actual 24-bit integrators used in the implemented modulator. Figure C.4 shows the in-band RMS noise as function of the output DC-amplitude. This graph is obtained from simulations of the modulator with an extremely slowly increasing ramp input. Notice the absence of spikes and irregularities on this curve. This type of plot is rather revealing for the modulators ability to decorrelate the noise from the input. Furthermore, MSAR can be found using these ramp input simulations. In this case MSAR was 0.412. The SNR for sinusoidal input with amplitude 0.4 is estimated to 97 dB using Figure C.3.

C.3.2 Modulator Implementation

Figure C.2 shows that for the multiple feedback type modulator topology, only one-bit values, i.e., -1 or 1 are multiplied on the b_i-coefficients and, consequently, only an ADD or a SUB operation is needed depending on the quantizer output. Furthermore, when the a_i-coefficients are designed as power of two values, binary shift operations can replace general multiplications. The absense of general multiply operations makes the multiple feedback modulator very well suited for implementation in programmable logic.

The 8th order modulator was implemented in a single 3000 gate equivalent Xilinx XC4003 FPGA circuit using hardware efficient bit-serial arithmetic. The implementation consists of four almost identical second order modulator sections, a shared control circuit and a signal input conditioner. A fully synchronous design was used operating on a $384 f_{in}$ input clock. This gives 12 clock cycles for processing of each output sample when using 32 times oversampling.

An other FPGA implementation of a Sigma-Delta modulator has been reported in [27]. In [27] a fourth order multiple feedback modulator was implemented in a single 9000 gate equivalent Xilinx XC3090 FPGA using bit-parallel arithmetic. However, the local feedback introduced by the a_i-coefficients was not implemented.

The arithmetic operations are performed serially operating on 2-bits for each clock cycle allowing an integrator resolution of $2 \cdot 12 = 24$bits. Each integrator consists of a 12-bit deep and 2-bit wide shift register and a number of 2-bit adders. This provides a state-space of totally $8 \cdot 24 = 192$-bits which ensures extremely long and complex modulator idle pattern sequences. Binary scaling shifts are used between the integrators in order to fully utilize the dynamic range of the integrators. The binary (right) shift operations are implemented by taking intermediate outputs from the integrator shift registers. Sign extension of the right shifted data stream is implemented using a multiplexer which retransmits the integrator sign bit when the most significant bit is reached.

Figure C.5 shows a simplified diagram of one of the four second order sections. Each section performs five ADD/SUB-operations and three shift operations i.e., one for each of the integrators and a shift for the a_i-coefficient multiplication. This corresponds to a total computational capacity of approx. 50 mill. 24-bit operations per second for the entire modulator.

The control circuit generates a number of shared controls signals which are used to control the modulator. A modulo 12 counter genererates a 4-bit address sequence which is fed to a number of look-up ROM tables producing different control signals. For instance, the b_i-coefficients are supplied bit-serially to the integrators from eight ROM-tables. Since the used FPGA holds programming information in a SRAM memory, new modulator coefficients can be programmed in miliseconds. In fact, during the design and test phase of the project, the FPGA was programmed by downloading information from a PC. A PC-program was developed which translates loop filters designed in a high-level environment (MATLAB) into boolean equations for the FPGA. This enables many loop filters to be tested under realistic conditions without redesigning silicon.

The used FPGA is organized as a 10 by 10 array of Configurable Logic Blocks (CLBs) which each contains two D-type latches and two four input arbitrary logic function generators. Additionally, the chip contains a large number of input/output blocks (IOBs) which each contains an input and an output latch connected to a pin. The ADD/SUB blocks in Figure C.5 use each two CLBs and the MUX and ROM functions use each one CLB. A total of 15 CLBs are used for each second order section. In addition, $2 \cdot 2 \cdot 12 = 48$ latches are needed for each second order section. These latches are partly taken from the remaining CLBs and partly from the IOBs of unused pins.

The input conditioner accepts the 8 times oversampled 18-bit serial input signal from the interpolation filter and loads a parallel input register. In order to upsample to 32 times oversampling, the content of the input register is retransmitted serially 4 times to the modulator for each input sample.

No hardware precausions against modulator instability were taken in the implementation except that proper coefficient scaling ensures that the output amplitude is limited to 0.4 relative to full scale for maximum digital input. However, special precausions have been taken in the design of the loop filter [54] such that the modulator operates reliably up to the 0.4 amplitude limit. The modulator has successfully been tested extensively with both full scale sinusoidal input and music programme input.

C.4 Analog Section

The objective of the analog design was to achieve superior sound quality. Two different one-bit DACs and two different analog post filters have been designed using standard discrete components.

C.4.1 One-bit DAC overview

The one-bit DAC converts the binary bit-stream from the modulator into an analog waveform which can be interpreted as a convolution between the digital one-bit signal $p(k)$ and an analog pulse, $g(t)$:

$$d(t) = \sum_k p(k) g(t - k \cdot T) \tag{C.2}$$

where T is the sampling time interval and $p(k) \in \{-1, 1\}$.

In the frequency domain, the convolution with $g(t)$ will act as a filtering of the digital one-bit signal. Fundamentally, two one-bit DAC types can be discriminated, i.e., the pulse type and the hold type. Pulse type DACs generate concentrated current pulses by charging or discharging capacitors between reference voltage levels [38, 61]. The hold type DACs hold a reference voltage or current level in a certain period of time. Usually the hold time is derived from the system clock. The hold type DAC can further be divided into return-to-zero (RTZ) types [23] and non-return-to-zero (NRZ) type. The latter holds the reference level for precisely one sampling time interval, and RTZ types return to zero output before the next input sample. Consequently, all three types can be destinguished by their corresponding $g(t)$ pulses.

C.4.2 Spectral DAC Properties

The three types of DACs have a number of different properties. The most apparent differences are the spectral properties induced by the $g(t)$ waveforms. For the pulse type DAC, $g(t)$ approaches a delta function which has a very wide-band spectrum. Consequently, the periodic spectrum of the digital DAC-input found at multiples of the sampling frequency is preserved up to high frequencies in the analog DAC-output. This means that the succeeding analog post filter must be able to handle wide-band signal and noise components without generating intermodulation components folded back into the audible range [28].

The hold type DACs have generally more intrinsic low-pass filtering due to longer $g(t)$ pulses. The amplitude response for the Fouriertransform of a unity amplitude square pulse of duration τ is given by:

$$|G(f)| = \tau \frac{\sin \pi f \tau}{\pi f \tau} \tag{C.3}$$

In particular, for $\tau = T$, $G(f)$ has zeros for multiples of the sampling frequency $f_s = 1/T$. This means that NRZ DACs have a first order suppression of the image spectra at multiples of the sampling frequency. For 32 times oversampling, the image spectrum at f_s will be attenuated at least $37\,dB$. The author believes that this image spectrum attenuation is very beneficial for the sound quality because much high-frequency distortion is avoided in the the succeeding analog signal processing. In fact, the usual multibit type converters use NRZ conversion; hence, they have the same attenuation of image spectra. It is very likely that the improvement in sound quality obtained when oversampling multibit DACs were introduced is attributable to the improved image spectra attenuation which is a consequence of oversampling.

For a RTZ DAC, τ will be less than T and the zeros of $G(f)$ will generally not be at multiples of f_s. Furthermore, the output signal amplitude scales directly with the hold time τ, and consequently, the analog SNR will deteriorate when other analog noise sources are present.

C.4.3 Clock Jitter Sensitivity

When clock timing jitter is taken into consideration, the pulse type and hold type DACs will show different properties [14]. Clock jitter is the designation for timing errors of the sampling instants which cause errors in the reproduced analog waveforms. It will be assumed that the sampling instants are displaced in time by the jitter sequence $t(k)$, i.e., the sampling instants are: $s(k) = k \cdot T + t(k)$. For the NRZ DAC, the hold time will be modeled by the jitter sequence, since the output signal only changes state at the sampling instants. It can be shown [14] that the error induced by the jitter sequence can be modeled as the following digital error sequence:

$$e_{\mathrm{NRZ}}(k) = (p(k) - p(k-1))\frac{t(k)}{T} \tag{C.4}$$

This equation shows that the error is proportional to the modulator output frequency due to the differencing operation. Notice that the equation also applies for usual multibit converters. The problem when dealing with one-bit DACs is the extreme content of high-frequency quantization noise in the $p(k)$-signal which has more noise power than signal power. When the jitter sequence is a wide-band and uncorrelated noise, the error signal of (C.4) will be almost white. The variance of the error sequence is then approximately the product of the jitter variance and the variance of $p(k) - p(k-1)$. Simulations have shown that the variance of the difference signal $p(k) - p(k-1)$ is approx. 2.7 for the used 8th order modulator with zero input. The in-band SNR for R-times oversampled sinusoidal output with amplitude A and RMS jitter t_{RMS} is consequently:

$$\mathrm{SNR}_{\mathrm{NRZ}} = 10 \log \frac{R\ A^2/2}{2.7\left(\frac{t_{\mathrm{RMS}}}{T}\right)^2}\ \mathrm{dB} \tag{C.5}$$

In order to obtain a SNR of e.g., 100 dB with 32 times oversampling ($f_s = 1.4\,\mathrm{MHz}$) and sinusoidal output with amplitude 0.4, the RMS wide-band jitter must be less than 7 picoseconds !

The pulse type DACs have a different sensitivity to clock jitter due to the fact that the pulse waveform is typically the result of a capacitor charge or discharge which is independent of the sampling time instants.

In [14] it has been shown that the equivalent digital jitter error signal for pulse type DACs could be interpreted as the product signal $p(k)t(k)$ filtered by a differentiator with transfer function $H(f) = j2\pi f$. This means that the jitter error is noise-shaped and suppressed significantly for low frequencies. Consequently, pulse type DACs can tolerate much more wide-band clock jitter.

The in-band jitter noise power can be calculated by integrating the shaped noise power density in the audio band. When assuming that the error generating sequence $p(k)t(k)$ is white with variance t_{RMS}^2, one arrives at the following SNR figure:

$$\mathrm{SNR}_{\mathrm{pulse}} = 10 \log \frac{3\ R^3\ A^2/2}{\pi^2 \left(\frac{t_{\mathrm{RMS}}}{T}\right)^2}\ \mathrm{dB} \tag{C.6}$$

In order to obtain a SNR of e.g., 100 dB with 32 times oversampling ($f_s = 1.4\,\mathrm{MHz}$) and sinusoidal output with amplitude 0.4, the RMS wide-band jitter must be less than 200 picoseconds. This shows that the pulse type DACs are considerably less sensitive to wide-band clock jitter. However, it should be emphasized that the shown calculations only have taken the interaction between the quantization noise and the jitter into account. It has been pointed out in [14] that pulse type DACs can be more sensitive to deterministic jitter than NRZ DACs for low frequency signal output.

C.4.4 Sensitivity to Unequal Rise and Fall Times

A major drawback of NRZ one-bit DACs is their high sensitivity to unequal rise and fall times of the output signal. The problem is that the $g(t)$ pulses will depend slightly on the previous output code in case of unsymmetrical transition times. This phenomenon gives rise to noise and distortion of the analog output.

The resulting error could be interpreted as a kind of 'auto jitter', i.e, the resulting transition time or sampling instant depends on the DAC input code. If the rise time is e.g., longer than the fall time, the equivalent transition time will be delayed for a positive input code. This corresponds to a jitter sequence of $t(k) = \alpha T(p(k)+1)/2$ and this results in an error sequence of $\alpha(p(k) - p(k-1))(p(k)+1)/2$ cf. (C.4). This signal is equal to 2α when $p(k)$ changes from low-to-high and zero otherwise. The error signal contains obviously second harmonic distortion due to the squaring of the signal $p(k)$. Furthermore, the error signal contains wide band noise due to intermodulation distortion of the high-frequency noise in $p(k)$. The error signal spectrum can be asessed by simulations. It is estimated by simulations that the rise and fall time difference must be less than 24 picoseconds for 100 dB SNR with the actual modulator. The error can typically be reduced considerably by a differential design giving symmetric transitions.

Pulse type and RTZ type DACs do not have this intrinsic error mechanism as far as the DACs are able to settle completely between the output pulses; however, incomplete and unsymmetric settling might cause a similar error.

C.4.5 Deterministic Components Near $f_s/2$

Almost all Sigma-Delta modulators produce variuos tones near $f_s/2$. If the modulator is fed with a constant giving a mean output value of m_p, a spectral peak will be present in the output spectrum at the frequencies $(1 \pm m_p)f_s/2$ (see Figure C.3) [49, 24]. The peaks might intermodulate in the analog circuitry and cause audio-band tones at $m_p f_s$.

This mechanism is a kind of frequency modulation. Consequently, if the modulator input is sinusoidal with frequency f, the modulator output should contain peaks or sidebands at $f_s/2 \pm nf$ for integer n. This conjecture is confirmed by Figure C.6 which shows an output spectrum for the used modulator with sinusoidal input with frequency $0.025 f_s/2$ and amplitude 0.2 relative to full scale. Again, nonlinearities or clock jitter could turn these peaks into harmonic distortion in the audible range. Obviously, interfering signals with frequency $f_s/2$ should be avoided as stated in [24].

In [54] it is revealed how Sigma-Delta modulators generally can be designed such that thay do not produce spectral peaks near $f_s/2$.

C.4.6 Pulse DAC Implementation

A pulse type DAC was implemented using discrete transistors and CMOS switches. Figure C.7 shows a simplified schematic of the pulse DAC. The DAC operates with only one capacitor which is either charged to a positive or a negative reference level depending on the digital one-bit input code. In between each sample, the capacitor is discharged to ground. The capacitor charge/discharge and reset operations are controlled by two non overlapping clock phases. The current pulses used for the capacitor charge or discharge are taken directly from the DAC output using bipolar transistors coupled as current conveyers. These transistors are furthermore isolated from the DAC output by two cascode transistors. This ensures that the DAC output signal cannot modulate the reference levels used for the switched capacitor. The entire DAC is thus implemented in current mode topology, and furthermore, the design is totally based on feedforward coupling. This ensures a high bandwidth and a good high frequency linearity.

C.4.7 Current Switch NRZ DAC Implementation

A NRZ current mode DAC was implemented using discrete bipolar transistors. A simpified schematic is found in Figure C.8. The heart of the DAC is a current switch based on a diode bridge. The incoming TTL-signal from the CMOS latch can either draw current from an upper or a lower reference current cource depending on the logical state. The current source which is not feeding the latch-output must feed the input transistors Q1 and Q2 which transmit the current to the output node through folded cascodes (Q3 and Q4). This arrangement ensures that the DAC output voltage does not modulate the reference currents. Again, the circuit is free from feedback and operates primarily in current mode.

The threshold level for the current switch can be adjusted by varying the base voltages of the input transistors Q1 and Q2. This feature makes it possible to adjust the delays of the low to high and high to low transitions relative to the clock signal. Consequently, this trimming option can null out the noise and distortion due to unequal rise and fall times. The adjustment can be done manually simply by means of ear-phones. However, the adjustment seems to be slightly temperature dependent.

C.4.8 Analog Post Filter

The analog post filter ‘decodes’ the one-bit signal, i.e., removes the out of band quantization noise components which could cause distortion in the succeeding analog circuitry. The post filter should be able to handle noise and deterministic components far into the MHz range without generating intermodulation components. These requirements call either for a passive design or the use of very high bandwidth and non slew-rate limiting active devices.

The use of 8th order noise-shaping makes it very difficult to suppress the quantization noise which rises very abruptly just over the audible range. This is the major drawback of low oversampling and high-order DACs. The filter characteristic used is a 5th-order Butterworth filter with a cut-off frequency at 25 kHz. Additionally, a transmission zero at approx. 58 kHz has been introduced in order to improve the noise suppression. Figure C.9 shows the amplitude response. The filter has approx. 1.5 dB amplitude loss at 20 kHz. The phase response can be corrected in the digital domain due to the minimum phase characteristic.

The filter results in a wide-band SNR of only 48 dB due to residual noise between 20 and 70 kHz which is difficult to suppress. This residual out-of-band noise is considered as harmless to the succeeding amplifiers and speakers, though a better suppression would be desirable.

The post filter has been implemented both as a passive LCR design and as an active design (Figure C.10). The passive design uses inductors with large air-gapped ferrit cores and high voltage polypropylene film capacitors for best linearity. The active design consists of a third order all-pole Sallen-Key filter followed by a Sallen-Key twin-T filter giving the remaining two complex poles and zeros. The active elements are based on AD811 current feedback op-amps with 140 MHz bandwidth, 2500 V/μS slew-rate and 100mA output current capability. The impedance level for both the active and passive filter is 300 Ω. Notice, the input capacitors for both filters which limit the DAC slew-rate.

C.4.9 Jitter Attenuator

A narrow band Phase Locked Loop (PLL) has been implemented which suppresses timing jitter on the recovered clock. It has been shown [14, 36] that the clock recovered from a SP/DIF interface can contain large amounts of deterministic jitter components due to both poor integrated circuit decoupling and insufficient interface bandwidth. It is strongly suspected that these jitter components can deteriorate the sound quality of a DAC.

The PLL is based on a Voltage Controlled Crystal Oscilator (VCXO) operating at $384f_{in}$. The used phase detector (AD9901) and PLL loop filter gives a closed loop cut-off frequency at 20 Hz. Jitter is asymptotically attenuated by 18 dB/Octave above approx. 100 Hz. The purified clock is fed directly to the one-bit DAC latches.

C.5 Measurement Results

An Audio Precision System One measurement system was used to meausure the converter performance. The test system supplies a digital test signal (sinewave) to the SP/DIF input of the converter and the resulting analog output is analysed. Figure C.11 shows the output spectra for a 0 dB full scale 1 kHz dithered 18-bit input for the pulse DAC with the active post filter and the NRZ DAC with the passive post filter. Notice the low harmonic distortion of the NRZ DAC with passive filtering. With −60 dB input, the signal to noise ratio below 20 kHz was measured to 34.1 dB and 31.4 dB, for the pulse and NRZ DAC, respectively. This corresponds to dynamic ranges of 94.1 dB and 91.4 dB. The figures include quantization noise present in the dithered digital input.

Figure C.12 shows the level linearity of the pulse DAC and active filter.

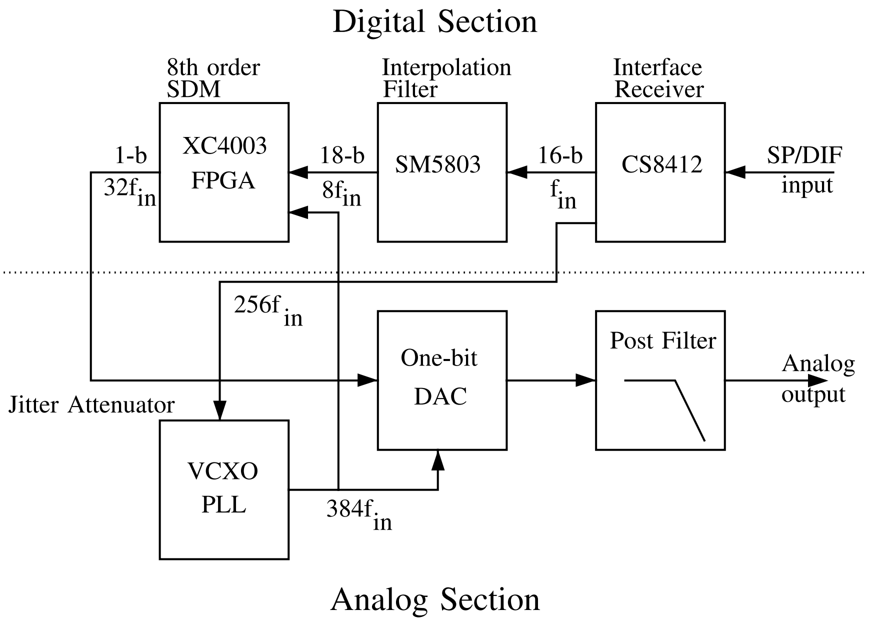
Digital Section
8th order SDM
Interpolation Filter
Interface Receiver
XC4003 FPGA
SM5803
CS8412
1-b
32f_in
18-b
8f_in
16-b
f_in
SP/DIF input
256f_in
One-bit DAC
Post Filter
Analog output
Jitter Attenuator
VCXO PLL
384f_in
Analog Section

Figure C.1: D/A converter overview showing three digital and three analog building blocks.

C.6 Conclusions

The implementation of an entire 8th order and 32 times oversampling one-bit Sigma-Delta audio DAC system has been addressed. The advantage of high-order encoding is better quantization noise decorrelation, and the advantage of a low oversampling ratio is a better linearity in the analog section. The disadvantage of this approach is the demand for a high-order analog post filter in order to suppress the sharply rising quantization noise. However, digital phase correction can be implemented.

Furthermore, it has been demonstrated that the digital modulator can be implemented in a single general purpose FPGA circuit using bit-serial arithmetic.

The sound quality of this DAC system has been judged subjectively as being superior to a number of tested commercial designs including both multibit and one-bit types. Especially, the sound from the NRZ DAC with the passive filter is very transparant. The author belives that this is attributable to both the choice of modulator and the careful and simple feedforward analog design , including the use of jitter attenuation and passive filters.

Acknowledgements

The author is very grateful to Thomas Mørch who inspired the author to use the diode bridge current switch and designed the printed circuit boards for the converter.

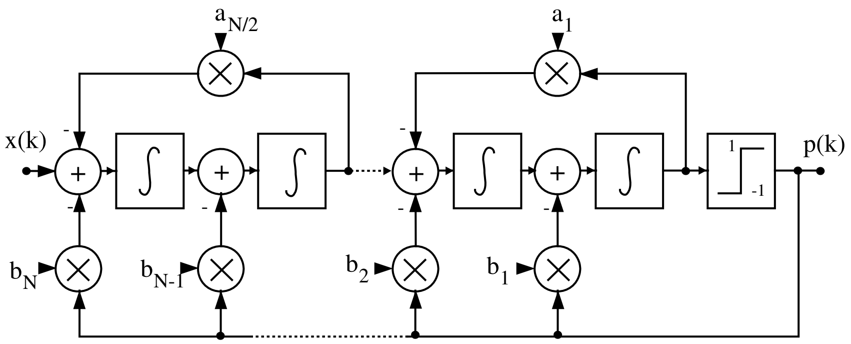

Figure C.2: Sigma-Delta modulator with multiple feedback topology for even order N. $x(k)$ is the input and $p(k)$ is the output. The blocks marked $\int$ are time-discrete delaying integrators with transfer function $1/(z+1)$.

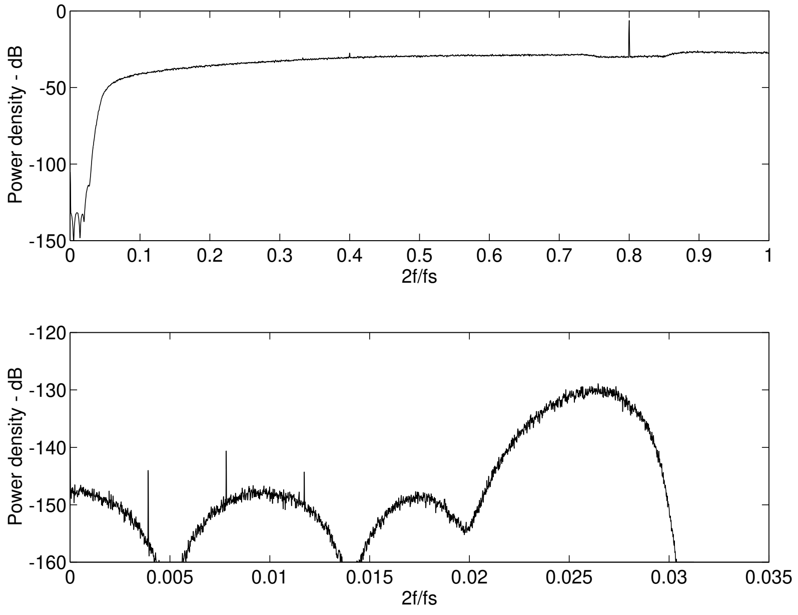

Figure C.3: Power density spectra estimated using averaged 8k FFTs with Kaiser-Bessel window on bit-exact 24b simulations of the modulator output (0 dB corresponds to a full-scale sinewave). Upper graph: Full spectrum of modulator with DC-input 0.2 (2 mill. samples). Lower graph: 32 times decimated modulator output with DC-input 1/512 (500.000 decimated samples).

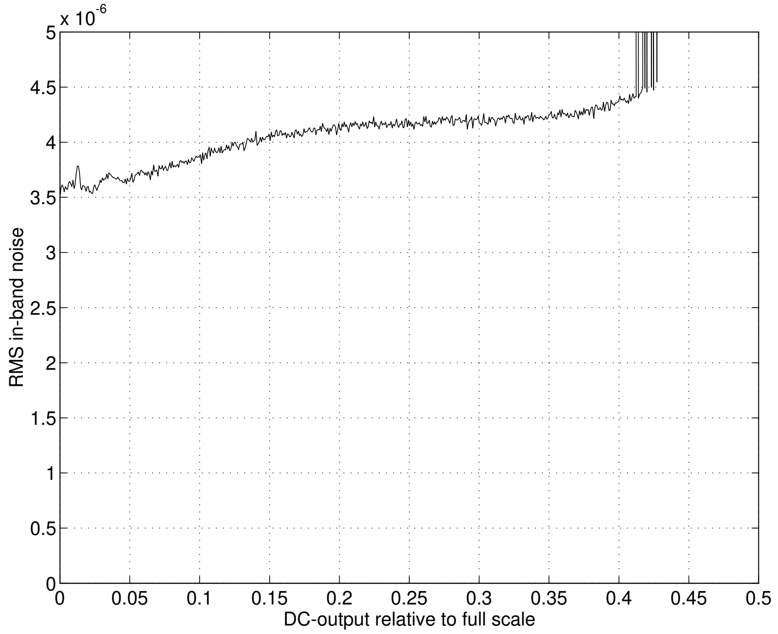

Figure C.4: In-band modulator output RMS noise versus DC-output measured by bit-exact 24b simulations and decimation. Each of the approx. 512 data points is based on decimation of 1.6 mill. modulator samples. The spikes at the end of the plot are the onset of instability.

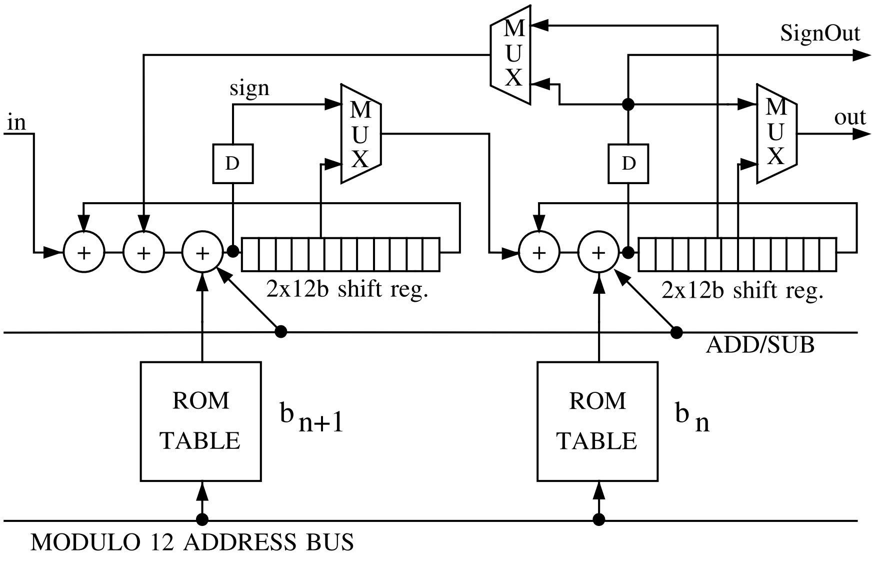
MUX
SignOut
sign
MUX
out
MUX
in
D
D
+
+
+
+
+
2x12b shift reg.
2x12b shift reg.
ADD/SUB
ROM TABLE
b n+1
ROM TABLE
b n
MODULO 12 ADDRESS BUS

Figure C.5: FPGA implementation of a second order multiple feedback modulator section. Four of these sections are cascaded to produce an 8th order modulator. The 'SignOut' signal of the last section is the output, $p(k)$, which is fed back to the ADD/SUB control line.

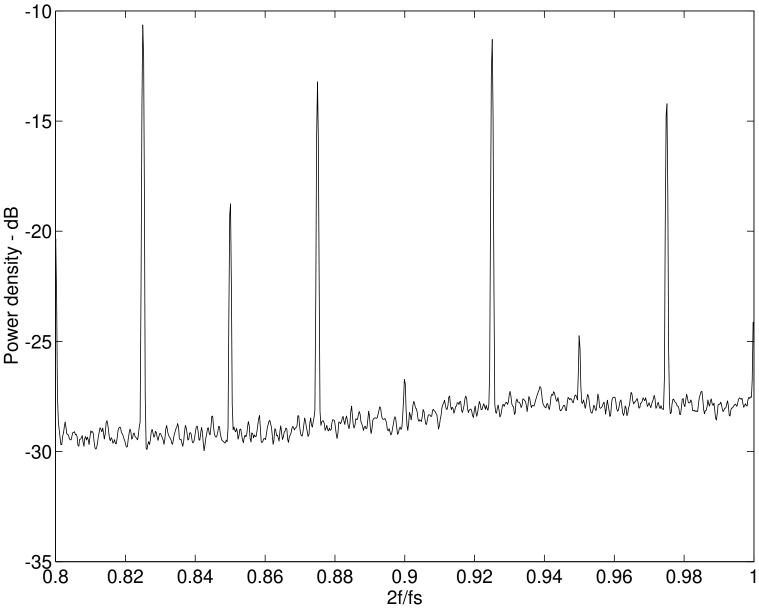

Figure C.6: Modulator power spectrum for sinusoidal input with frequency $0.025f_s/2$ and amplitude 0.2. Notice the side-bands at $2f/f_s = 0.975,\ 0.95,\ 0.925$ etc.

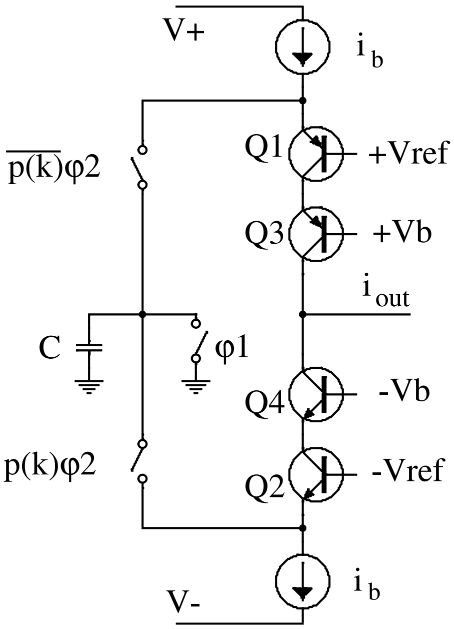
V+
i_b
Q1
+Vref
$\overline{p(k)}\varphi 2$
Q3
+Vb
i_out
C
φ1
Q4
-Vb
p(k)φ2
Q2
-Vref
V-
i_b

Figure C.7: The implemented pulse type one-bit DAC. The circuit is controlled by the modulator output $p(k)$ and two non-overlapping clocks, φ_1 and φ_2.

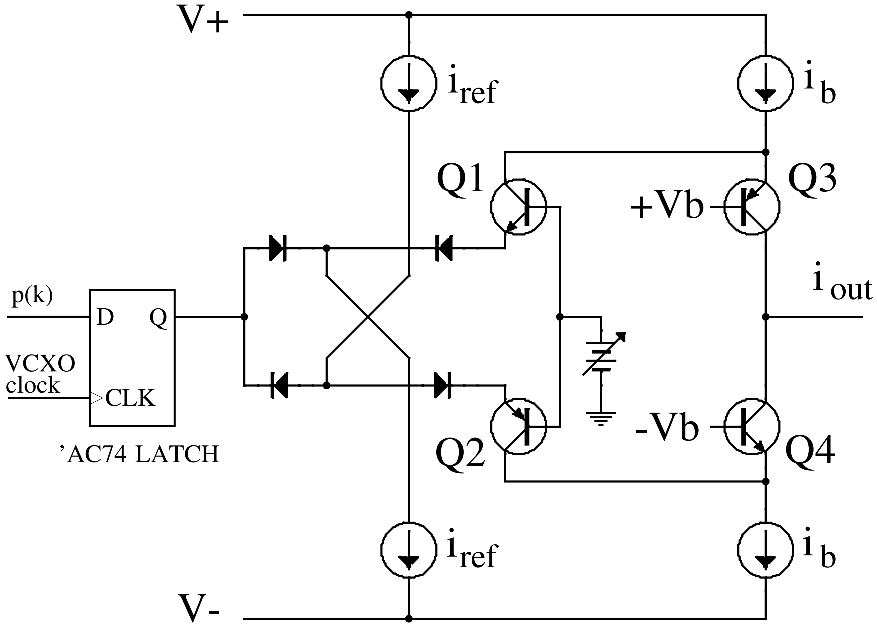
V+
V-
i_{ref}
i_{ref}
i_b
i_b
Q1
Q2
Q3
Q4
+Vb
-Vb
i_{out}
p(k)
D
Q
VCXO
clock
CLK
'AC74 LATCH

Figure C.8: The implemented NRZ current switch one-bit DAC. The output current is either $+i_{\text{ref}}$ or $-i_{\text{ref}}$ depending on the digital input code $p(k)$. The adjustable voltage source adjusts the skewness of the low to high and high to low transitions relative to the clock.

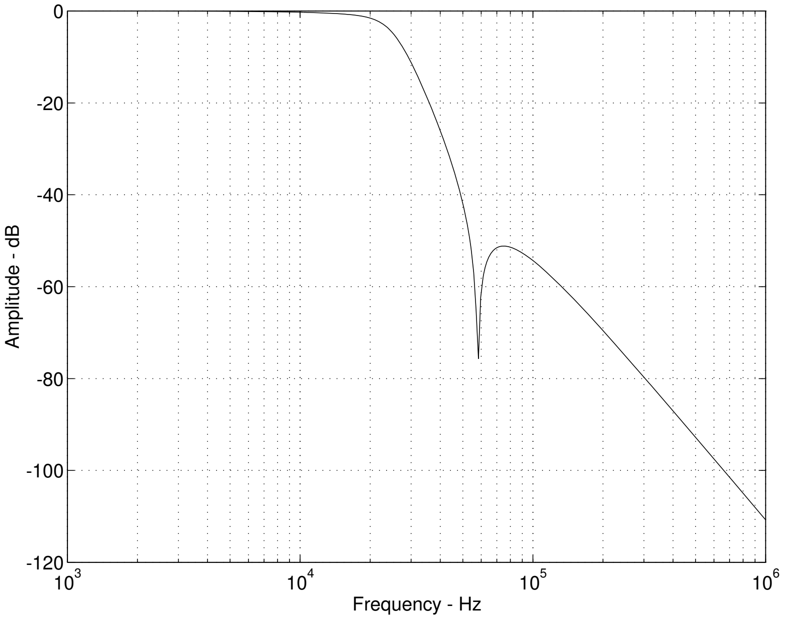

Figure C.9: Amplitude response for the analog post filters.

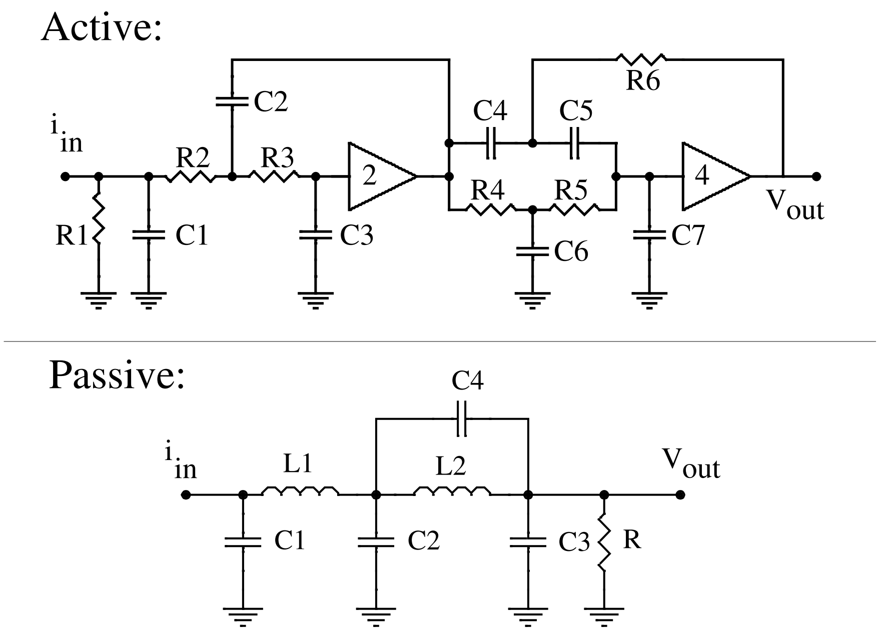
Active:
i_{in}
R1
C1
R2
C2
R3
C3
2
C4
C5
R4
R5
C6
R6
4
C7
V_{out}
Passive:
i_{in}
L1
C1
L2
C2
C4
C3
R
V_{out}

Figure C.10: The implemented passive and active analog post filters.

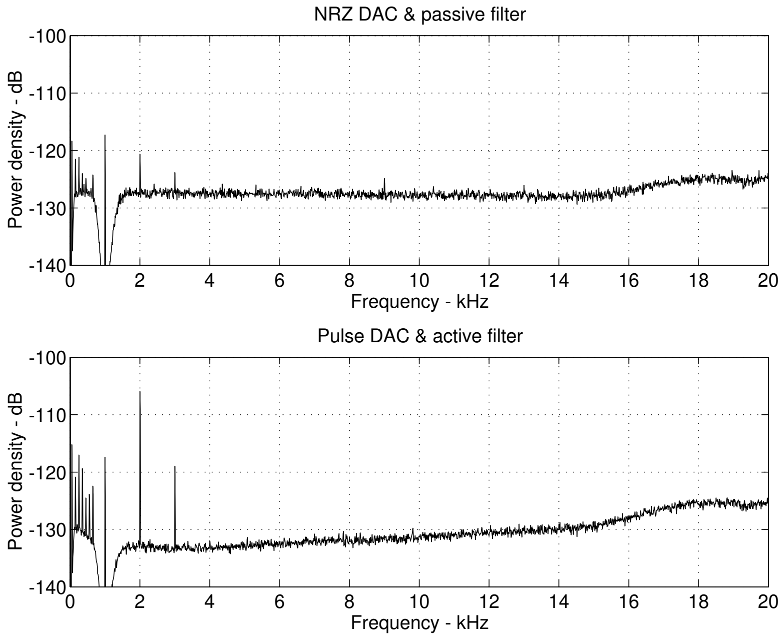

Figure C.11: Output spectra of the entire converter with 0dB 1kHz sinewave input quantized to 18 bit using triangular dither. The spectra are measured using a notch filter and 64 times averaged 16k FFTs at sample rate 44.1 kHz. Upper graph: NRZ DAC and passive post filter. Lower graph: Pulse DAC and active filter.

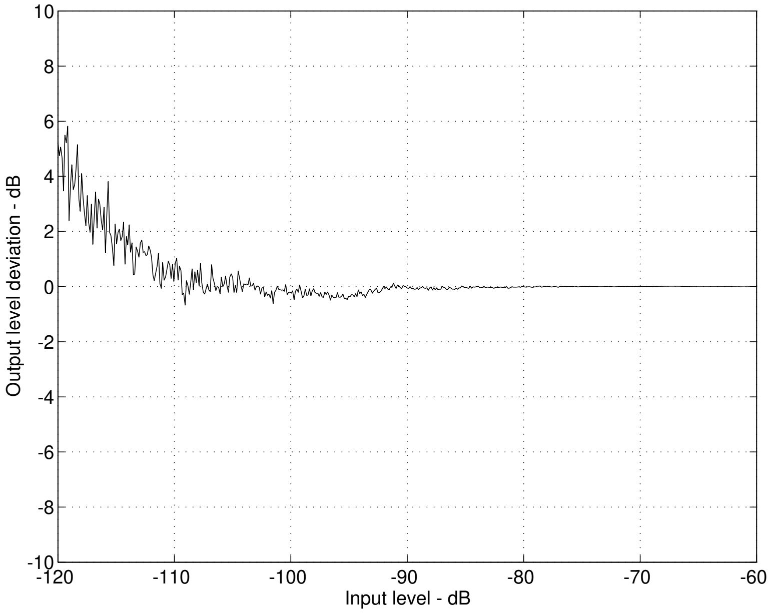

Figure C.12: Level linearity of the pulse type DAC and active filter with 18-bit triangular dithered sine input at 500 Hz.

Bibliography

[1] R. W. Adams *et. al.,* "Theory and practical implementation of a 5th-order Sigma-Delta A/D converter," *J. Audio Eng. Soc.,* vol. 39, pp. 515–528, July/Aug. 1991.

[2] R. W. Adams, *private communication*, Nov. 1992.

[3] B. P. Agrawal, K. Shenoi, "Design methodology for $\Sigma\Delta$M," *IEEE Trans. on Communication*, vol. 31, no. 3, pp. 360–369, March 1983.

[4] D. Anastassiou, "Error diffusion coding for A/D conversion," *IEEE Trans. on Circuits and Systems*, Vol. 36, No. 9, pp. 1175–1186, Sept. 1989.

[5] S. H. Ardalan and J. J. Paulos, "An analysis of nonlinear behavior in delta-sigma modulators," *IEEE Trans. Circuits Sys.*, vol. CAS–34, pp. 593–603, June 1987.

[6] T. Berger, *Rate distortion theory — a mathematical basis for data compression*, Prentice-Hall, 1971.

[7] G. Burra, K. S. Chao, "A high-speed high-resolution oversampled A/D converter," *IEEE Proc. ISCAS 1993*, pp. 1282–1285, May 1993.

[8] J. C. Candy, "A use of double integration in sigma-delta modulation," IEEE Trans. on Communications, vol. 33, no. 3, pp. 249–258, March 1985.

[9] J. C. Candy, G. C. Temes, eds., *Oversampled delta-sigma data converters. Theory, design and simulation*, IEEE Press, 1992.

[10] K. C.-H Chao, S. Nadeem, W. L. Lee and C. G. Sodini, "A higher order topology for interpolative modulators for oversampling A/D conversion," *IEEE Trans. Circuits. Sys.*, vol. CAS–37, pp. 309–318, Mar. 1990.

[11] L. O. Chua, T. Lin, "Chaos in digital filters," *IEEE Trans. Circuits and Systems*, vol. 35, no. 6, pp. 648–658, June 1988.

[12] W. Chou, R. M. Gray, "Dithering and its effect on sigma-delta and multistage sigma-delta modulation," *IEEE Trans. on Information Theory*, vol. 37, no. 3, pp. 500–513, May 1991.

[13] P. Cvitanović, "Chaos for cyclists," printed in: *Noise and chaos in nonlinear dynamical systems,* F. Moss, L. Lugiato and W. Schleich, eds. Cambridge Univ. Press, 1990, pp. 270–288.

[14] C. Dunn and M. O. J. Hawksford, "Is the AESEBU/SPDIF Digital Audio Interface Flawed?," *Proc. 93rd AES Convention*, Preprint # 3360, Oct. 1992.

[15] O. Feely, L. O. Chua, "Nonlinear dynamics of a class of analog-to-digital converters," *Int. journal of Bifurcation and Chaos*, Vol. 2, No. 2, pp. 325–340, 1992

[16] O. Feely, L. O. Chua, "The effect of integrator leak in Σ-Δ modulation," *IEEE Trans. on Circuits and Systems*, Vol. 38, No. 11, pp. 1293–1305, Nov. 1991.

[17] I. Fujimori, K. Hamashita, E. J. Swanson, "A fifth-order delta-sigma modulator with 110 dB audio-band dynamic range," *Proc. 93rd AES Convention*, Preprint # 3415, Oct. 1992.

[18] R. A. Gabel, R. A. Roberts, *Signals and linear systems*, Wiley, 1987.

[19] A. Gersho, R. M. Gray, *Vector quantization and signal compression*, Kluwer Academic Publishers, 1992.

[20] M. Gerzon, P. G. Craven, "Optimal noise shaping and dither of signals," *Proc. 87th AES Convention*, Preprint # 2822, Oct. 1989.

[21] A. Grace, *Optimization TOOLBOX – user's guide*, The Mathworks Inc, 1992.

[22] C. Grebogi, E. Ott and J. A. Yorke, "Chaotic attractors in crisis," *Phys. Rev. Lett.*, vol. 48, pp. 1507–1510, 1982.

[23] S. R. Green, S. Harris and B. Wilson, "An 18-bit Delta-Sigma D/A Processor System Achieving Full-Scale THD+N>100dB," *Proc. 93rd AES Convention*, Preprint # 3416, Oct. 1992.

[24] S. Harris, "How to achieve optimum performance from Delta-Sigma A/D and D/A converters," *J. Audio Eng. Soc.*, vol. 41, no. 10, pp. 782–789, Oct. 1992.

[25] S. Hein and A. Zakhor, *Sigma-Delta modulators – nonlinear decoding algorithms and stability analysis*, Kluwer Academic Publishers, 1993.

[26] S. Hein, "Exploiting chaos to suppress spurious tones in general double-loop Σ-Δ modulators," *IEEE Trans. on Circuits and Systems –II*, Vol. 40, No. 10, pp. 651–659, Oct. 1993.

[27] J. Isoaho, H. Tenhunen, J. Heikkilä and L. Lipasti, "High Resolution DAC Design based on FPGAs," *Proc. Oxford 1991 Int. Workshop on FPGA Logic and Applications*, pp. 343–352, Sep. 1991.

[28] T. Karema, T. Ritoniemi and H. Tenhunen, "Intermodulation in Sigma-Delta D/A Converters," *Proc. IEEE ISCAS'91*, pp. 1625–1628.

[29] J. G. Kenney, L. R. Carley, "Design of multibit noise-shaping data converters," *Analog Integrated Circuits and Signal Processing*, 3, pp. 99–112, 1993.

[30] C. Knudsen, *Aspects of noninvertible dynamics and chaos,* Ph.D. dissertation, Physics Department, The Technical University of Denmark, Jan. 1994.

[31] B. M. J. Kup, E. C. Dijkmans, P. J. A. Naus, J. Sneep, "A bitstream digital-to-analog converter with 18-b resolution," *IEEE Journal of Solid-state Circuits*, vol. 26, No. 12, pp. 1757–1763, Dec. 1991.

[32] J. L. LaMay, H. T. Bogard, "How to obtain maximum practical performance from state-of-the art delta-sigma analog-to-digital converters," *IEEE Trans. on Instrumentation and Measurements*, vol. 41, no. 6, pp. 861–867, December 1992.

[33] R. C. Ledzius, J. Irwin, "The basis and architecture for the reduction of tones in a sigma-delta DAC," *IEEE Trans. on Circuits and Systems –II*, Vol. 40, No. 7, pp. 429–439, July 1993.

[34] W. L. Lee & C. Sodini, "A topology for higher order interpolative coders," *IEEE Proc. ISCAS'87*, pp. 459–462.

[35] J. N. Little and L. Shure, *Signal processing TOOLBOX - user's guide*, The MathWorks, Inc., 1993.

[36] E. Meitner and R. Gendron, "Time Distortions Within Digital Audio Equipment Due to Integrated Circuit Logic Induced Modulation Products," *Proc. 91st AES Convention*, Preprint # 3105, Oct. 1991.

[37] M. Motamed, A. Zakhor and S. Sanders, "Tones, saturation and SNR in double loop $\Sigma - \Delta$ modulators," *Proc. IEEE ISCAS'93,* pp. 1345–1348, Chicago, USA, May 1993.

[38] P. J. A. Naus *et. al.*, "A CMOS Stereo 16-bit D/A Converter for Digital Audio," *IEEE J. Solid-State Circuits*, vol. SC–22, pp. 390–394, June 1987.

[39] Truong-Thao Nguyen, *Deterministic analysis of oversampled A/D conversion and sigma-delta modulation, and decoding improvements using consisten estimates*, Ph.D. dissertation, Center for Telecommunications Research, Columbia University, New York, 1993.

[40] S. R. Norsworthy and D. A. Rich, "Idle channel tones and dithering in Delta-Sigma modulators," *Proc. 95th AES Convention*, preprint # 3711, Oct. 1993.

[41] S. R. Norsworthy, "Optimal nonrecursive noise shaping filters for oversampling data converters," *IEEE Proc. ISCAS'93,* pp. 1353–1356, Chicago, USA, May 1993.

[42] S. R. Norsworthy, "Effective dithering of sigma-delta modulators," *IEEE Proc. ISCAS 1992*, pp. 1304–1307, May 1992.

[43] T. Okamoto, Y. Maruyama, A. Yukawa, "A stable high-order delta-sigma modulator with an FIR spectrum distributor," *IEEE Journal of Solid-State Circuits*, vol. 28, no. 7, pp. 730–735, July 1993.

[44] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing*, Prentice Hall, 1989.

[45] S. J. Park, R. M. Gray, "Sigma-Delta modulation with leaky integration and constant input," *IEEE Trans. on Information Theory*, Vol. 38, No. 4, pp. 1512–1533, Sep. 1991.

[46] T. S. Parker, L. O. Chua, *Practical numerical algorithms for chaotic systems*, Springer Verlag, 1989.

[47] S. C. Pinault, P. V. Lopresti, "On the behavior of the double-loop sigma-delta modulator," *IEEE Trans. on Circuits and Systems –II*, Vol. 40, No. 8, pp. 467–479, Aug. 1993.

[48] L. Risbo, "FPGA based 32 times oversampling 8th order Sigma-Delta audio DAC," *Proc. 96th AES Convention*, preprint # 3808, Feb. 1994, Reprinted in Appendix C.

[49] L. Risbo, *Synthesis and simulations of stable high-order Sigma-Delta modulators*, M.sc. Thesis (in Danish), Electronics Institute, Technical University of Denmark, Aug. 1991.

[50] L. Risbo, *Stability analysis of a class of chaotic second-order* Σ-Δ *modulators*, Report prepared in course 1751 (in Danish), MIDIT, Technical University of Denmark, June 1993.

[51] L. Risbo, "Stability predictions for high-order Sigma-Delta modulators based on quasilinear modeling," *IEEE Proc. ISCAS'94,* pp. 5.361–5.364, London, UK, May 1994.

[52] L. Risbo, "Improved stability and performance from Σ-Δ modulators using one-bit vector quantization," *IEEE Proc. ISCAS'93,* pp. 1365–1368, Chicago, USA, May 1993, Reprinted in Appendix B.

[53] L. Risbo, "On the design of tone free Σ-Δ modulators," Preliminary manuscript, march 1994, Accepted for *IEEE Trans. on Circuits and Systems –II*, Reprinted in Appendix A.

[54] L. Risbo, *This Dissertation.*

[55] T. Ritoniemi, T. Karema and H. Tenhunen, "Design of Stable High Order Sigma-Delta Modulators," *Proc. IEEE ISCAS'90*, pp. 3267–3270, May 1990.

[56] T. Ritoniemi, T. Karema, H. Tenhunen, "Modelling and performance estimation of sigma-delta modulators," *IEEE Proc. ISCAS 1991*, pp. 2705–2708, May 1991.

[57] R. Schreier, "Destabilizing limit cycles in Delta-Sigma modulators with chaos," *IEEE Proc. ISCAS'93,* pp. 1369–1372, Chicago, USA, May 1993.

[58] R. Schreier, "An empirical study of high-order single-bit delta-sigma modulators," *IEEE Trans. Circuits and Syst.—II: Analog and Digital Signal Processing*, vol. 40, no. 8, pp. 461–466, Aug. 1993.

[59] R. Schreier, Y. Yang, "Stability tests for single-bit sigma-delta modulators with second-order FIR noise transfer functions," *IEEE Proc. ISCAS 1992*, pp. 1316–1319, May 1992.

[60] R. Schreier, M. Snelgrove, "Stability in a general Σ-Δ modulator," *IEEE Proc. ICASP 1991*, vol. 3 pp. 1769–1772, May 1991.

[61] N. S. Sooch, J. W. Scott, T. Tanaka, T. Sugimoto and C. Kubomura, "18-bit Stereo D/A Converter with Integrated Digital and Analog Filters," *Proc. 91st AES Convention*, Preprint # 3113, Oct. 1991.

[62] H. A. Spang, P. M. Schultheiss, "Reduction of quantization noise by use of feedback," *IRE Trans. on Communications Systems*,pp. 285–317, Dec. 1962.

[63] E. F. Stikvoort, "Some remarks on the stability and performance of the noise shaper or sigma-delta modulator," IEEE Trans. on Communications, vol. 36, no. 10, pp. 1157–1162, Oct. 1988.

[64] J. T. Stonick, S. H. Ardalan, J. K. Townsend, "An improved analysis of sigma-delta modulators with bandlimited Gaussian input," *IEEE Proc. ICASSP'90*, pp. 1691–1694, 1990.

[65] J. T. Stonick, J. L. Rulla, S. H. Ardalan, J. K. Townsend, "A new architecture for second order $\Sigma\Delta$ Modulation," *IEEE Proc. ISCAS'90*, pp. 360–363, 1990.

[66] S. K. Tewksbury, R. W. Hallock, "Oversampled, linear predictive and noise-shaping coders for order $N > 1$," *IEEE Trans. Circuits and Systems*, vol. 25, no. 7, pp. 436–447, July 1978.

[67] H. Wang, "A geometric view of $\Sigma\Delta$ modulation," *IEEE Trans. on Circuits and Systems –II*, Vol. 39, No. 6, pp. 402–405, June 1992.

[68] H. Wang, *A study of sigma-delta modulations as dynamical systems*, Ph.D. dissertation, Center for Telecommunications Research, Columbia University, New York, 1993.

[69] C. Wolff, L. R. Carley, "Modeling the quantizer in higher-order delta-sigma modulators," *IEEE Proc. ISCAS'88,* pp. 2335–2338, Helsinki, Finland, June 1988.

Dansk Resumé

Σ-Δ modulation anvendes i udstrakt grad i A/D og D/A omsættere til krævende audio formål. Denne modulationsform indkoder et tidsdiskret signal med høj opløsing om til et to-niveau (dvs. et-bit) signal med en samplingfrekvens væsentligt højere end krævet af sampling sætningen (oversampling). Herved opnås en høj amplitudeopløsning på bekostning af lavere opløsning i tid. Denne virkning er bedre kendt som "noise-shaping", dvs. fejlen ved et-bit signalet er fjernet fra et lavfrekvent bånd og koncentreret ved høje frekvenser.

En Σ-Δ modulator består af en et-bit kvantisator indeholdt i en modkoblingssløjfe med et lineært filter. Modulatoren er dermed et ulineært dynamisk system. For nogle typer filtre bliver modulatoren tillige kaotisk. En alvorlig ulempe ved Σ-Δ modulatoren er muligheden for ustabilitet. Dette er specielt et praktisk problem for modulatorer med høj filterorden.

Stabilitetsproblemet er hovedemnet for denne afhandling. Problemet behandles i første omgang fra et system-dynamisk synspunkt. En af konlusionerne er, at stabiliteten af kaotiske modulatorer forsvinder ved en såkaldt *boundary crisis*, der opstår, når attraktoren kolliderer med sit attraktionsbasin. Graden af ustabilitet kan beskrives ved den såkaldte escape-rate.

Stabilitet kan også analyseres vha. symbolsk dynamik. Konlusionen er, at stabiliteten er tæt knyttet til antallet af mulige grænsecykler.

Nogle modulatorer bliver karakteriseret som værende upålidelige. Dette vil sige, at ustabiliteten ikke indtræffer ved et veldefineret punkt. Sådanne modulatorer kan være tilsyneladende stabile ved kortere simuleringer. Ustabiliteten viser sig kun ved enten meget lange simuleringer eller ved realtidsimplementeringer.

Afhandlingen præsenterer værktøjer til optimering af både signal/støj forhold og pålidelighed. Metoden kombinerer eksakt analyse med approksimativ lineariseret analyse. Metoden har været anvendt til design af en 32 gange oversamplende 8. ordens Σ-Δ D/A-omsætter til audio brug. Denne omsætter er implemeteret vha. Field Programmable Gate Array kredse.

Afhandlingen behandler også metoder til undertrykkelse af de meget kraftige toner omkring den halve sampling frekvens. Brugen af kaos og dither-støj viser sig at være nogenlunde lige effektive og ækvivalente på flere måder. Med hensyn til brug af kaotiske modulatorer, demonstreres en metode, hvor støjoverføringsfunktionen forlænges med et alpasled.

Til slut præsenteres en ny klasse af modulatorer, der anvender et-bit vektor kvantisering. Det vises, at introduktionen af en vektorkvantisator kan forbedre stabiliteten betydeligt. Samtidig kan styrken af tonerne ved den halve samplingfrekvens reduceres.