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:

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

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

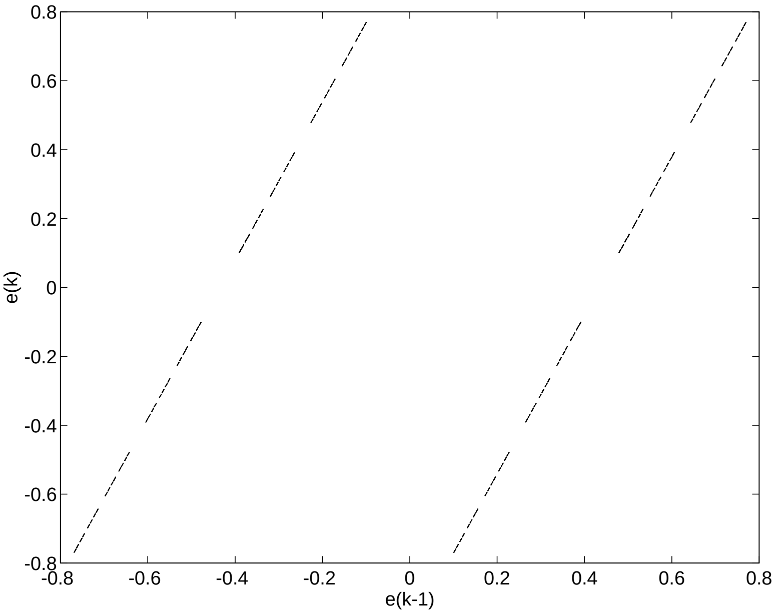

Figure 4.1: Limit cycles of period up to 12 for the modulator cf. Eq. (4.10) with $a = 2.3$.

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 $\square$

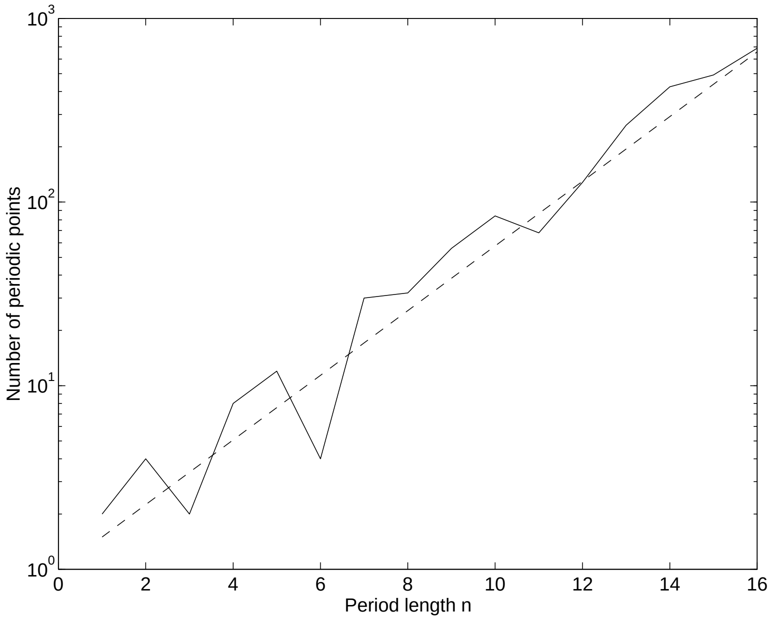

Figure 4.2: Number of periodic points l_n with period up to 16 versus the period length n for the modulator cf. Eq. (3.5) with with $h_1 = 0.8$, $h_2 = -1.5$ and zero input. The dashed line corresponds to the function $\Lambda^n = |h_2|^n$.

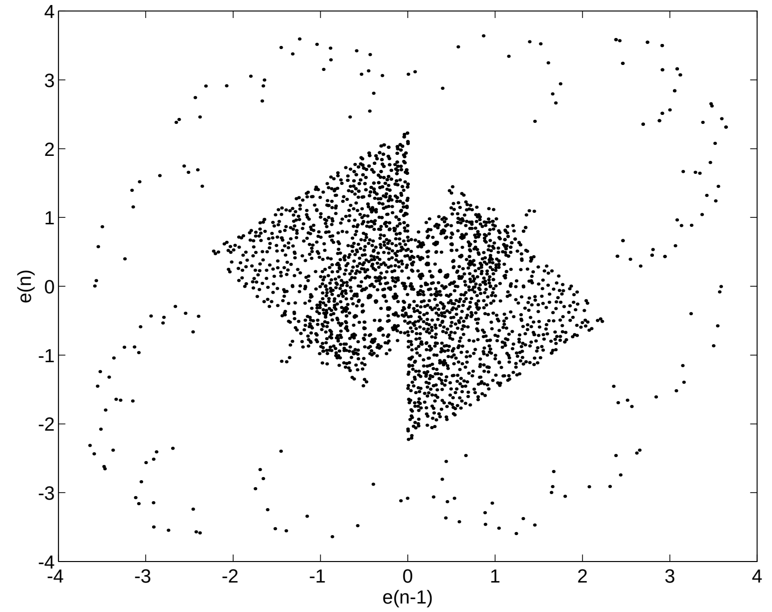

Figure 4.3: Limit cycles with period up to 16 for the modulator cf. Eq. (3.5) with $h_1 = 0.8$, $h_2 = -1.5$ and zero input.

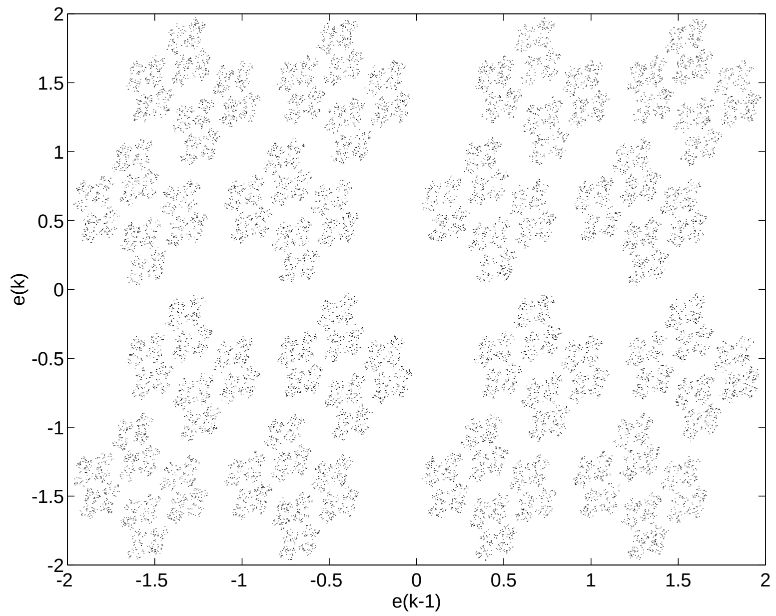

Figure 4.4: Reverse time simulation of the system cf. Eq. (4.14) with zero input.

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 $W = \{\boldsymbol{x} \in \mathbb{R}^2 | \|\boldsymbol{x}\|_\infty < 1\}$ 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:

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

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}.

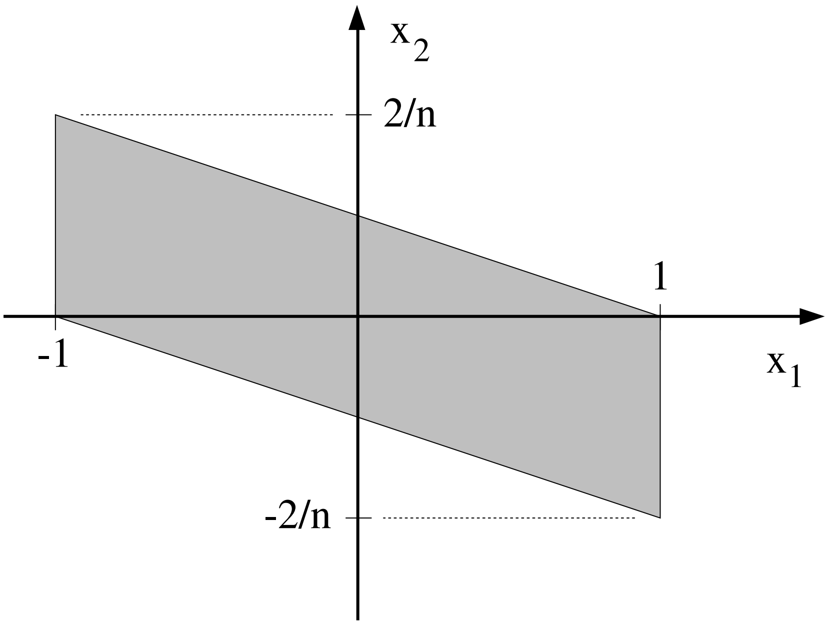

Figure 4.5: The hatched area is the set of survivors Ω_n for the nonhyperbolic system cf. Eq. (4.18)

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:

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

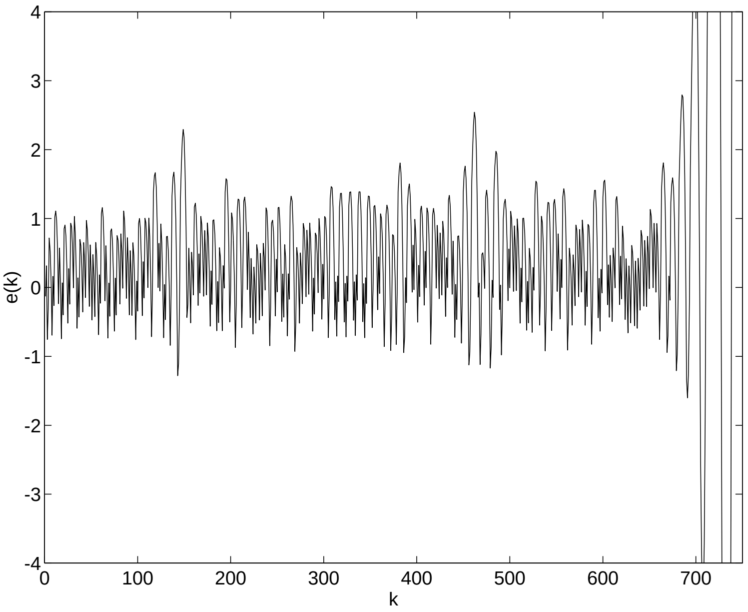

Figure 5.1: Transient leading to an overload cycle for the modulator cf. Eq. (5.1) with $m = 0.7$, $v = 0.09$ and constant input 0.4. Quantizer input $e(k)$ is plotted versus k

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 $\square$

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.

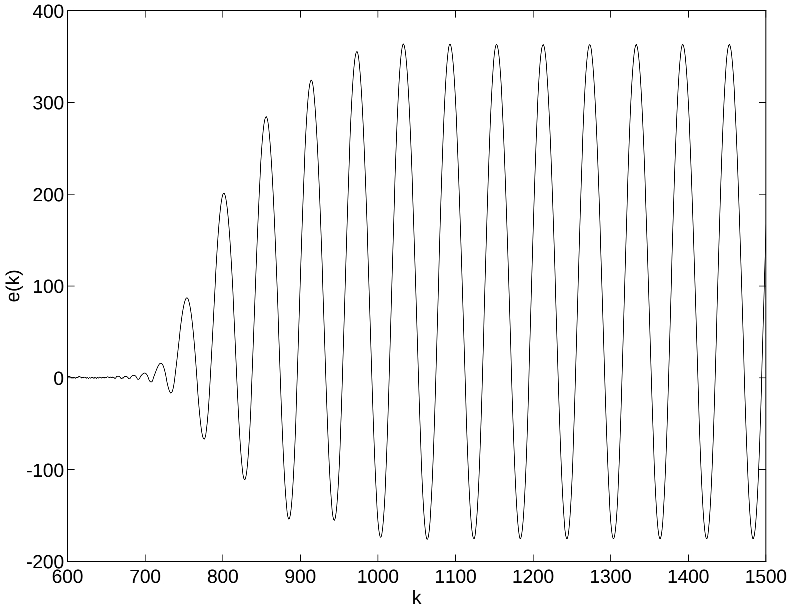

Figure 5.2: The onset of an overload cycle for the modulator cf. Eq. (5.1) with $m = 0.7$, $v = 0.09$ and constant input 0.4. The plot is the same as in Fig. 5.1 but with different axes.

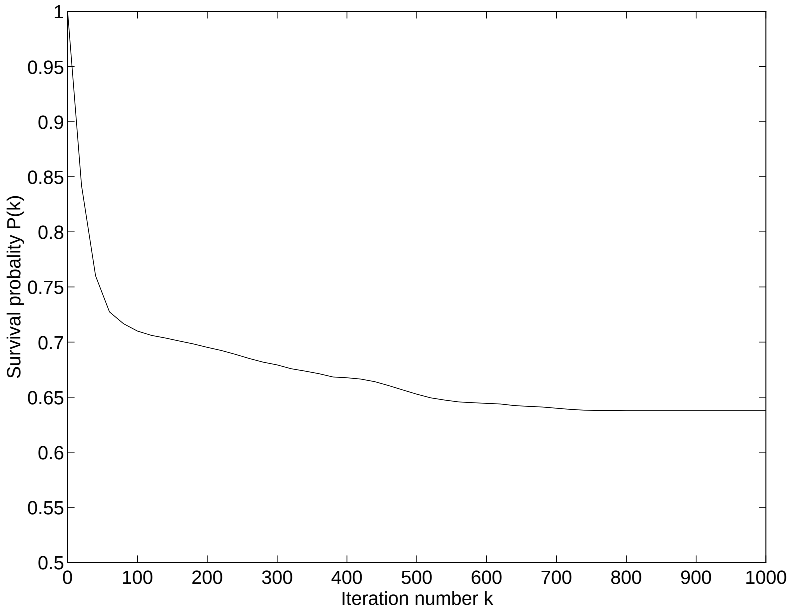

Figure 5.3: Survival plot for the modulator cf. Eq. (5.1) with $m = 0.7$, $v = 0.09$ and constant input 0.4.

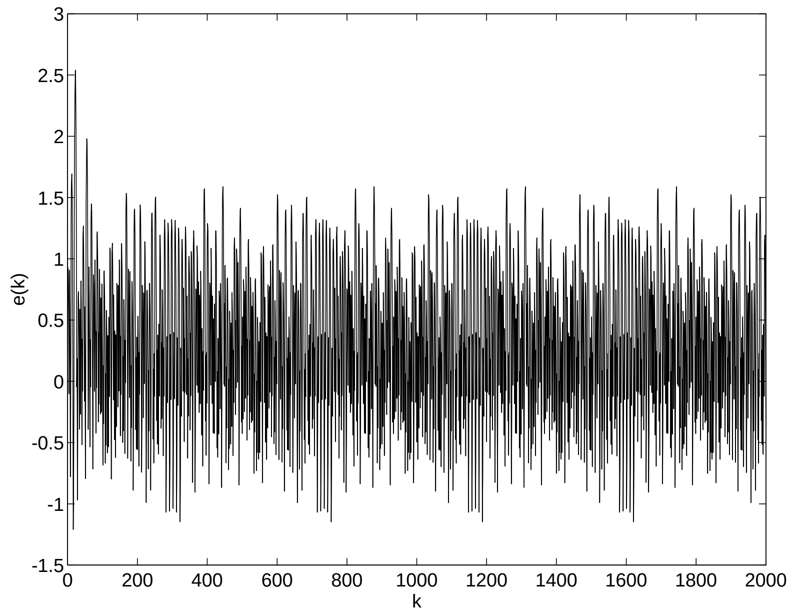

Figure 5.4: Example of a granular limit cycle.

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

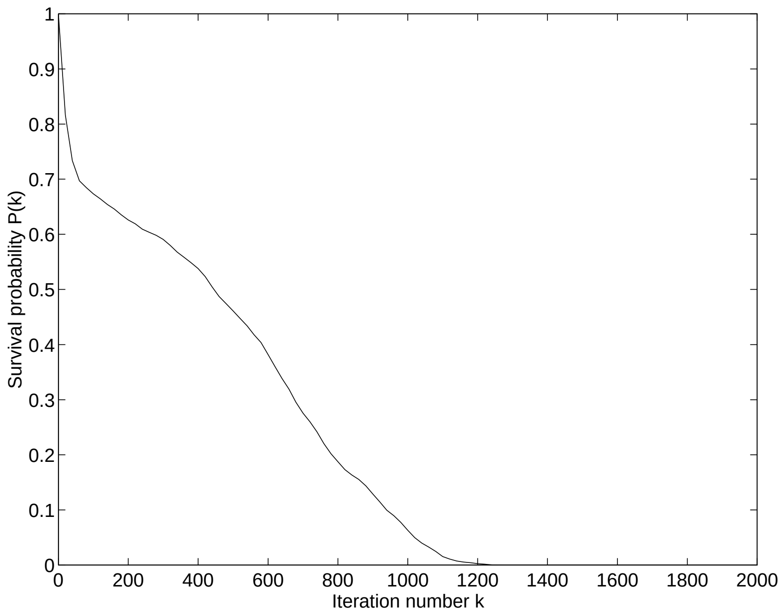

Figure 5.5: Survival plot for the modulator cf. Eq. (5.1) with $m = 0.7$, $v = 0.09$ and constant input 0.5.

$z = 1$, 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

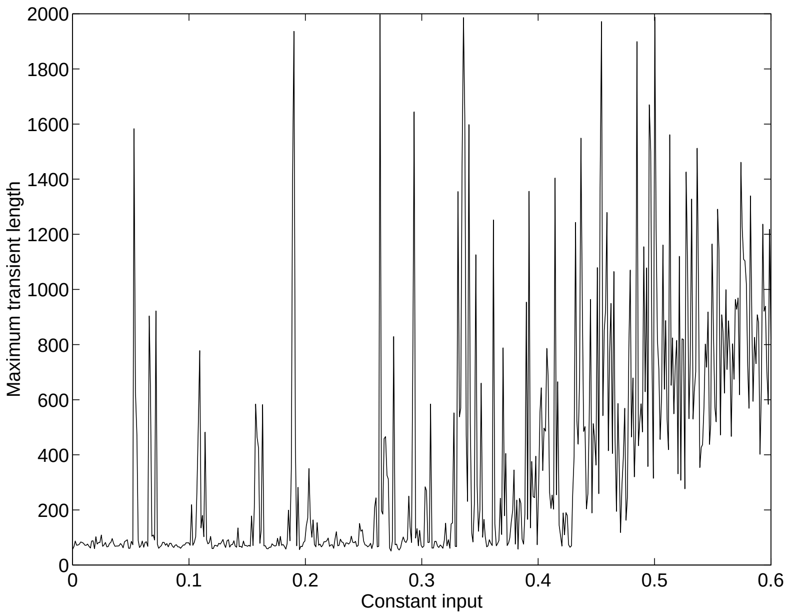

Figure 5.6: Maximum length of unstable transients versus constant input for the third order modulator cf. Eq. (5.1) with $m = 0.7$ and $v = 0.09$.

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

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.

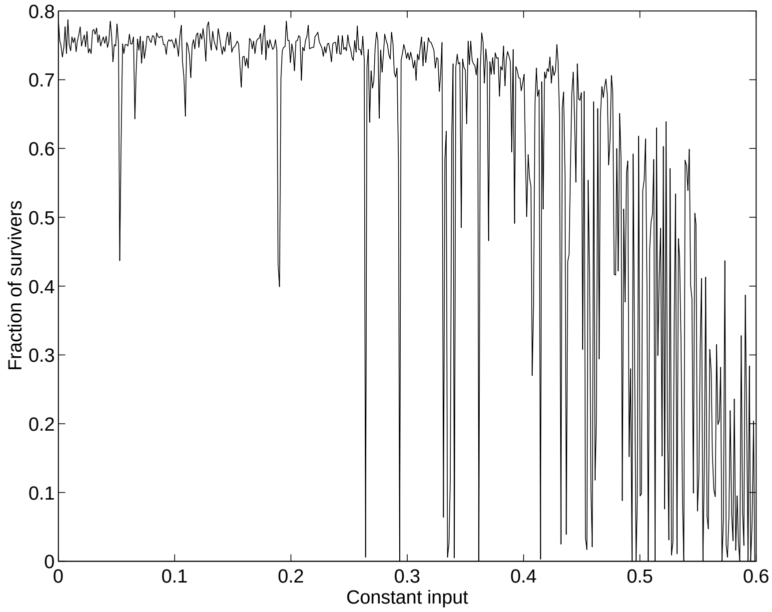

Figure 5.7: The fraction of survivors (i.e. initial conditions leading to granular cycles) for the modulator cf. Eq. (5.1) with $m = 0.7$, $v = 0.09$.

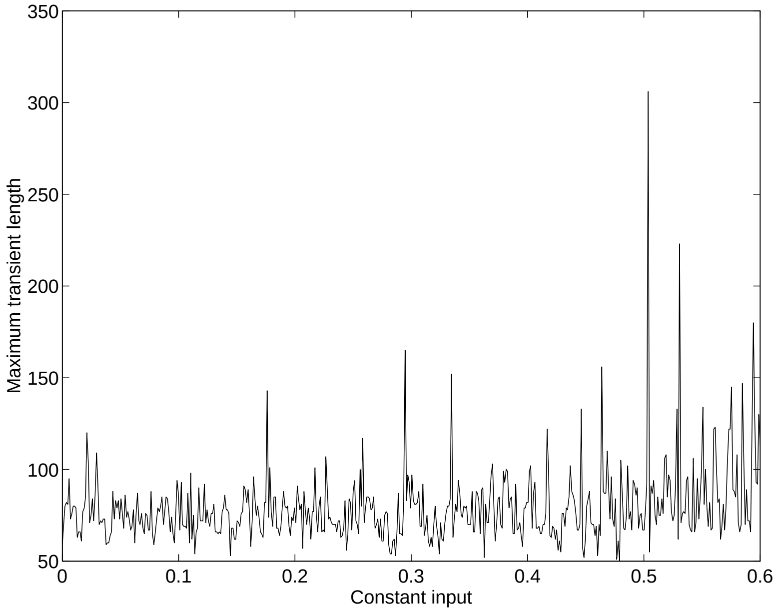

Figure 5.8: Maximum length of unstable transients versus constant input for the third order modulator cf. Eq. (5.1) with $m = 0.75$ and $v = 0.09$.

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

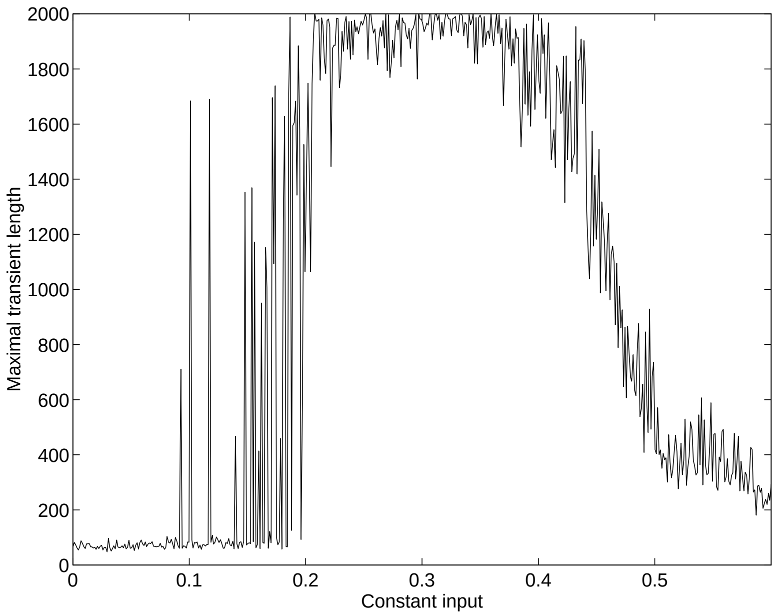

Figure 5.9: Maximum length of unstable transients versus constant input for the third order modulator cf. Eq. (5.1). The poles are mirrored outside the unit circle in order to get a chaotic system and the parameters $m = 0.9$ and $v = 0.11$ were used.

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

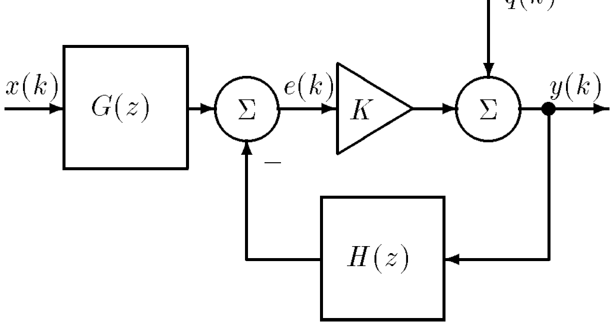

Figure 6.1: Generic Σ-Δ modulator with linearized quantizer model

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 $\mathrm{NTF}(z)$ *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:

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

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

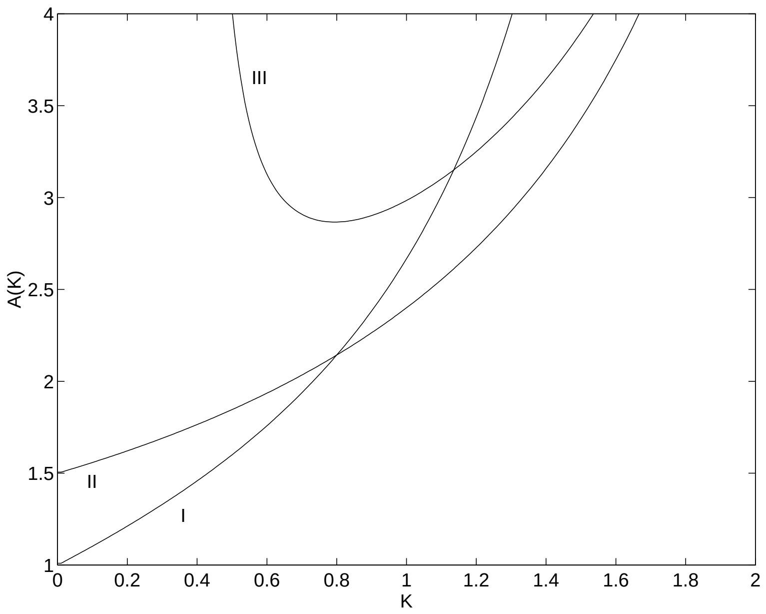

Figure 6.2: A(K)-curves for the type I, II and III loop filter examples in Table 6.1.

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 $m_y = \mathrm{E}\{\mathrm{sgn}(e(k))\} = \Pr\{e(k) > 0\} - \Pr\{e(k) < 0\}$ as parameter, i.e., the quantizer input is DC-shifted such that the mean output is m_y. The result for

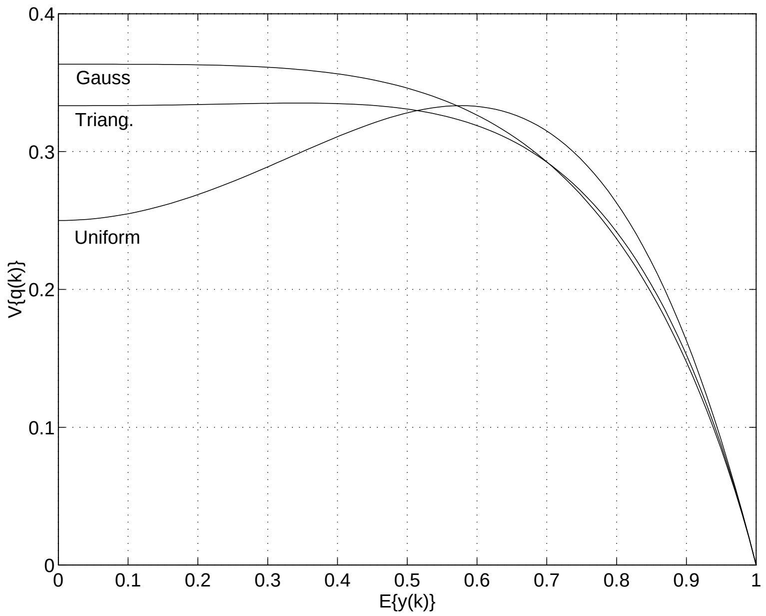

Figure 6.3: σ_q^2 versus $m_y > 0$ for three types of probability distributions, cf. Eq. (6.15), (6.16) and (6.17).

Gaussian pdf is [5]:

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

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.

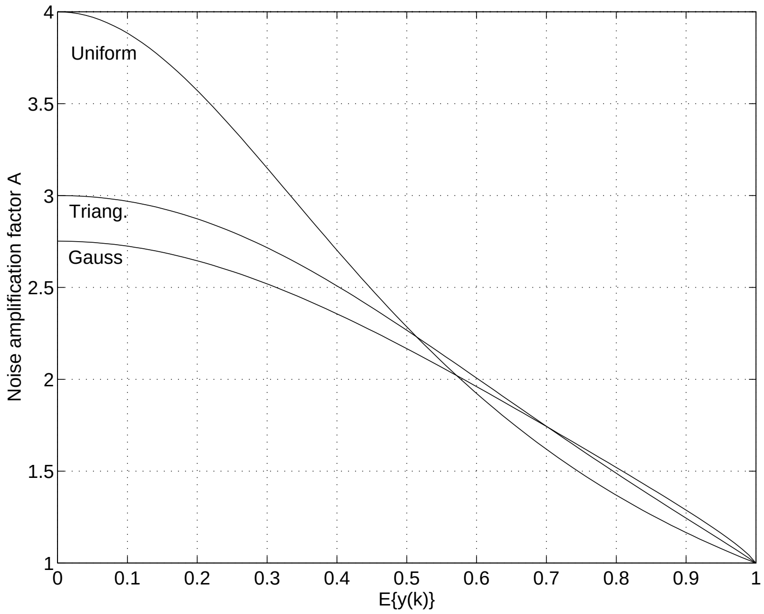

Figure 6.4: Noise amplification factor A versus m_y for these three types of probability distributions, cf. Eq. (6.15), (6.16) and (6.17).

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 $\mathrm{A}(m_y)$ 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 $\mathrm{A}(K)$-curve. The knowledge of the location on the $\mathrm{A}(K)$-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

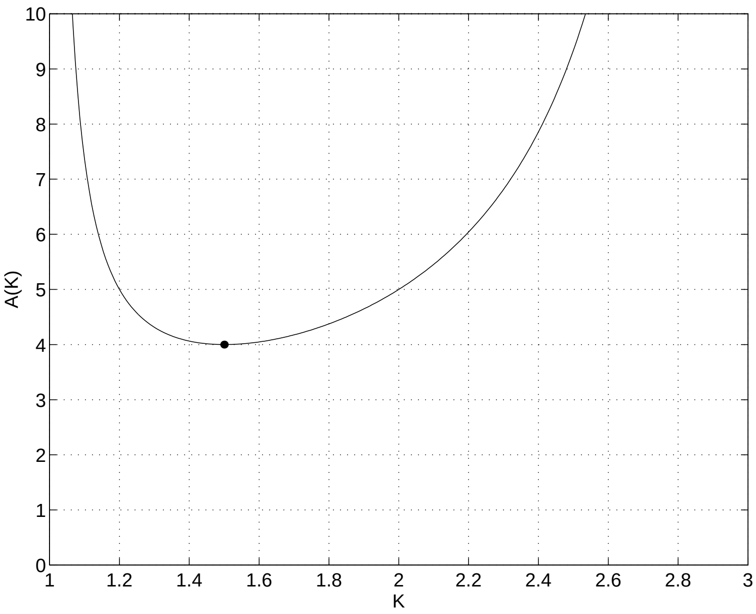

Figure 6.5: Noise amplification factor A versus K for the first order feedback filter $H(z) = z^{-1}/(1-2z^{-1})$. The A_{min}-point is shown with a dot.

reaches always equilibrium thanks to the increasing $\mathrm{A}(K)$-curve. Real type II systems have in fact the property that $\max(|e(k)|) \to \infty$ as $m_y \to 1$, [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 $\mathrm{A}(K)$-curve with $\mathrm{A}_{min} =$

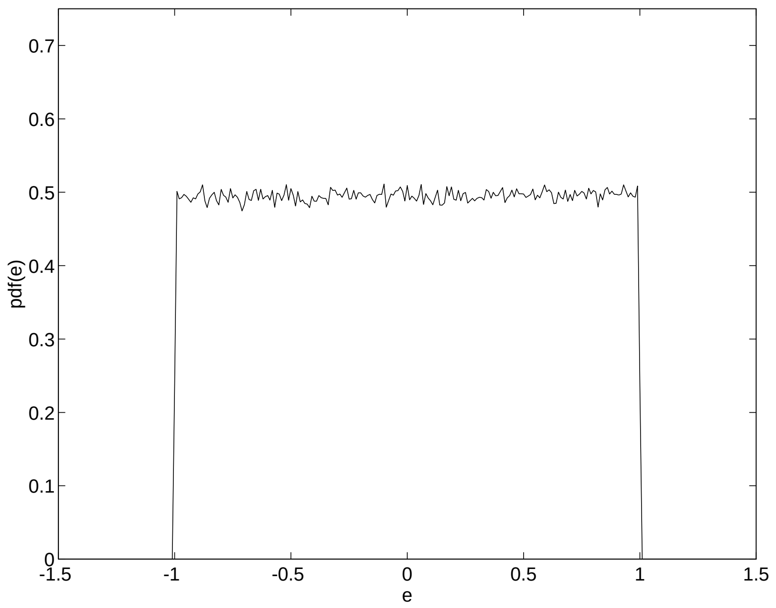

Figure 6.6: Estimated quantizer input pdf for the first order modulator with $H(z) = z^{-1}/(1-2z^{-1})$. Obtained using 200 uniformly spaced bins from -1 to 1 and 1 mill. samples.

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 $\mathsf{fb} = 1/64$. 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.

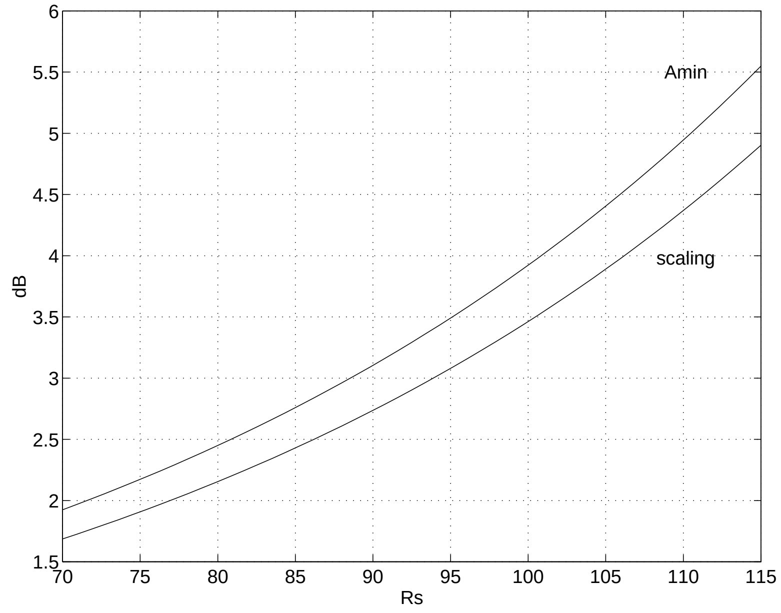

Figure 6.7: Scaling and A_{min} versus Rs (in dB) for fifth order Chebychev II filters with $\mathsf{fb} = 1/64$.

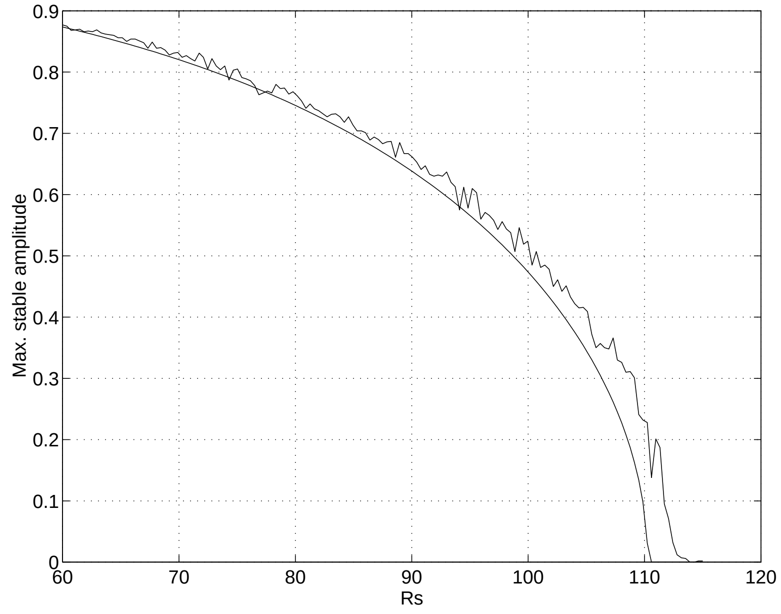

Figure 6.8: Simulated maximum stable input amplitude versus Rs for fifth order Chebychev II filters with $\mathsf{fb} = 1/64$. The smooth curve is the Gaussian A_{min}-criterion of Eq. (6.20).

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):

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

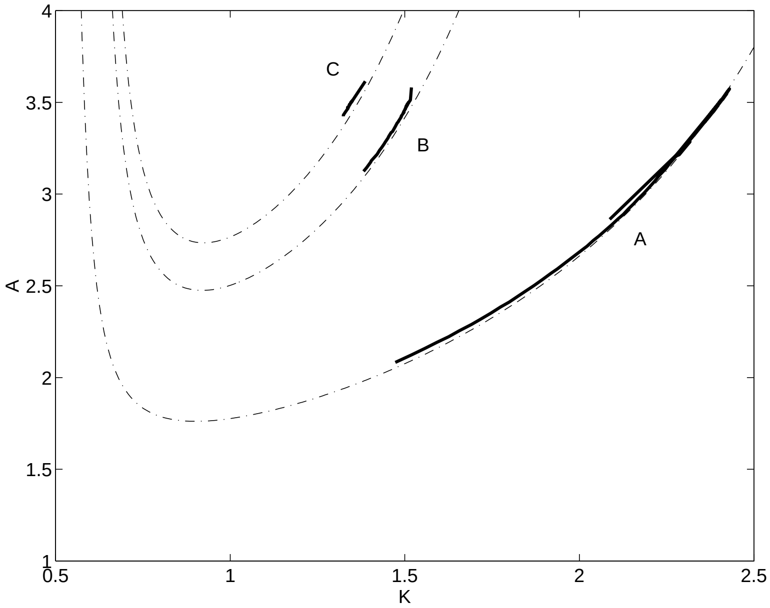

Figure 6.9: Simulated and theoretical (dashed lines) A(K)-curves for filter A, B and C, cf. Table 6.2.

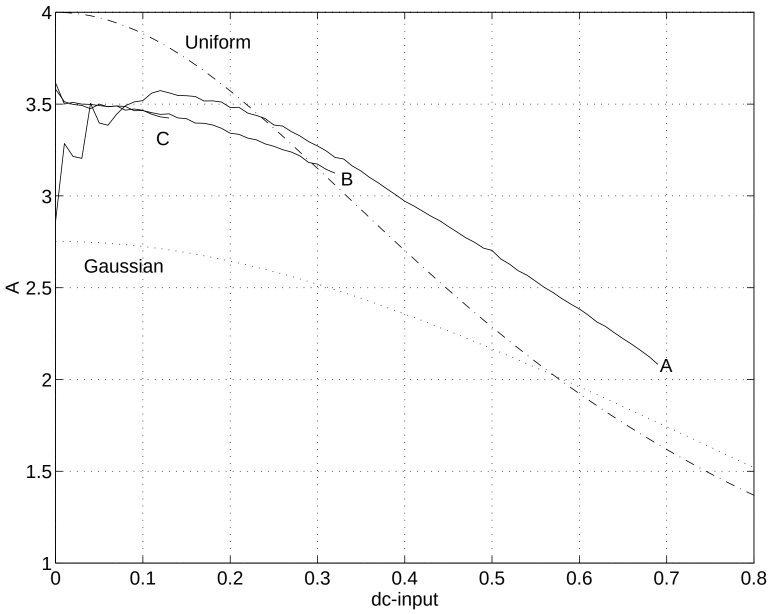

Figure 6.10: Noise amplification factor A versus dc-input found by simulations of filter A, B and C, cf. Table 6.2. Theoretical curves for Gaussian and uniform quantizer input pdf are shown with dashed lines.

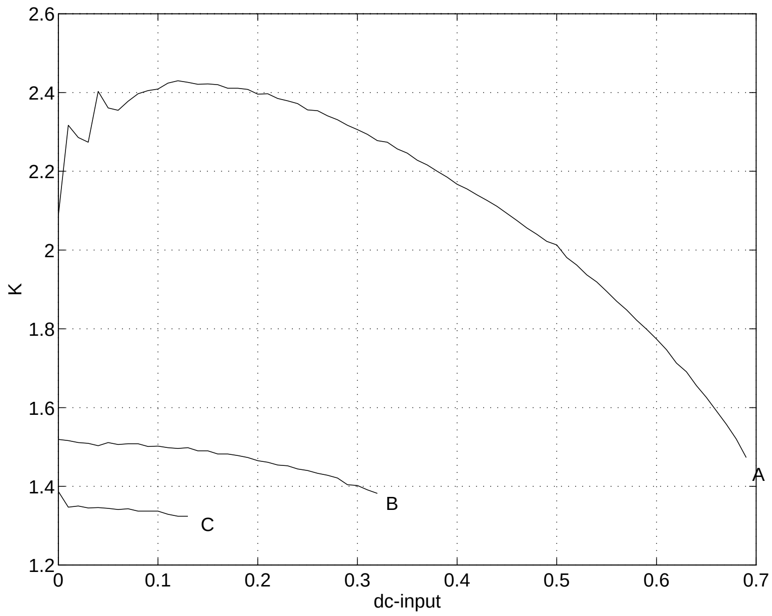

Figure 6.11: Quantizer gain K versus dc-input found by simulations of filter A, B and C, cf. Table 6.2.

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 pass-band 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

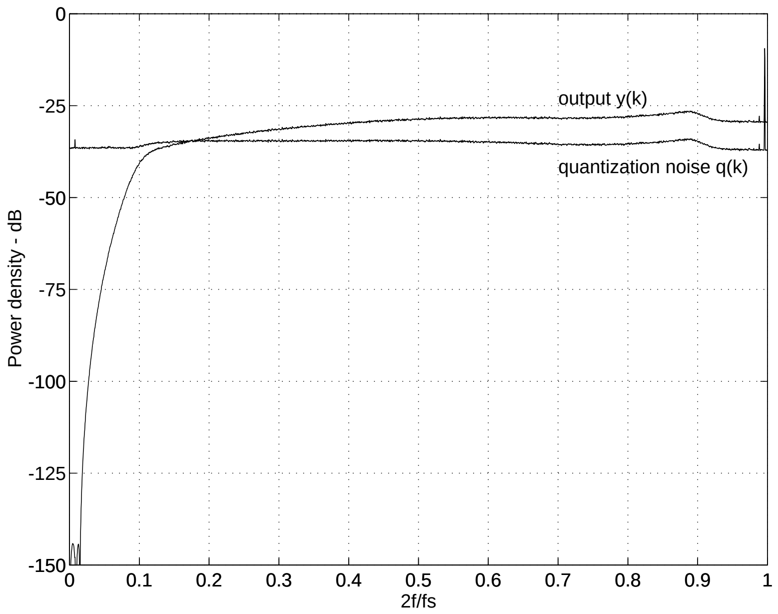

Figure 6.12: Power spectra of the modulator output $y(k)$ and the quantization noise $q(k)$ for a 10 million time step simulation of filter B, cf. Table 6.2 with a dc-input of 1/256. Note the frequency scale normalized in respect to half the sample rate.

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 $\mathrm{q}(e)$ is non-invertible in the

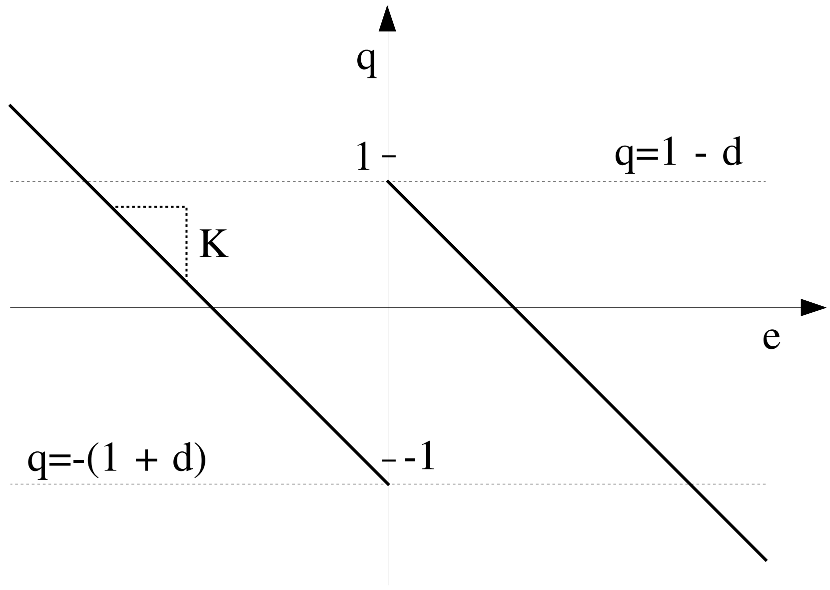

Figure 6.13: Quantization noise q versus quantizer input e for the quasilinear quantizer model, cf. Eq. (6.23) with displacement factor d> 0.

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

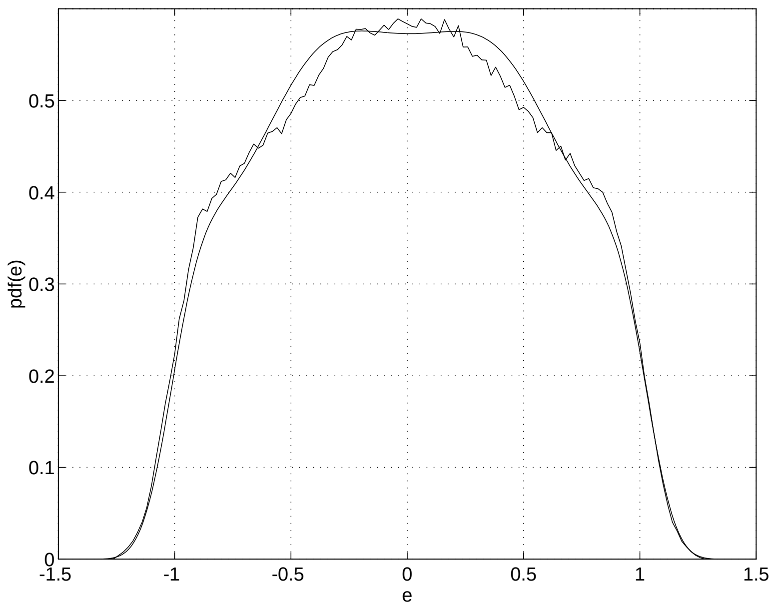

Figure 6.14: Predicted (smooth curve) and simulated quantizer input pdf for a modulator with filter B, cf. Table 6.2 and zero input.

> 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:
>
> $$\|\mathrm{etf}_{1.5}\|_1 = \sum_k |\mathrm{etf}_{1.5}(k)| = 2 \qquad (6.29)$$

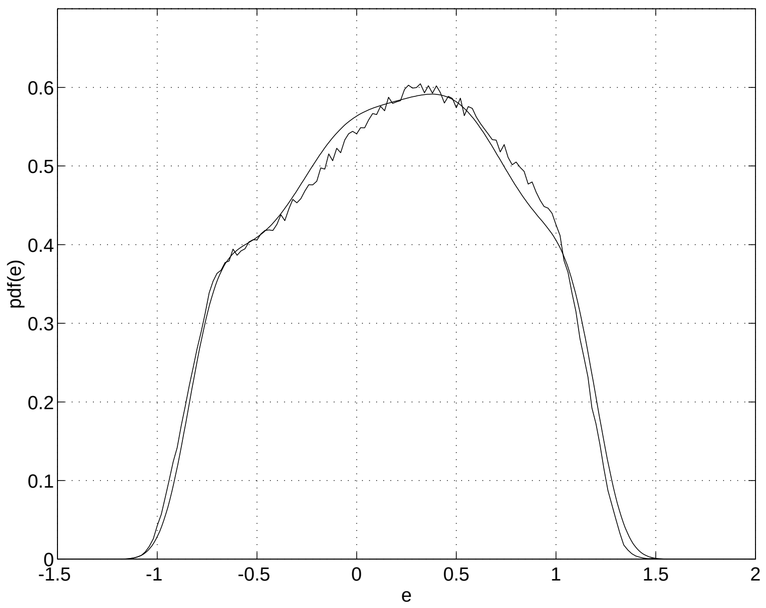

Figure 6.15: Predicted (smooth curve) and simulated quantizer input pdf for a modulator with feedback filter B, cf. Table 6.2 with $m_e = 0.2$ corresponding to $m_y = 0.2443$.

> 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,

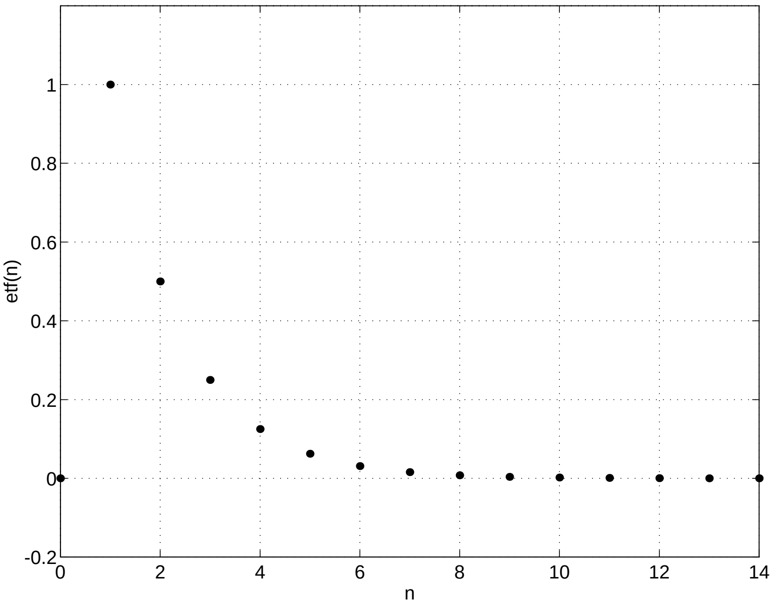

Figure 6.16: The impulse response $\mathrm{etf}_K(n)$ for the feedback filter $H(z) = z^{-1}/(1 - 2z^{-1})$ and $K = 1.5$.

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 $F(z)$