On the Design of Stable, High Performance Sigma Delta Modulators

by

Brett Christopher Hannigan

B.A.Sc. (Hons), Simon Fraser University, 2015

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Applied Science

in

THE FACULTY OF GRADUATE STUDIES
(Biomedical Engineering)

The University of British Columbia
(Vancouver)

November 2018

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

On the Design of Stable, High Performance Sigma Delta Modulators

submitted by **Brett Christopher Hannigan** in partial fulfillment of the requirements for the degree of **Master of Applied Science** in **Biomedical Engineering**.

Examining Committee:

Guy Dumont, Electrical and Computer Engineering
Supervisor

Abstract

This document provides brief instructions for using the `ubcdiss` class to write a **UBC!**-conformant dissertation in LaTeX. This document is itself written using the `ubcdiss` class and is intended to serve as an example of writing a dissertation in LaTeX. This document has embedded **URL!**s (**URL!**s) and is intended to be viewed using a computer-based **PDF!** (**PDF!**) reader.

Note: Abstracts should generally try to avoid using acronyms.

Note: at **UBC!** (**UBC!**), both the **GPS!** (**GPS!**) Ph.D. defence programme and the Library's online submission system restricts abstracts to 350 words.

Lay Summary

The goal of this work was a method to better design analog-to-digital converters with special interest to recording weak bio-signals, such as those from electroencephalography and electrocardiography.

The sigma delta architecture of analog-to-digital converters is known for having high resolution for signals of this class while requiring fewer expensive analog circuit components. However, as its performance is increased, it tends to become unstable, a point at which the digitized signal no longer accurately represents the original.

To this end, a theory and set of software tools were developed that use mathematical optimization and control theory to design sigma delta circuits with varying degrees of performance and stability. It is even possible to generate a design that is guaranteed to be stable. The method is generalizable to any kind of signal, medical or otherwise. These developments were used to analyze and synthesize designs and will hopefully inspire future high-resolution analog-to-digital converters.

Preface

At **UBC!**, a preface may be required. Be sure to check the **GPS!** guidelines as they may have specific content to be included.

Table of Contents

List of Tables

List of Figures

List of Symbols

K Variable quantizer gain (linearized model).

P_Q In-band quantization noise power.

$S(\lambda)$ Sensitivity function.

$T(\lambda)$ Complementary sensitivity function.

$\Delta_Q P$ Quantization step size.

λ Placeholder for the continuous-time Laplace variable s or discrete-time z-transformation variable z.

d Quantization noise source (linearized model).

e Feedback error signal.

n Filter order.

r Analog reference input signal.

u Quantizer input signal.

y Digital bitstream output signal.

Glossary

A/D analog-to-digital.

AAF antialiasing filter.

BIBO bounded-input bounded-output.

BMI bilinear matrix inequality.

BP band pass.

CCF controllable canonical form.

CLANS closed-loop analysis of noise shaper.

CT continuous-time.

D/A digital-to-analog.

DRF digital reconstruction filter.

DSP digital signal processing.

DT discrete-time.

ECG electrocardiography.

EEG electroencephalography.

FIR finite impulse response.

GKYP generalized Kalman-Yakubovič-Popov.

HP high pass.

IIR infinite impulse response.

LF loop filter.

LFT linear fractional transformation.

LMI linear matrix inequality.

LP low pass.

MSIA maximum stable input amplitude.

NTF noise transfer function, equivalent to the sensitivity function.

OSR oversampling ratio.

PDF probability density function.

PPG photoplethysmography.

S/H sample-and-hold.

SDP semidefinite program.

SQNR signal-to-quantization-noise ratio.

STF signal transfer function, equivalent to the complementary sensitivity function.

Acknowledgments

I would like to thank my supervisor Prof. Guy Dumont for his support of my research and openness to help as well as the other members of the BC Children's Hospital Research Institute's Digital Health Innovation Lab team.

I gratefully acknowledge the funding recieved from industry partner ESS Technologies whose participation in the Mitacs Accelerate program allowed me to write this thesis. The staff at ESS have been especially helpful, many thanks to Martin Mallinson and Chris Petersen in particular whose enthusiastic technical guidance and anecdoes about electronics, mathematics, and the early days of computing were a source of inspiration. Finally, thank you to the others who attended my progress meetings and provided valuable feedback.

Chapter 1

Introduction

The conversion of signals between analog and digital domanis is an often encountered problem in signal processing. For an analog signal to be represented digitally, it must undergo the processes of sampling and quantization (Figure 1.1). The former is the conversion from continuous-time (CT) to discrete-time (DT) and can be done without loss of information by the Nyquist-Shannon sampling theorem, given a sufficiently high sample rate. The latter is the mapping from an infinite set of possible values to a finite number of quantization levels. Unlike sampling, the process of quantization is non-injective and thus irreversible. The design of signal conversion circuits that minimize the error introduced by quantization is a major problem in mixed signal electronics.

Sigma delta modulation is a widely used technique for analog-to-digital (A/D) and digital-to-analog (D/A) conversion of signals that provides high resolution through the techniques of oversampling and noise shaping. Oversampling trades throughput for resolution, thus the sigma delta modulator generally lies between integrating converters, which are specialized for near-DC signals, and high-speed architectures, such as successive approximation and flash. The sigma delta quantization scheme is especially applicable to signals with low to moderate frequency content. Signals with these properties include most biosignals such as those recorded electrically (electroencephalography (EEG), electrocardiography (ECG)) or through other means using transducers (photoplethysmography (PPG)), as well as audio signals.

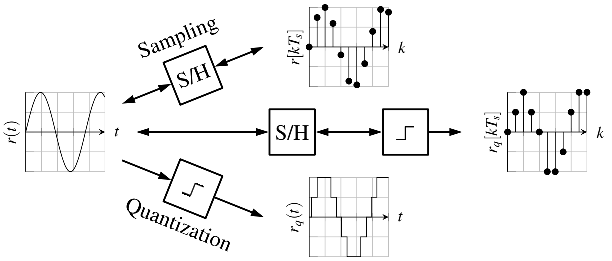

Figure 1.1: A continuous-time, continuous-value signal $r(t)$ is sampled to produce a discrete-time, continuous-value signal $r[kT_s]$. $r(t)$ independently undergoes quantization to yield a continuous-time, discrete-value signal $r_q(t)$. When both processes are applied in sequence, a discrete-time, discrete-value signal $r_q[kT_s]$ is the result.

1.1 Oversampling and Noise Shaping

Oversampling is simply the process where the analog signal is sampled at a rate higher than what the sampling theorem would dictate for perfect reconstruction, expressed as the oversampling ratio (OSR) relative to the Nyquist frequency. It may seem that this does not have a direct benefit *per se*, but it allows a less demanding analog antialiasing filter (AAF) to be used, saving circuit area. It also permits the quantization error to be spread across a larger bandwidth to increase resolution. Assuming quantization error can be modelled by white noise, oversampling reduces the in-band quantization noise power P_Q by a factor directly proportional to OSR [3] as seen in Equation 1.1, where $\Delta_Q P$ is the difference between quantization levels. These two advantages — reducing analog circuit complexity and increasing resolution — are common goals in sigma delta modulator design.

$$P_Q = \frac{\Delta^2}{12 \cdot OSR} \tag{1.1}$$

It may appear that oversampling alone quickly becomes impractical because one must approach very high sampling frequencies to increase the signal-to-quantization-noise ratio (SQNR) substantially. However, this assumes that the quantization noise is evenly distributed across the spectrum. Noise shaping is the use of a filter operating on the oversampled signal to push quantization noise out of the signal band where it can be removed by digital reconstruction filter (DRF). This

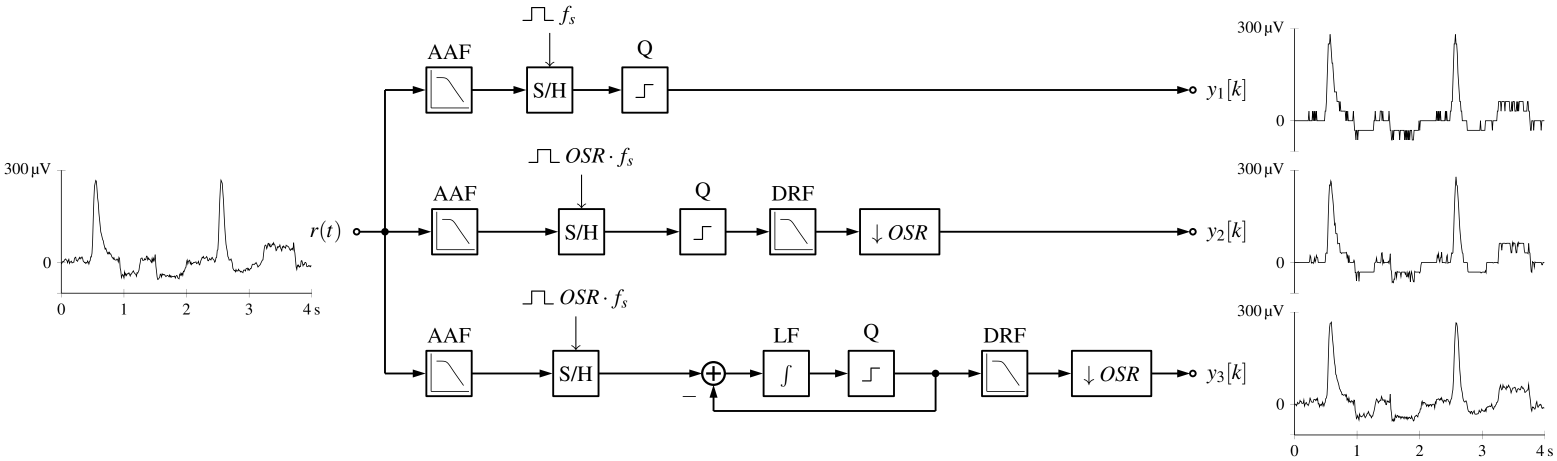

Figure 1.2: A comparison between naïve quantization (top), 10 times oversampled quantization (middle), and first order sigma delta modulation (bottom). The graphs on the right show the increasing quality of an EEG signal [1] sampled to a final rate of 100 Hz and quantized by Q with 5 bits by each scheme.

behaviour is implemented by wrapping the filter and quantizer in a feedback loop. With the same white noise assumption, the tradeoff between in-band shaped quantization noise and OSR is improved for ideal loop filters when order n is increased as shown in Equation 1.2 [3]. The effect of oversampling and noise shaping is demonstrated in Figure 1.2.

$$P_Q = \frac{\Delta^2 \pi^{2n}}{12\,(2n+1) \cdot OSR^{2n+1}} \tag{1.2}$$

1.2 Basic Structure

The basic block diagram of a sigma delta modulator and nomenclature that will be used herein is now introduced. For brevity, the scope is limited to sigma delta A/D converters but the concepts are easily transferrable to the D/A case. Modulators can be one of two main classes, CT or DT referring to the nature of the loop filter (LF).

1.2.1 Discrete-Time Modulator

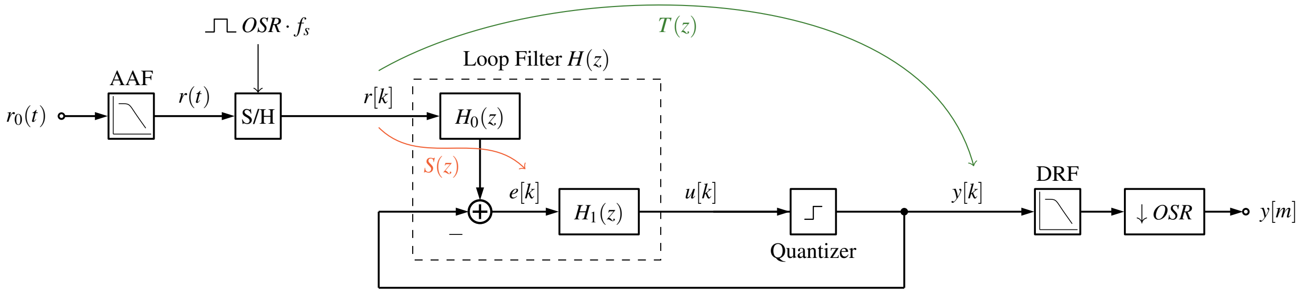

Figure 1.3: The basic block diagram of a DT sigma delta A/D converter.

Consider the DT modulator block diagram shown in Figure 1.3. The analog front-end includes the AAF and sample-and-hold block. This subsystem conditions the input signal $r_0(t)$ and samples it outside the loop to produce DT signal $r[k]$. In the modulator loop, the 2-input 1-output LF operates on $r[k]$ and the feedback signal, producing intermediate signal $u[k]$ with shaped noise. Then, $u[k]$ undergoes quantization producing discrete-value output $y[k]$. The quantizer output is fed back to the LF and also passed along. The final subsystem filters the signal from the shaped noise in the digital domain with a downsampling DRF to yield the final digital output $y[m]$.

From a control systems perspective, there are a couple of transfer functions that will be used to analyze and synthesize loop filters. The sensitivity function $S(\lambda)$, where $\lambda = z$, is known as the noise transfer function (NTF) of the modulator because it shows how the quantization error is filtered in the linearized model. The complementary sensitivity function $T(\lambda)$ is known as the signal transfer function (STF) of the modulator and shows how the signal is transformed by the modulator loop.

1.2.2 Continuous-Time Modulator

For the CT class of modulators, consider the structure of Figure 1.4. Both types are similar except the LF operates directly on analog input $r(t)$ in the CT domain and sampling is done inside the loop. The AAF is no longer necessary in most

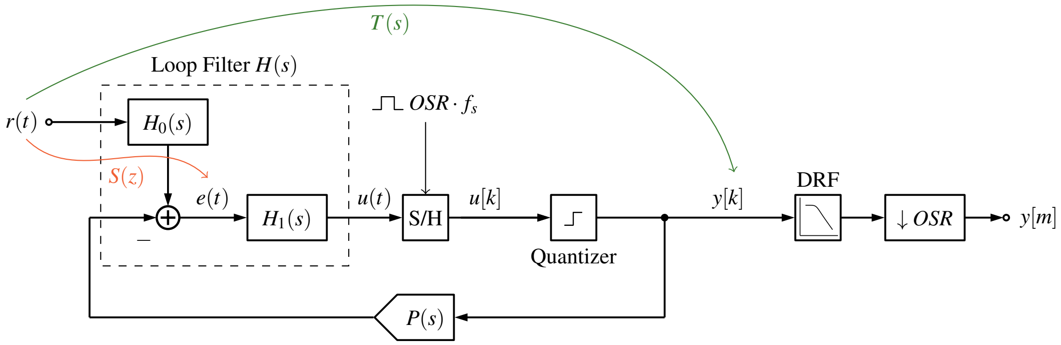

Figure 1.4: The basic block diagram of a CT sigma delta A/D converter.

cases as the LF precedes the sampling block and implicitly attenuates components of the signal that would result in aliasing. Finally, signal $y[k]$ must undergo D/A conversion during feedback, modelled witht the pulse transfer function $P(s)$.

The NTF and STF of a CT sigma delta modulator are more difficult to define because they are transfer functions involving both CT and DT signals. The DT equivalence principle states that there is a DT modulator model that exactly describes the CT design at the sampling instants, because the modulator is overall a sampled data system [4, Sec. 3.2]. Thus, DT transfer functions can be derived for this purpose. However, these equivalent transfer functions may be difficult to manipulate due to their dependence on $P(s)$. For the purposes of this analysis, we omit the sampling block during design and use the simplification that $S(\lambda)$ and $T(\lambda)$ are CT ($\lambda = s$) transfer functions mapping $t(t) \rightarrow e(t)$ and $r(t) \rightarrow y(t)$, respectively.

1.3 Loop Filter

Together, quantization and noise shaping permit a coarser quantizer element to be used. A common design pattern is to use a high (> 2) order LF paired with a 1-bit quantizer, which is advantageous from a circuit design perspective because a quantizer with just two levels is inherently linear. In addition, low order sigma delta loops often suffer from spurious tones [5, Sec. 2.6.1]. Unfortunately, as LF order is increased, the tendency of the loop to become unstable does as well. While first and second order designs are provably stable for DC inputs [6], high order filters require careful design to avoid instability. Ensuring stability while maintaining performance is a difficult task due to the presence of the highly nonlinear quantizer. The nonlinearity makes analysis complicated, a stable linear model does not imply a stable modulator while an unstable model can even result in a stable modulator known as the chaotic type [2].

The design of the noise shaping loop filter is the focus of this thesis. Modelling the loop filter as a 2-input 1-output system as shown in Section 1.2 allows the NTF to be determined by $H_1(\lambda)$ alone while the STF can be modified independently with filter $H_0(\lambda)$, without loss of generality:

$$S(\lambda) = \frac{1}{1 - H_1(\lambda)} \tag{1.3}$$

$$T(\lambda) = \frac{H_0(\lambda)}{1 - H_1(\lambda)}. \tag{1.4}$$

A desirable NTF is one that results in a stable linear model, rejects noise in the signal band as much as possible, and has low gain in the out-of-band region to promote stability. The STF is less important as $H_0(\lambda)$ can be interpreted as a pre-filter to modify the STF, but we prefer unity gain in the signal band.

For a first order modulator, a pure integrator can be used as the loop filter $H_0(\lambda)$. For higher orders, it is common to choose a prototype NTF from a family of filters. For example, the popular Delta Sigma Toolbox for MATLAB [5, Appx. B] uses a Chebyshev type II filter for this purpose. The choice of filter greatly affects the stability of the loop, so the traditional design procedure involves extensive simulation under varying input conditions to ensure instability is unlikely during normal operation. Once unstable, the filter states must be reset in order to restore operation. Various schemes to detect the onset of instability [7] and avoid it with gain scaling [8], internal linear feedback [9], and automatic resetting schemes [10].

1.4 Related Works

Optimization techniques have been used to design NTFs with more degrees of freedom than those made with a single filter prototype. A simple example is that from

[5, Sec. 4.3], where the zeros of the prototype NTF are optimized by approximating the integral of the NTF in the pass-band, then minimizing it analytically by equating its derivative to zero. The procedure results in an optimal spreading of zeros across the signal bandwidth for the given NTF poles. One of the first optimization-based approaches to NTF design was the closed-loop analysis of noise shaper (CLANS) methodology that minimizes P_Q under the white quantization noise assumption [11]. This is done using nonlinear optimization to find stable NTF pole locations that minimize the accumulation of quantization error subject to some stability and realizability constraints.

Using the principles from $\mathscr{H}_\infty$ control and its associated linear matrix inequality (LMI) methods, one can define the quantizer as a very simple feedthrough plant and introduce weighting filters on the feedback error signal e, loop filter output u, and quantizer output y to design the loop filter as a controller for various performance and stability constraints [12]. However, the system is bound to the order of the plant augmented with weighting filters and relies on the designer to choose the weights. Choosing weighting filters that are ideal is almost as difficult a task as just choosing the prototype NTF directly. Despite this, if a known AAF or DRF is specified in advance, it may be used as a sort of weighting filter and an optimal LF can be designed around it [13]. Applications for this method could be optimizing the STF to a psychoacoustic model or making use of existing filters in the signal path.

More recently, the generalized Kalman-Yakubovič-Popov (GKYP) lemma has been applied to sigma delta modulator design. The lemma provides a link between a finite frequency domain inequality, such as specifications on the NTF gain, and a linear matrix inequality condition, which can be solved using efficient interior point methods. Using this lemma, the techniques of $\mathscr{H}_\infty$ control can be applied to a transfer function but restricted to a frequency band. This eliminates the need for weighting filters that specify a select band of interest. Unfortunately, the problem becomes non-convex and hard to solve if both poles and zeros are to be optimized simultaneously as is the case with an infinite impulse response (IIR) filter. As a workaround, the NTF poles may be fixed to a prototype design and just the zeros optimized [14], similar to what was described above. Alternatively, a finite impulse response (FIR) NTF form may be assumed [15, 16] then possibly converted to IIR form using approximate methods such as least-squares or Yule-Walker [17]. Aside from the large delay introduced, the FIR form is not the optimal choice according to [18]. Iterative methods have shown promise in providing a workaround to the non-convexity associated with direct IIR design. A survey of some of these methods is presented in [19] while Table 1.1 summarizes the major contributions of each and differences between them.

Table 1.1: A comparsion of some recent work on sigma delta modulator design as a control optimization problem.

Reference	**Optimized norms**	**NTF Type**	**Performance goal**	**Stability criteria (see Chapter 3)**
Oberoi (2004) [12]	$\mathscr{H}_\infty$, $\mathscr{H}_2$, ℓ_1	IIR	Weighting filters	Uses heuristic bounds on $\mathscr{H}_\infty$, $\mathscr{H}_2$ norms
Osqui & Megretski (2007) [14]	$\mathscr{H}_\infty$	IIR[1]	GKYP lemma	Not reported
Nagahara & Yamamoto (2012) [15]	$\mathscr{H}_\infty$	FIR	GKYP lemma	ℓ_1 criterion mentioned, but Lee criterion used in design
Li, Yu, & Gao (2014) [20]	$\mathscr{H}_\infty$	IIR	GKYP lemma	Lee criterion
Tariq & Ohno (2016) [16]	$\mathscr{H}_\infty$, $\mathscr{H}_2$, ℓ_1	FIR	Weighting filters	ℓ_1 criterion mentioned but Lee criterion used in design

[1] Only the zeros of the IIR filter are optimized.

1.5 Organization of this Thesis

Having established some background on the workings and nomenclature of a sigma delta modulator, Chapter 2 expands on this to show modifications to the general sigma delta model based on control theory that will permit it to be used in an optimization framework. In Chapter 3, various stability criteria are introduced, ranging from heuristics to sufficient conditions and their impact on performance. Following the discussion of the role of optimization in loop filter design, Chapter 4 bridges the model and stability criteria chapters by introducing a semidefinite programming framework that supports the aforementioned criteria. The design process is discussed in Chapter 5, with emphasis on simulation results as well as an empirical study of the tradeoff between performance and stability when designing to different criteria. Finally, Chapter 6 concludes the thesis with some discussion about the merits and shortcomings of this method of sigma delta modulator design and possible directions for future work.

Chapter 2

Modelling the Sigma Delta Modulator

In order to apply an optimization framework to the design of the LF, the system from Figures 1.3 and 1.4 must be placed in a form that allows tractable application of the desired performance and stability targets. This includes omission of blocks that have minimal or no effect on the loop as well as linearization of the quantizer. The AAF (when present) can be considered as a pre-filter operating on the input signal. The filter $H_0(\lambda)$ serves as an additional degree of freedom for the STF can be set to unity for the purposes of the model. These two filters are not required in stabilitiy analysis, because the NTF depends only on $H_1(\lambda)$ as seen in Equation 1.4. After noise rejection performance has been optimized, $H_0(\lambda)$ can be tuned as necessary to ensure that the combined gain of the AAF and LF is close to unity in the signal band. In a similar way, the digital signal processing (DSP) in the output path serves only to filter out the signal and decimate to the original sampling frequency which may be dealt with separately without impacting loop stability.

2.1 Linearization of the Quantizer Element

Next, the nonlinear nature of the quantizer is dealt with. As mentioned before, a common linearization approach is to replace the quantizer with an additive noise source d. Furthermore, the linear model can incorporate a variable gain K. The

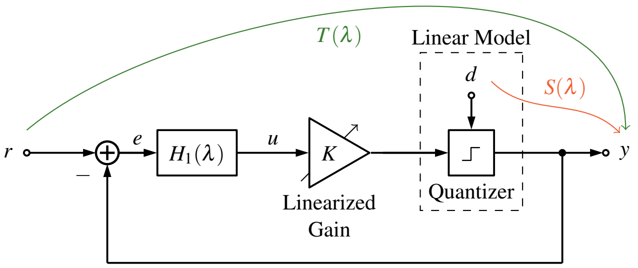

Figure 2.1: The linearized sigma delta loop block diagram with omission of extraneous filters and the quantizer replaced by a variable gain and additive quantization noise signal.

inclusion of K has uses in linearization, stability, and performance that will be expanded upon in Chapter 3. After these simplifications, the block diagram in Figure 2.1 is obtained, which is applicable to DT or CT designs. In the DT case, the loop is operating entirely in the oversampled domain and the sample-and-hold (S/H) block is not shown. In the CT case, the S/H block in the loop is neglected so that $S(\lambda)$ and $T(\lambda)$ are CT transfer functions[1].

2.2 Well-Posedness and Internal Stability

The meaningful application of feedback to reduce an uncertainty (in this case, error introduced by the nonlinear quantizer) requires that the system be well-posed in order for a solution to exist. Figure 2.1 can undergo block diagram mainpulation bringing it into the standard feedback form shown in Figure 2.2 with signals r, e, d, and y.

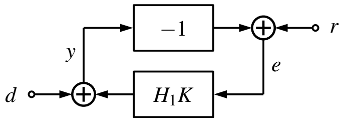

Figure 2.2: The linearized model converted into standard feedback form.

[1] Note that regarding Figure 2.1 and Figures 1.3/1.4, the NTF $S(\lambda)$ is the same transfer function whether interpreted from $d \rightarrow y$ or $r \rightarrow e$.

The equations describing this loop are:

$$\begin{bmatrix} r \\ d \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -H_1K & 1 \end{bmatrix} \begin{bmatrix} e \\ y \end{bmatrix}. \tag{2.1}$$

A feedback system is considered well-posed if the inverse of the transfer matrix in Equation 2.1 exists and each of its elements are proper. Equation 2.2 shows that this is the case if both $S(\lambda)$ and $T(\lambda)$ are proper transfer functions.

$$\begin{bmatrix} e \\ y \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -H_1K & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{1+H_1K} & \frac{-1}{1+H_1K} \\ \frac{H_1K}{1+H_1K} & \frac{1}{1+H_1K} \end{bmatrix} \begin{bmatrix} r \\ d \end{bmatrix} = \begin{bmatrix} S & -S \\ T & S \end{bmatrix} \begin{bmatrix} r \\ d \end{bmatrix}. \tag{2.2}$$

The principle of internal stability is stricter than bounded-input bounded-output (BIBO) stability because it guarantees that the internal states of the system remain bounded. The system in Equation 2.2 is internally stable if each element of the transfer matrix belongs to the set $\mathscr{RH}_\infty$, that is the set of stable real rational proper transfer functions.

2.2.1 Constraints on the Noise Transfer Function

A sufficient condition for $S(\lambda)$ and $T(\lambda)$ to be proper is that transfer function $H_1(\lambda)$ is a strictly proper real rational transfer function. Internal stability of the system follows if $S(\lambda)$ and $T(\lambda)$ are stable. This leads to the following constraints on the NTF:

1. $S(\lambda)$ is stable, and ,
2. The following equivalent conditions hold:
   (a) $S(\infty) = 1$,
   (b) If $S(\lambda)$ is in state-space form, the feedthrough matrix $D = 1$, and,
   (c) The first element of the impulse response of $S(\lambda)$ is one.

Most prior work in the area performs optimization directly on the NTF of the system. This is effective because it is a relatively accurate model of the noise shaping performance. In addition, constraint 2 enforces causality on the feedback loop ensuring the system is physically realizable.

2.3 Modelling Uncertain Quantizer Gain

Having established conditions to ensure the closed-loop system is realizable and internally stable, there remains a nonlinear gain block K. K can be understood as a time-varying gain dependent on the quantizer input. For example, a 1-bit quantizer ($\Delta_Q P$= 2) with output $\{-1,1\}$ would have instantaneous gain $K(t)=\frac{1}{u(t)}$. As the value of u at each sample time is not known in advance, K may be modelled as a multiplicative uncertainty. The upper linear fractional transformation (LFT) allows K to be separated into a constant gain matrix $M_{2\times2}$ and a normalized, $\mathscr{H}_\infty$ norm-bounded uncertain block Δ by Expression 2.3.

$$K \leftrightarrow \mathscr{F}_U\{M,\Delta\} \quad ||\Delta||_\infty \leq 1 \tag{2.3}$$

The model from Figure 2.2 is shown in Figure 2.3 with the quantizer and variable gain replaced by this LFT interconnection. In Chapter 4, it is of interest to ensure the robustness of the system to Δ, which may be achieved using this form.

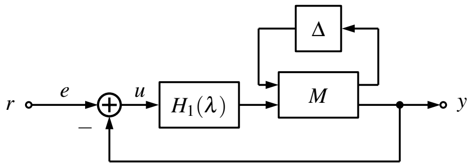

Figure 2.3: The linearized block diagram with the quantizer replaced by a multiplicative uncertainty extracted via LFT.

2.4 Derivation of Augmented System

2.4.1 Extraction of Performance and Stability Channels

Finally, the model is abstracted into an augmented form where all desired input and output channels are present and all unnecessary ones hidden. Let the LFT input to M and output from M be w and z, respectively. These channels are required to

Table 2.1: Input and output channels of interest for the augmented system.

Output Input	z	e	u	y
r	Not used	NTF performance channel	Constraint on quantizer input signal	STF constraints for CT design
w	Quantizer gain robustness channel	Not used	Not used	Not used

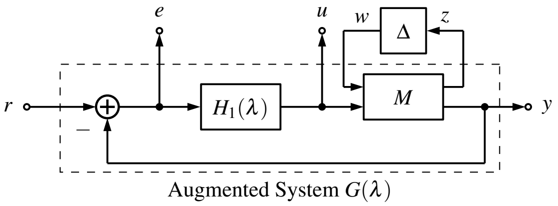

Figure 2.4: The augmented plant is derived by setting $H_0(\lambda) = 1$, taking the LFT of the uncertain gain, extracting the signals of interest, and writing the closed-loop equations.

be accessed in addition to r, e, u, and y for the purposes listed in Table 2.1. The augmented system $G(\lambda)$ is shown as the dashed block in Figure 2.4.

2.4.2 Derivation of State-Space Model

Now that the desired input and output signals are captured by the model, it is a simple exercise to write the system in state-space form. To begin, let filter $H_1(\lambda)$ be the transfer function of order n in variable $\lambda = z$ in the DT case (or $\lambda = s$ in the CT case). The numerator and denominator coefficients are shown in Equation 2.4 which has the equivalent state-space representation of Equation 2.5.

$$H_1(\lambda) = \frac{U(\lambda)}{E(\lambda)} = \frac{b_{n-1}\lambda^{n-1} + b_{n-2}\lambda^{n-2} + \ldots + b_1\lambda + b_0}{\lambda^n + a_{n-1}\lambda^{n-1} + a_{n-2}\lambda^{n-2} + \ldots + a_1 z\lambda + a_0} \tag{2.4}$$

$$= C_H(\lambda I - A_H)^{-1} B_H \tag{2.5}$$

Naturally, $H_1(\lambda)$ is a strictly proper transfer function and state-space feedthrough matrix $D_H = 0$ due to the constraints proposed in Section 2.2. The constant gain matrix M may be split into its constituent parts as shown in Equation 2.6.

$$M = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \tag{2.6}$$

With some algebra, the augmented system $G(\lambda)$ from Figure 2.4 may be written in state-space form with notation from Equations 2.5 and 2.6 by introducing state vector x. We use the notation $G_{qp}(\lambda)$ to indicate the transfer function of $G(\lambda)$ from some input channel p to some output channel q and name the closed-loop state-space matrix blocks with cursive symbols as shown in Equation 2.8.

$$G : \begin{bmatrix} \dot{x} \\ z \\ e \\ u \\ y \end{bmatrix} = \left[\begin{array}{c|cc} A_H - m_{22}B_HC_H & -m_{21}B_H & B_H \\ \hline m_{12}C_H & m_{11} & 0 \\ -m_{22}C_H & -m_{21} & 1 \\ C_H & 0 & 0 \\ m_{22}C_H & m_{21} & 0 \end{array}\right] \begin{bmatrix} x \\ w \\ r \end{bmatrix} \tag{2.7}$$

$$= \left[\begin{array}{c|cc} \mathscr{A} & \mathscr{B}_w & \mathscr{B}_r \\ \hline \mathscr{C}_z & \mathscr{D}_{zw} & \mathscr{D}_{zr} \\ \mathscr{C}_e & \mathscr{D}_{ew} & \mathscr{D}_{er} \\ \mathscr{C}_u & \mathscr{D}_{uw} & \mathscr{D}_{ur} \\ \mathscr{C}_y & \mathscr{D}_{yw} & \mathscr{D}_{yr} \end{array}\right] \begin{bmatrix} x \\ w \\ r \end{bmatrix} \tag{2.8}$$

With the channels of interest exposed and the system in a state-space form, one can express design goals as constraints on these channels. In Chapter 3, various stability measures and performance goals are discussed from which those that are ideal from an optimization perspective are selected. In Chapter 4, the framework is introduced to allow the targets from the previous chapter to be applied to the augmented system in a way that allows the optimization problem to be efficiently solved.

Chapter 3

Stability Criteria and Performance Goals

Due to the nonlinear effects of the quantizer, the stability of the sigma delta feedback loop is difficult to prove. An excellent exploration into the mechanisms of instability may be found in [2]. There are no known necessary conditions for stability of sigma delta modulators but there are several heuristic and sufficient conditions with various degrees of conservativenenss. In Section 3.1.1, some theory from relay feedback control is introduced to establish formal methods for ensuring stability. There are some shortcomings of these methods when applied to practical sigma delta modulator design, therefore Sections 3.1.2 to 3.1.5 describe some stability criteria that may be less robust but allow a greater performance-stability tradeoff. For completeness, additional stability methods of interest that are not compatible with this optimization framework are presented in Section 3.2. Finally. the performance goal is discussed in Section 3.3.

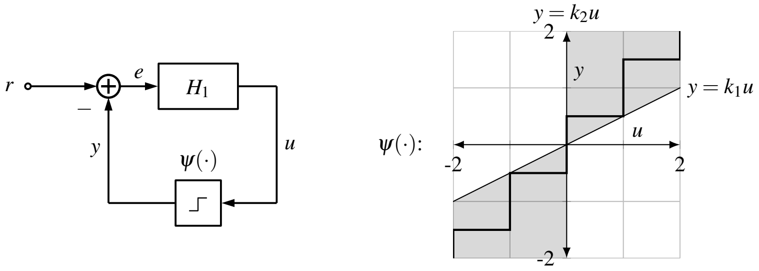

Figure 3.1: A Lur'e system (left) with the example nonlinear transfer curve of an infinite quantizer ($\Delta = 1$) shown with a shaded sector bounded region (right).

3.1 Stability Criteria Used by this Optimization Framework

3.1.1 Ideas from Nonlinear Control

An early theoretical treatment of nonlinear control is the circle criterion, which provides a graphical frequency domain method for evaluating the stability of a CT Lur'e system. A Lur'e system is a simplified negative feedback loop consisting of a linear plant $L(s)$ with a nonlinear element $\psi(\cdot)$ in the feedback path such as the one shown in Figure 3.1. The transfer curve of the nonlinear element may be time-varying and even non-monotonic but is bounded by a sector condition, a set of two lines passing through the origin with slopes k_1, k_2 that bound the curve.

Theorem 3.1.1 (Circle criterion [21, Sec. 7.1.1])**.** *Given the Lur'e feedback system in Figure 3.1 where the denominator of $H_1(s)$ is Hurwitz and $\psi(t,\cdot)$ is a memoryless function sector bounded by $[k_1,k_2]$, the closed-loop system $L(s)$ is globally asymptotically stable if one of the following cases is true:*

1. *Case $\psi \in [k_1,\infty)$: The inequality in Equation 3.1 is satisfied.*

$$\Re\left\{\frac{H_1(s)}{1+k_1H_1(s)}\right\} > 0 \tag{3.1}$$

2. *Case $\psi \in [k_1,k_2], \quad k_2 - k_1 > 0$: The inequality in Equation 3.2 satisfied.*

$$\Re\left\{\frac{1+k_2H_1(s)}{1+k_1H_1(s)}\right\} > 0 \tag{3.2}$$

The graphical interpretation for the circle criterion[1] is that the Nyquist plot of $H_1(s)$ does not enter the disk passing through the points $-\frac{1}{k_1}+j0$ and $-\frac{1}{k_2}+j0$ if $0 < k_1 < k_2$. When $0 = k_1 < k_2$, the Nyquist plot must lie to the right of the vertical line $\Re\{s\} = -\frac{1}{k_2}$. If a single-bit quantizer is used as is the scope of this thesis, the sector bounds include the entire first and third quadrants. Case 1 from Theorem 3.1.1 then applies and the Nyquist plot of $H_1(s)$ must lie entirely in the right half-plane. The optimization framework presented in Chapter 4 may be used with the circle criterion although with a single-bit quantizer, the method is too restrictive for practical use.

The Popov criterion in Theorem 3.1.2 is a slightly less conservative approach that restricts the problem to time-invariant nonlinearities.

Theorem 3.1.2 (Popov criterion [21, Sec. 7.1.2])**.** *Given the Lur'e feedback system in Figure 3.1 where the denominator of $H_1(s)$ is Hurwitz and $\psi(\cdot)$ is a time-invariant memoryless function sector bounded by $[0,k_2]$, the closed-loop system $L(s)$ is globally asymptotically stable if there exists a scalar $\gamma \geq 0$ such that the following inequality is satisfied:*

$$\frac{1}{k_2} + \Re\{H_1(j\omega)\} - \gamma\omega\Im\{H_1(j\omega)\} > 0 \quad \forall\omega \in [0,\infty). \tag{3.3}$$

The graphical interpretation for this criterion is that the Popov plot of $\omega\Im\{H_1(j\omega)\}$ versus $\Re\{H_1(j\omega)\}$ remains to the right of a line passing through point $-\frac{1}{k_2}+j0$ with slope $\frac{1}{\gamma}$.

The DT version is the Tsypkin criterion, which has cases valid for time-varying and time-invariant nonlinearities. The analog to the circle criterion is shown in Theorem 3.1.3 and the analog to the Popv criterion is shown in Theroem 3.1.4.

[1] The circle criterion has different graphical interpretations for the cases where $k_1 < 0$ and where $H_1(s)$ has zeros in the open right half-plane, but these are omitted because they are not valid quantizer transfer curves or because nonminimum phase $H_1(s)$ are not considered here (see [21, Sec. 7.1.1] for a full treatment).

Theorem 3.1.3 (Tskypin criterion for time-varying nonlinearities [22, Sec. 4.6])**.** *Given the Lur'e feedback system in Figure 3.1 where the denominator of $H_1(z)$ is Schur and $\psi(t,\cdot)$ is a memoryless function sector bounded by $[0,k_2]$, the closed-loop system $L(z)$ is globally asymptotically stable if the following inequality is satisfied:*

$$\frac{1}{k_2} + \Re\{H_1(z)\} \geq 0 \quad \forall |z| = 1. \tag{3.4}$$

Theorem 3.1.4 (Tskypin criterion for time-invariant nonlinearities [22, Sec. 4.7])**.** *Given the Lur'e feedback system in Figure 3.1 where the denominator of $H_1(z)$ is Schur and $\psi(\cdot)$ is a time-invariant memoryless function sector bounded by $[0,k_2]$, the closed-loop system $L(z)$ is globally asymptotically stable if there exists a scalar $\gamma \geq 0$ such that the following inequality is satisfied:*

$$\frac{1}{k_2} + \Re\left\{\left(1+\gamma\left(1-z^{-1}\right)\right) H_1(z)\right\} \geq 0 \quad \forall |z| = 1. \tag{3.5}$$

The Jury-Lee criteria are less strict cases of the Tsypkin criteria requiring that the nonlinearity be slope bounded and monotonic. However, this is not applicable to quantizer feedback, where the slope may go to infinity. The above techniques from nonlinear control are sufficient conditions and are related to important results from passivity theorem.

3.1.2 $\mathscr{H}_\infty$ Stability Criterion

The $\mathscr{H}_\infty$ stability criterion, commonly known as Lee's rule, is a heuristic predictor of stability stating that a modulator is likely to be stable if the NTF out-of-band gain, or $||S(\lambda)||_\infty$, does not exceed a benchmark value. The rule was initially based on the empirical study of a fourth-order DT sigma delta modulator with single-bit quantization [23]. The criterion is not necessary nor sufficient for stability and must be verified with extensive simulations. Despite this, the rationale for its use as a suggestion of stability comes from the Bode sensitivity integral shown in Equation 3.6 for Schur stable $H_1(z)$ [24, Thm. 1].

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \log|S(\omega)|\,d\omega = 0 \tag{3.6}$$

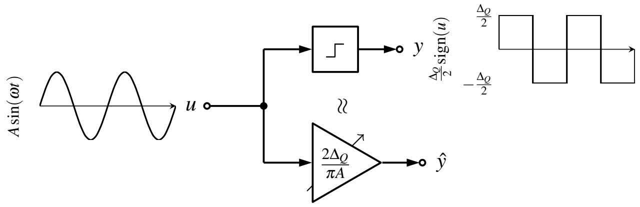

Figure 3.2: An ideal 1-bit quantizer (above) and its describing function approximation (below).

The integral enforces that the total area under the curve of the NTF log-magnitude versus frequency is equal to zero. Applied to the sigma delta linear model, if the sensitivity of the closed-loop system to the quantization error is suppressed in the signal band, it must be compensated for by an equal area of amplified sensitivity outside the signal band. Because the quantization error is nonlinear and signal dependent, the higher the gain of the sensitivity function, the greater chance there is for a limit cycle at that frequency to destabilize the loop. Thus, the Lee's rule is a indicator of the performance-stability tradeoff. In practice, $||S(z)||_\infty \leq 2$ is often used, but this has been found to be conservative for low-order and inadequate for high-order designs [25]. However, the criterion is used extensively as a starting point for practical design due to its inclusion in popular software tools. A recent extension of Lee's rule produces a more complex $\mathscr{H}_\infty$ bound based on empirical results that may be used as an alternative $\mathscr{H}_\infty$ stability goal [26]. This stability criterion is easy to formulate as part of an optimization problem as it may be applied with an $\mathscr{H}_\infty$ constraint on the $r \rightarrow e$ channel of Equation 2.7:

$$||G_{er}(\lambda)||_\infty \leq \gamma_\infty. \tag{3.7}$$

3.1.3 Describing Function Approximation and Root Locus Stability

Two closely related stability methods are the describing function approximation and the root locus approach. These both rely on the variable gain K introduced in Section 2.1 but use different interpretations of it to stabilize the sigma delta modulator.

The describing function method [27] is an approximate technique of linearization often applied to steady-state electrical circuits or to nonidealities in mechanical systems. As shown in Figure 3.2, a zero-mean sinusoidal input to the quantizer is assumed: $u(t) = A\sin(\omega t)$. The Fourier series of the output is truncated at the first odd coefficient because the quantizer transfer curve is also an odd function. For a single-bit quantizer, the first coefficients are:

$$a_1 = \frac{1}{\pi}\int_0^{2\pi} \psi(t)\cos(\omega t)d(\omega t) \tag{3.8}$$

$$b_1 = \frac{1}{\pi}\int_0^{2\pi} \psi(t)\sin(\omega t)d(\omega t), \tag{3.9}$$

where $\psi(t)$ is the nonlinear quantizer function. Integral 3.8 evaluates to zero while the period of Integral 3.9 may be split into two parts:

$$b_1 = \frac{1}{\pi}\left(\int_0^{\pi} \psi(t)\sin(\omega t)d(\omega t) + \int_{\pi}^{2\pi} \psi(t)\sin(\omega t)d(\omega t)\right).$$

In the interval $\omega t \in (0, \pi)$, the single-bit quantizer outputs $\frac{\Delta_Q}{2}$ whereas in the interval $\omega t \in (\pi, 2\pi)$, the single-bit quantizer outputs $-\frac{\Delta_Q}{2}$. By symmetry, this integra simplifies to Equation 3.10.

$$b_1 = \frac{\Delta_Q}{\pi}\int_0^{\pi} \sin(\omega t)d(\omega t) = \frac{2\Delta_Q}{\pi} \tag{3.10}$$

Using the Fourier series approximation $\hat{y}(t) \approx y(t)$, the describing function is derived as follows:

$$\begin{aligned} K(A) &= \frac{\hat{y}(t)}{u(t)} \\ &= \frac{b_1 \sin(\omega t)}{A\sin(\omega t)} \\ &= \frac{2\Delta_Q}{\pi A}. \end{aligned}$$

Thus, the describing function of the single-bit quantizer is a variable gain $K(A)$ dependent on the quantizer input amplitude. As expected, when the input amplitude approaches zero, the gain approaches infinity and when it approaches $\pm\frac{\Delta_Q}{2}$, the gain approaches one. In fact, instability in sigma delta modulators is often associated with low frequency, large amplitude limit cycles where the quantizer gain is low. The describing function method is a good approximation to large signal stability but often fails to predict small limit cycles because the higher harmonics of the output are neglected. The describing function has been applied to the design of sigma delta modulators [28] and extended to be dependent on phase in addition to gain to design a sixth-order modulator [29].

In the open loop, the describing function method approximated the quantizer as a variable gain. The root locus can determine stability by showing the position of the closed-loop poles as a function of this gain. One method to design stable sigma delta modulators is to position the poles and zeros of the loop filter such that the root locus remains in the stable region of the complex plane when sweeping through valid quantizer gain values [30–32]. Recall the LFT used to model the varying gain in Section 2.3. Equation 3.11 defines M when the gain is within a given range $K \in [k_l, k_h]$ with nominal value k_0.

$$M = \begin{bmatrix} \frac{k_h - 2k_0 + k_l}{k_h - k_l} & \frac{-2(k_0 - k_h)(k_0 - k_l)}{k_h - k_l} \\ 1 & k_0 \end{bmatrix} \tag{3.11}$$

The root locus stability criterion may be used in robust control fashion by choosing a range for K, e.g. $[k_l, k_h] = [1/||u||_\infty, \infty]$, then constraining the $\mathscr{H}_\infty$ norm to unity for the $z \rightarrow w$ channel as seen in Equation 3.12. This ensures that the linearized model is stable for the selected gain values.

$$||G_{zw}(\lambda)||_\infty \leq 1. \tag{3.12}$$

3.1.4 $\mathscr{H}_2$ Stability Criterion

Previously, the quantizer was replaced by an additive noise source and some performance estimations were presented assuming that the quantization noise was uncorrelated with input had a white spectrum. The white noise model is only a close approximation if the following hold [33, Ch. 6]:

1. The quantizer is not overloaded,

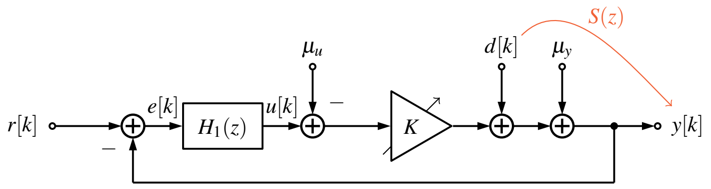

Figure 3.3: The block diagram for the $\mathscr{H}_2$ stability criterion.

2. There are a large number of quantization levels with small $\Delta_Q P$, and

3. The probability density function (PDF) of input samples is smooth.

In reality, especially with single-bit quantization, the approximation does not hold. The $\mathscr{H}_2$ stability criterion, sometimes called the power gain rule, uses a statistical look at the quantizer input [2]. The output $y[k]$ of a single-bit quantizer may be considered the superposition of three signals: a DC component μ_y, AC component amplified by the quantizer gain $K(u[k]-\mu_u)$, and the quantization noise $d[k]$, as seen in Figure 3.3. With these additional degrees of freedom, one can enforce that $d[k]$ is zero mean, white, and uncorrelated with the quantizer input $u[k]$ by setting K to that in Equation 3.13.

$$K = \frac{\text{cov}\{u[k], y[k]\}}{\sigma_u^2} \tag{3.13}$$

The gain K is now entirely dependent on μ_y and the distribution of u. Due to the fact that $y[k]$ has bounded output power (equal to one with $\Delta_Q = 2$ single-bit quantization), the variance at the output may be calculated:

$$\sigma_y^2 = K^2\sigma_u^2 + \sigma_d^2. \tag{3.14}$$

Functions relating μ_y to σ_q^2 have been derived for the Gaussian [34, Eq. 26], uniform, and triangular [2, Eq. 6.16, 6.17] distributions. Once a distribution has been chosen, the only free variable remaining is μ_y.

With reference to Figure 3.3, the NTF is the gain from quantization noise to the output. Because the $\mathscr{H}_2$ norm is a power gain, $||S(z)||_2^2$ is the amplification of σ_q^2 to σ_y^2. Altogether, μ_y defines σ_q^2, so for bounded output power, a maximum $||S(z)||_2$ is

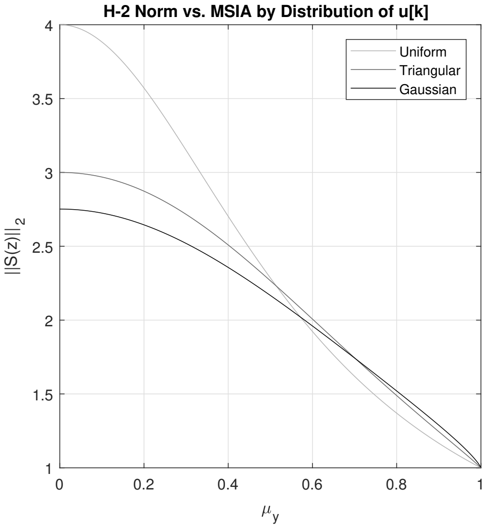

Figure 3.4: The choice of PDF for the quantizer input signal places bounds on the squared $\mathscr{H}_2$ norm of the NTF for a given MSIA [2].

established. The $\mathscr{H}_2$ stability criterion can be applied by choosing a quantizer input signal distribution,which sets the relationship between μ_y and $||S(z)||_2$. Because the STF gain is near one in the signal band, μ_y tracks input r and thus can be seen as the maximum stable input amplitude (MSIA) for that value of $||S(z)||_2$. The maximum permitted MSIA for a given $\mathscr{H}_2$ norm is shown in Figure 3.4 for the three probability density functions mentioned. None of the cases guarantee stability because the actual PDF of the quantizer input is not known. However, the Gaussian PDF has been shown to be a close approximation for high-order designs [2] despite being more conservative than the others.

To employ this stability criterion, the $\mathscr{H}_2$ norm of the sensitivity function, or $r \rightarrow e$ channel of Equation 2.7, is constrained to a value dependent on the desired MSIA $||u||_\infty$ and the choice of PDF. This criterion is seen in Equation 3.15 and is only applicable to DT designs because the CT sensitivity function has infinite $\mathscr{H}_2$ norm.

$$||G_{er}(z)||_2 \leq \gamma_2. \tag{3.15}$$

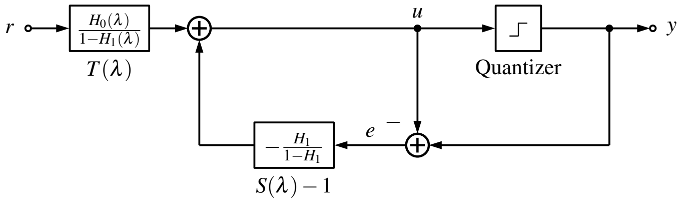

Figure 3.5: The block diagram of the error feedback form of a sigma delta modulator.

3.1.5 ℓ_1 Stability Criterion

The ℓ_1 stability criterion (also known as the Anastassiou's stability criterion) is best understood when the modulator is transformed into the equivalent error feedback structure shown in Figure 3.5. This form is primarily of theoretical importance because its implementation is extremely sensitive to filter coefficient values in the feedback path [5]. The advantage for analysis is that the NTF and STF are shown in independent blocks. We can use this structure to define the ℓ_1 norm stability criterion by writing the equations for signals u and y as follows:

$$\begin{aligned} u &= rT(\lambda) + e\left(S(\lambda) - 1\right) \qquad (3.16) \\ y &= e + u \\ &= rT(\lambda) + eS(\lambda). \end{aligned}$$

In the time domain, the input to the quantizer from Equation 3.16 can be bounded by the following (the DT version is shown):

$$|u[k]| \leq \left|\sum_{i=1}^{\infty} t[i]r[t-i]\right| + \left|\sum_{i=1}^{\infty} (s[i]-1)\, e[t-i]\right|, \qquad (3.17)$$

where $t[k]$ is the impulse response of the STF and $s[k]$ is the impulse response of the NTF. Because the STF only operates on the input as a pre-filter, we can assume $T(\lambda) = 1$ and Equation 3.17 simplifes to:

$$|u[k]| \leq |r[k]| + \left| \sum_{i=1}^{\infty} (s[i]-1)\, e[t-i] \right|. \tag{3.18}$$

With a single-bit quantizer, the output y is quantized to $\{-1,1\}$ so the quantization error signal is bounded to within $[0, \frac{\Delta_Q}{2}]$ if $|u[k]| \leq 2$, where $\Delta_Q P$ is the quantization step size, in this case 2.

$$|u[k]| \leq |r[k]| + \left| \sum_{i=1}^{\infty} (s[i]-1) \right| \frac{\Delta_Q}{2} \tag{3.19}$$

Taken over all time, the maximum magnitude of each signal is the ℓ_∞ norm and the summation over the impulse response of $S(\lambda)-1$ in Equation 3.19 is its ℓ_1 norm. The ℓ_1 norm is equivalent to the maximum peak-to-peak gain of the system. Substituting this into Equation 3.19 and simplifying $\frac{\Delta_Q}{2} = 1$ gives:

$$\begin{aligned} ||u||_\infty &\leq ||r||_\infty + ||S(\lambda)-1||_1 \\ &\leq ||r||_\infty + ||S(\lambda)||_1 - 1. \end{aligned}$$

Assuming a worst case quantizer input magnitude of $||u||_\infty = 2$, the expression for ℓ_1 stability based on maximum input magnitude is shown in Equation 3.20 [35].

$$||S(\lambda)||_1 \leq 3 - ||r||_\infty \tag{3.20}$$

Thus, a modulator with a single-bit quantizer is guaranteed to be stable for inputs of magnitude less than three minus the ℓ_1 norm of the NTF. If this value is negative, the modulator is not proven stable by this criterion, for any input. The ℓ_1 criterion is powerful because it is a sufficient condition of BIBO stability but it is very conservative as it assumes the worst-case input to the loop filter which may be impossible or extremely unlikely. Like the $\mathscr{H}_2$ and $\mathscr{H}_\infty$ cases, the ℓ_1 stability criterion may be applied to the $r \rightarrow e$ system in Equation 2.7 by the following:

$$||G_{er}(\lambda)||_1 \leq \gamma_1. \tag{3.21}$$

3.1.6 Scale Invariance of the Single-Bit Quantizer

Here an important result is mentioned that reduces conservatism in the stability criteria from Section 3.1.4 and Section 3.1.5 that are derived from a norm constraint on the NTF. If K is interpreted not a a time-varying gain that models the quantizer but as a fixed design variable, it is obvious that the value of K has no effect if it precedes the single-bit quantizer. The filter $H_1(\lambda)$ in expression for the NTF from Equation 1.3 may be substituted by the filter and gain in series, yielding:

$$S(K,\lambda) = \frac{1}{1 - KH_1(\lambda)}. \tag{3.22}$$

The fictitious gain may be swept in the region $(0,\infty)$ in a nonlinear search to minimize the norm of interest. Because all these NTFs are equivalent, the smallest achieved value $\min_K ||S(K,\lambda)||_p$ can be used in place of $||S(\lambda)||_p$ for $p = 1,2$ as a less conservative stability criterion.

3.2 Stability Concepts Not Used by this Optimization Framework

It may be worthwhile to introduce further methods of designing sigma delta modulators that are likely stable. The method in this section are presented for a greater understanding of ways to promote stability but are not able to be used in the optimization framework for the given reasons.

3.2.1 Methods Ensuring Bounded States

A different way of ensuring stability of a modulator system is the positive invariant set approach. This is a mixed analytical and simulation-based method to find positive invariant sets in the state-space of the system. These are regions of n-dimensional space for which, once entered, the states of the system x remain inside under given input conditions. The method is computationally intesive and relies on sampling to expand or contract the set bounds [5], which is not rigorous. However, given a random enough input, it may be a very close approximation to the actual positive invariant sets. A simpler version of this technique uses hypercubes as set boundaries and can be combined with the ℓ_1 rule to ensure stability [36].

The method is a good analysis of predicted stability because it also captures the integrator state values of the system which are important for the actual circuit implementation. However, it is not easy to apply as a design method.

3.2.2 Diagonal Modulators

It is shown in [37] that many modulator topologies can be converted into a state-space diagonal representation of the loop

filter. In these cases, an nth order modulator decomposes into n parallel 1st-order modulators interacting only through the quantizer nonlinearity. By examining the stability of the 1st-order modulator using its fixed points [38], the parallel modulators can be shown as a "shifted" version, where the shift indicates an offset at the quantizer input due to the other parallel paths. The equations for the

fixed points are rearranged to form constraints on the system's poles, inputs, and output matrix coefficients. The downside of this method is that the loop filter is restricted to a specific form and the mathematics become diffcult when complex conjugate poles are introduced [39].

3.3 Performance Goals

The performance goal for the design of sigma delta modulators is simply the attenuation of quantization error in the signal band. In a signals and systems context, this can be made solvable by the minimization of the $\mathscr{H}_\infty$ norm of the noise transfer function within the signal band. In the sigma delta literature, this is sometimes called min-max optmization and has some advantages in contrast to minimization of the power using the $\mathscr{H}_2$ norm [15]. To frame the problem, one must first specify a frequency range of interest. In the DT case, this is equal to $\frac{\pi}{OSR}$ and in the CT case, this is equal to the actual signal band. Using the GKYP expression, the $\mathscr{H}_\infty$ norm of the NTF in the signal band is minimized to either below a target value (feasibility problem) or as low as possible (optimization problem), subject to any of the above stability constraints. With reference to the augmented system in Equation 2.7, the GKYP constraint is placed on the $r \rightarrow e$ channel that exposes the sensitivity function:

$$\min_{\lambda \in [\omega_l, \omega_h]} ||G_{er}(\lambda)||_\infty. \tag{3.23}$$

In DT designs, the Bode integral (3.6) combined with the performance goal sets the roll-off of the NTF. For CT designs, the performance goal and any stability constraint does not capture the effect of the quantizer sampling frequency because the S/H block was omitted in Section 1.2.2. This leads to designs with very low roll-off. To rectify this and force high roll-off, the constraint in 3.24 is added for CT designs. This forces high roll-off by reducing the $\mathscr{H}_\infty$ norm of the complementary sensitivity function (STF) just outside the signal band. The appearance of this constraint also shows evidence of the implicit antialiasing feature of CT loop filters.

$$||G_{yr}(j\omega)||_\infty \leq \gamma_T \quad \forall \omega \in [\omega_h, f_s] \tag{3.24}$$

Chapter 4

Optimization of Loop Filter Design

In this chapter, the model of the sigma delta modulator system developed in Chapter 2 is combined with the stability and performance expressions from Chapter 3 in a framework to synthesize a loop filter satisfying the desired criteria. Modulator design is done by solving a multiobjective optimization problem with a singular performance goal and one or more stability criteria applied to different channels of the system in Equation 2.7. The optimization framework unifies the expression for the GKYP, $\mathscr{H}_2$, and ℓ_1 norm LMIs for the augmented system. The expressions for optimizing each norm are presented in Sections 4.1 to 4.3 followed by a method to make the problem convex in Section 4.4.

4.1 GKYP Lemma

It is common in control systems to design based on constraints in the frequency domain. The KYP lemma establishes an equivalence between these frequency domain inequalities and LMI expressions on the state-space matrices of the system. Frequency domain inequalities defined with the KYP lemma are valid across all frequencies and this necessitates the use of weighting filters or frequency gridding to target a specific frequency band. The generalized KYP lemma allows criteria to be applied to specific sections of the system's Nyquist plot. In the design of sigma delta loop filters, the GKYP lemma will primarily be used as a performance goal by setting constraints on the signal band, but may also be used in the design of CT modulators and those satisfying the stability criteria presented in Section 3.1.1. The GKYP lemma is presented in Lemma 4.1.1 below.

Lemma 4.1.1 (GKYP lemma [40])**.** *Given state-space matrices $\mathscr{A} \in \mathbb{R}^{n\times n}$, $\mathscr{B}_p \in \mathbb{R}^{n\times 1}$, $\mathscr{C}_q \in \mathbb{R}^{1\times n}$, $\mathscr{D}_{qp} \in \mathbb{R}^{1\times 1}$ of system $G_{qp}(\lambda)$, frequency range $[\omega_l, \omega_h]$, and symmetric matrix variables $P, Q \in \mathbb{R}^{n\times n}$, the finite frequency condition:*

$$||G_{qp}(\lambda)||_\infty \leq \gamma_\infty \quad \omega_l \leq \lambda \leq \omega_h$$

holds if and only if $Q \geq 0$ and the LMI:

$$\begin{aligned} -\begin{bmatrix} \mathscr{A} & \mathscr{B}_p \\ I & 0 \end{bmatrix}^T (\Phi \oplus P + \Psi \oplus Q) \begin{bmatrix} \mathscr{A} & \mathscr{B}_p \\ I & 0 \end{bmatrix} + & \\ -\begin{bmatrix} \mathscr{C}_q & \mathscr{D}_{qp} \\ 0 & I \end{bmatrix}^T \begin{bmatrix} 1 & 0 \\ 0 & -\gamma_\infty \end{bmatrix} \begin{bmatrix} \mathscr{C}_q & \mathscr{D}_{qp} \\ 0 & I \end{bmatrix} \geq 0 & \quad (4.1) \end{aligned}$$

is satisfied, where $\oplus$ denotes the Kronecker product. For the CT case with low pass (LP) or band pass (BP) designs:

$$\Phi = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad \Psi = \begin{bmatrix} -1 & j\omega_c \\ -j\omega_c & -\omega_1\omega_h \end{bmatrix} \qquad (4.2)$$

while for the DT, LP or BP case:

$$\Phi = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \qquad \Psi = \begin{bmatrix} 0 & e^{j\omega_c} \\ e^{-j\omega_c} & -2\cos\omega_0 \end{bmatrix} \qquad (4.3)$$

where:

$$\omega_1 = \begin{cases} -\omega_h & \omega_l = 0 \\ \omega_l & otherwise \end{cases}, \qquad \omega_c = \frac{\omega_h + \omega_1}{2}, \qquad \omega_0 = \frac{\omega_h - \omega_1}{2}.$$

For the high pass (HP) case where $\omega_h = \infty$ (CT), Equation 4.2 is modified to:

$$\Psi = \begin{bmatrix} 1 & 0 \\ 0 & -\omega_1^2 \end{bmatrix}, \tag{4.4}$$

while for the HP case where $\omega_h = \pi$ (DT), Equation 4.3 is modified to:

$$\Psi = \begin{bmatrix} 0 & -1 \\ -1 & -2\cos\omega_1 \end{bmatrix}. \tag{4.5}$$

4.1.1 $\mathscr{H}_\infty$ Minimization Across All Frequencies

Often, a conventional $\mathscr{H}_\infty$ constraint across all frequency is desired, as is for the stability criterion from Section 3.1.2. In this case, Lemma 4.1.1 may be modified by eliminating Q and adding an additional non-negative definiteness constraint for stability:

$$Q = \mathbf{0} \qquad\qquad P \geq 0. \tag{4.6}$$

4.1.2 Positive Real Constraints

The GKYP lemma presented above is valid for placing a $\mathscr{H}_\infty$ norm constraint on the gain of a transfer function within a region of frequency space. The matrix:

$$\Pi = \begin{bmatrix} 1 & 0 \\ 0 & -\gamma_\infty \end{bmatrix} \tag{4.7}$$

in Equation 4.1 accomplishes this by defining a circle in the complex plane with radius $\left(\sqrt{\gamma_\infty}\right)^{-1}$ and centre at the origin [41, Lem. 1]. Because the gain of a transfer function is represented by the distance from a point along the Nyquist curve to the origin, this circle captures gain constraints by the parameter γ_∞ using the small gain theorem. This technique may be extended to arbitrary conical regions of the Nyquist diagram. Recall that the circle criterion for CT systems and Theorem 3.1.3 for DT systems placed constraints on the Nyquist diagram defined by a vertical line. The matrix Π may be modified to the following:

$$\Pi = \begin{bmatrix} 0 & 1 \\ 1 & 2\gamma \end{bmatrix}, \tag{4.8}$$

which defines a section of the complex plane divided by the line $\Re\{\lambda\} = \gamma$ to enable a positive real constraint on the transfer function. Thus, the circle criterion (Tsypkin criterion) may be realized with the GKYP lemma using this Π with $\gamma = \frac{1}{k_2}$ applied to the $r \to e$ sensitivity channel.

4.2 $\mathscr{H}_2$ Semidefinite Expression

The $\mathscr{H}_2$ norm used in the stability constraint from Section 3.1.4 can be minimized between two channels by solving a pair of inequalities with some similarities to Lemma 4.1.1.

Theorem 4.2.1. *Given state-space matrices $\mathscr{A} \in \mathbb{R}^{n\times n}$, $\mathscr{B}_p \in \mathbb{R}^{n\times 1}$, $\mathscr{C}_q \in \mathbb{R}^{1\times n}$, $\mathscr{D}_{qp} \in \mathbb{R}^{1\times 1}$ of system $G_{qp}(\lambda)$, symmetric matrix variable $P \in \mathbb{R}^{n\times n}$ and Φ from Equation 4.2 or 4.3, the $\mathscr{H}_2$ condition:*

$$||G_{qp}(\lambda)||_2 \leq \gamma_2$$

holds if and only if the following LMIs are satisfied:

$$-\begin{bmatrix} \mathscr{A} & \mathscr{B}_p \\ I & 0 \end{bmatrix}^T (\Phi \oplus P) \begin{bmatrix} \mathscr{A} & \mathscr{B}_p \\ I & 0 \end{bmatrix} + \begin{bmatrix} \mathbf{0} & 0 \\ 0 & 1 \end{bmatrix} \geq 0 \tag{4.9}$$

$$\begin{bmatrix} \gamma_2 & \mathscr{C}_q & \mathscr{D}_{qp} \\ \mathscr{C}_q^T & P & 0 \\ \mathscr{D}_{qp}^T & 0 & 1 \end{bmatrix} \geq 0. \tag{4.10}$$

Proof. Simplifying Equation 4.9 by multiplying outer factors and summing yields:

$$\begin{bmatrix} P\mathscr{A}+\mathscr{A}^TP & P\mathscr{B}_p \\ \mathscr{B}_p^TP & 1 \end{bmatrix} \geq 0 \tag{4.11}$$

for CT. Assuming $\mathscr{D}_{qp}=0$ as is necessary for the CT case, (4.10) simplifies to:

$$\begin{bmatrix} \gamma_2 & \mathscr{C}_q \\ \mathscr{C}_q^T & P \end{bmatrix} \geq 0. \tag{4.12}$$

Equations 4.11 and 4.12 comprise the well-known $\mathscr{H}_2$ LMI for CT systems [42, 43]. For the DT case, the simplification of Equation 4.9 along the same lines results in:

$$\begin{bmatrix} -\mathscr{A}^TP\mathscr{A}+P & -\mathscr{A}^TP\mathscr{B}_p \\ -\mathscr{B}_p^TP\mathscr{A} & -\mathscr{B}_p^TP\mathscr{B}_p+1 \end{bmatrix} \geq 0 \tag{4.13}$$

which can be manipulated into the form:

$$\begin{bmatrix} P & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} \mathscr{A}^T \\ \mathscr{B}_p^T \end{bmatrix} P \begin{bmatrix} \mathscr{A} & \mathscr{B}_p \end{bmatrix} \geq 0. \tag{4.14}$$

By Schur complement around P^{-1}, Equation 4.14 becomes:

$$\begin{bmatrix} P^{-1} & \mathscr{A} & \mathscr{B}_p \\ \mathscr{A}^T & P & 0 \\ \mathscr{B}_p^T & 0 & 1 \end{bmatrix} \geq 0. \tag{4.15}$$

Finally, after a congruent transformation of Equation 4.15 by $\operatorname{diag}(P,I,1)$, combining it with Equation 4.10 matches the well-known $\mathscr{H}_2$ LMI for DT systems [43]:

$$\begin{bmatrix} P & P\mathscr{A} & P\mathscr{B}_p \\ \mathscr{A}^TP & P & 0 \\ \mathscr{B}_pP & 0 & 1 \end{bmatrix}.$$

□

4.3 ℓ_1 Semidefinite Expression

The computation of the ℓ_1 norm is not generally possible in a semidefinite programming context. Instead, it is made tractable by minimizing the $\star$-norm, an upper bound on the ℓ_1 norm [44]. The $\star$-norm minimization is a set of two LMIs and one bilinear matrix inequality (BMI) with a scalar parameter that enters nonlinearly. The BMI can be solved by running a simple one-dimensional constrained minimization problem where the parameter α is minimized and the semidefinite program in Theorem 4.3.1 is solved for that α at each iteration.

Theorem 4.3.1. *Given state-space matrices $\mathscr{A} \in \mathbb{R}^{n\times n}$, $\mathscr{B}_p \in \mathbb{R}^{n\times 1}$, $\mathscr{C}_q \in \mathbb{R}^{1\times n}$, $\mathscr{D}_{qp} \in \mathbb{R}^{1\times 1}$ of system $G_{qp}(\lambda)$, symmetric matrix variable $P \in \mathbb{R}^{n\times n}$, auxiliary scalar variables $\mu \geq 0$, $\nu \geq 0$, and $\alpha \in (0,1)$, and Φ from Equation 4.2 or 4.3, the $\star$-norm condition:*

$$||G_{yx}(\lambda)||_\star \leq \gamma_\star$$

holds if and only if $P \geq 0$, the following LMI and BMI:

$$-\begin{bmatrix} \mathscr{A} & \mathscr{B}_p \\ I & 0 \end{bmatrix}^T \left(\left(\Phi + \begin{bmatrix} 0 & 0 \\ 0 & \alpha \end{bmatrix} \right) \oplus P \right) \begin{bmatrix} \mathscr{A} & \mathscr{B}_p \\ I & 0 \end{bmatrix} + \begin{bmatrix} \mathbf{0} & 0 \\ 0 & 1 \end{bmatrix} \geq 0 \tag{4.16}$$

$$\begin{bmatrix} \alpha P & 0 & \mathscr{C}_q \\ 0 & \mu - 1 & \mathscr{D}_{qp} \\ \mathscr{C}_q^T & \mathscr{D}_{qp}^T & \nu \end{bmatrix} \geq 0 \tag{4.17}$$

are satisfied for some α, and the following LMI is also satisfied:

$$\begin{bmatrix} \gamma_\star & \mu & \nu \\ \mu & 1 & 0 \\ \nu & 0 & 1 \end{bmatrix} \geq 0. \tag{4.18}$$

Proof. The proof proceeds in a similar way to that of Theorem 4.2.1 by transforming Equation 4.16 as was done with Equation 4.9. Then, combined with Equations 4.17 and 4.18, the matrix inequalities are in the form of the $\star$-norm semidefinite program reported in literature [45, 46]. □

4.4 Convexification

The LMIs presented in this Chapter are convex in optimization variable P when the state-space matrices of the system are known. For synthesis, Equations 4.1, 4.9, and 4.16 are non-convex as there are products between the state-space matrices $\mathscr{A}$, $\mathscr{B}_p$ and the optimization variable P. There are a number of approaches to make the problem convex in the design parameters. As described in Section 1.4, a common workaround is to define $S(\lambda)$ as an FIR filter. In FIR form, the $\mathscr{A}$ and $\mathscr{B}_p$ matrices are constant, thus $\mathscr{C}_q$ may be optimized in a convex fashion [15, 16]. A different approach is to use weighting filters where the "controller" can be extracted from the "plant" to restore convexity as is done with $\mathscr{H}_\infty$ control [12, 16]. The GKYP lemma may also be applied with convex constraints on the gain only. Another option is to work around the non-convexity with an iterative scheme [47]. For this thesis, the controller-plant approach was attempted using the extended controller parameterizations [48]. The GKYP gain constraints were examined although there wasn't any clear ways to enforce the closed-loop constraints from Section 2.2.1. In the end, the iterative workaround method resulted in better modulator designs. Before the iterative procedure is presented, a change of variables and congruent transformation must be done on the non-convex LMIs. As a first step, the number of product terms can be reduced by assuming the state space system in Equation 2.5 is in controllable canonical form (CCF) as shown below:

$$\dot{x} = \overbrace{\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix}}^{A_H} x + \overbrace{\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}}^{B_H} e \tag{4.19}$$

$$u = \underbrace{\begin{bmatrix} b_0 & b_1 & b_2 & \cdots & b_{n-1} \end{bmatrix}}_{C_H} x, \tag{4.20}$$

where $a_H \equiv \begin{bmatrix} a_0 & a_1 & a_2 & \cdots & a_{n-1} \end{bmatrix}^T$ and $b_H \equiv \begin{bmatrix} b_0 & b_1 & b_2 & \cdots & b_{n-1} \end{bmatrix}^T$

are defined for brevity. These vectors are the numerator and denominator coefficients of the loop filter in ascending powers of λ. Note the important fact that when Equation 2.5 is in CCF, sub-systems $G_{er}(\lambda)$, $G_{yr}(\lambda)$, and $G_{zw}(\lambda)$ from Equation 2.7 are in CCF as well, possibly after a trivial state-space transformation by $T = \frac{1}{-m_{21}} I_n$ for the latter to ensure that the lower element of $T\mathscr{B}_w$ is unity. In the following, we seek to design a_H and b_H by solving an optimization problem in different variables consisting of one or more of the constraints described in Chapter 3.

4.4.1 Change of Variables

Let $a \equiv a_H + m_{22} b_H$, the negative transpose of the lower row of $\mathscr{A}$. Define $b \equiv -m_{22} b_H$ as one of the following depending on the subsystem $G_{qp}(\lambda)$:

$$b = \begin{cases} -m_{22} b_H & G_{qp}(\lambda) = G_{er}(\lambda) \\ m_{12} b_H & G_{qp}(\lambda) = G_{zw}(\lambda) \\ m_{22} b_H & G_{qp}(\lambda) = G_{yr}(\lambda) \\ b_H & G_{qp}(\lambda) = G_{ur}(\lambda) \end{cases} \tag{4.21}$$

The vectors a and b are the denominator and numerator coefficients, respectively, of the closed-loop transfer function in descending powers of λ. The semidefinite program is redefined in terms of these variables to simplify nomenclature. The implementation concern of deriving a, b from the closed-loop state-space matrices of Equation 2.8 is given in Theorem 4.4.1.

Theorem 4.4.1. *Given state-space matrices $\mathscr{A} \in \mathbb{R}^{n\times n}$, $\mathscr{B}_p \in \mathbb{R}^{n\times 1}$, $\mathscr{C}_q \in \mathbb{R}^{1\times n}$, $\mathscr{D}_{qp} \in \mathbb{R}^{1\times 1}$ of system $G_{qp}(\lambda)$, a subsystem of Equation 2.8, it is assumed that there exists a transformation matrix T that places $G_{qp}(\lambda)$ into CCF. The vectors b and a are equal to:*

$$a = -T^{-T} \mathscr{A}^T T^T T \mathscr{B}_p \tag{4.22}$$

$$b = T^{-T} \mathscr{C}_q^T \tag{4.23}$$

Proof. For (4.22), because CCF is assumed, $T\mathscr{B}_p = \begin{bmatrix} 0_{1\times n-1} & 1 \end{bmatrix}^T$ extracts the negative transpose of the bottom row of the transformed $\mathscr{A}$, which is equal to the definition $a \equiv a_H + m_{22} b_H$ by Equation 4.19 and Equation 2.8.

Equation 4.23 is simply the transformed $\mathscr{C}_q$ which is different for each subsystem. Since the transformation T ensures the subsystems are in CCF, C_H may be replaced by b_H in the expression for $\mathscr{C}_q$ from Equation 2.7, obtaining the cases from Equation 4.21. □

4.4.2 Sensitivity Shaping

Addressing the performance goal in Section 3.3, the authors of [20, Th. 1] have shown that a congruent transformation of Equation 4.1 by the matrix:

$$\begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix} \tag{4.24}$$

on the left and its transpose on the right eliminates any products between a, b and P, Q, restoring linearity in the first summation term. This leaves only products between a and b in the second term of Equation 4.1. Simplifying and using a Schur complement results in only one non-convex term, that is aa^T in the upper-left block. The procedure in [20] is only applicable to shaping the sensitivity function $G_{er}(\lambda)$ because it assumes $\mathscr{D} = 1$. A full derivation that is valid for any $\mathscr{D}$, such as that encountered when solving Equation 3.12, is produced in Appendix A.1.

4.4.3 $\mathscr{H}_2$, ℓ_1 Optimization

The congruent transformation procedure from Section 4.4.2 does not depend on the centre expression in the relevant LMI (that may be a function of any of Φ, Ψ, P, Q) so it is applicable to Equations 4.9 and 4.16, which have the same outer factors, restoring linearity to the first summation term. The second term of both is the same. Equation 4.25 shows this second term with the congruent transformation from Equation 4.24 applied:

$$\begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{0} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix}^T = \begin{bmatrix} aa^T & a \\ a^T & 1 \end{bmatrix} \tag{4.25}$$

It is seen that, like Equation 4.1, the other semidefinite programs can undergo a change of variables to have the same, single nonlinear term aa^T. The full expression is given in Appendix A.2.

4.4.4 Iterative Procedure

The non-convex inequalities have now been converted into a form where one quadratic term exists. The solving of a quadratically constrained LMI is a difficult problem. Several methods of solving Equation A.1 and Equation A.8 were attempted but had poor results. These included using a general non-convex solver directly, using a rank-constrainted LMI solver, and using Shor's relaxation to linearize the problem. Instead, we use the iterative method from [47] which, for problems with simple non-convexities, appears to be similar to the method used in [20] except with an extra parameter to guarantee finite convergence.

In short, the iterative method proceeds following Algorithm 1, where the quadratic term is separated from the rest of the LMI by splitting the optimization problem into the form:

$$\begin{gathered} \min f(\lambda) \text{s.t.} \\ C_i(a, \lambda, \cdot) + Q(a) \geq 0 \quad \forall i \end{gathered} \tag{4.26}$$

where $f(\gamma)$ is the objective function, $C_i(a, \gamma, \cdot)$ is the ith convex LMI constraint, and $Q(a) = \begin{bmatrix} aa^T & 0 \\ 0 & 0 \end{bmatrix}$ is the quadratic part involving a. A maximum number of iterations `maxIter` is defined, as well as termination crtieria ε and an optional weight κ to penalize wandering in the neighbourhood of a solution.

The iterative LMI problems generated with this method were solved numerically using the YALMIP Toolbox for MATLAB [49] with the LMILAB solver [50] wrapped by the Nelder-Mead simplex algorithm to solve the ℓ_1 BMI (if applicable) [51]. Curiously, other semidefinite program (SDP) solvers seem to converge on inferior solutions and often encounter numerical problems. Parameters κ and ε were tuned as necessary by observing the convergence of a by iteration.

Algorithm 1 requires an feasible intialization in the form of a loop filter transfer function. In many cases, passing a simple transfer function such as:

Algorithm 1 Iterative convexification

```
procedure QMISOLVE($a_{init}$)
    $a_f = \text{CHECKFEAS}(a_{init})$
    $(a, b, \gamma) = \text{CVXITER}(a_f)$
    return $a, b, \gamma$
end procedure
function CHECKFEAS($a_{in}$)
    given starting value $a_{in}$, find feasible $a$ s.t.:        ▷ Convex feas. problem
        $C_i(a + a_{in}, \gamma, \cdot) + Q(a + a_{in}) - Q(a)\ \forall i$
    $a_{out} \leftarrow a + a_{in}$
    return $a_{out}$
end function
function CVXITER($a_{in}$)
    $k \leftarrow 0$
    $a^{(0)} \leftarrow a_{in}$
    while $k <$ maxIter and $\Delta_a > \varepsilon$ do
        solve:                                                    ▷ Convex opt. problem
            $\min f(\gamma) + \kappa ||a||_2^2$ s.t.
            $C_i(a + a^{(k)}, \gamma, \cdot) + Q(a + a^{(k)}) - Q(a)\ \forall i$
        $k \leftarrow k + 1$
        $a^{(k)} \leftarrow a^{(k-1)} + a$
        $\Delta_a \leftarrow ||a^{(k)} - a^{(k-1)}||_2$
    end while
    return $a^{(k)}, b, \gamma$
end function
```

$$H_1(z) = \frac{z^{n-1}}{z^n}$$

for the DT case, or:

$$H_1(s) = \frac{(s+1)^{n-1}}{(s+1)^n}$$

for the CT case results in convergence. For more difficult cases with higher order or more aggressive OSR, it may be necessary to use a more appropriate starting point such as the poles and zeros chosen from a specific region by the authors of [20, Fig. 2] or a LF derived from the `synthesizeNTF` function in the Delta Sigma Toolbox [5, Appx. B]. A third option explored is using the Shor convex

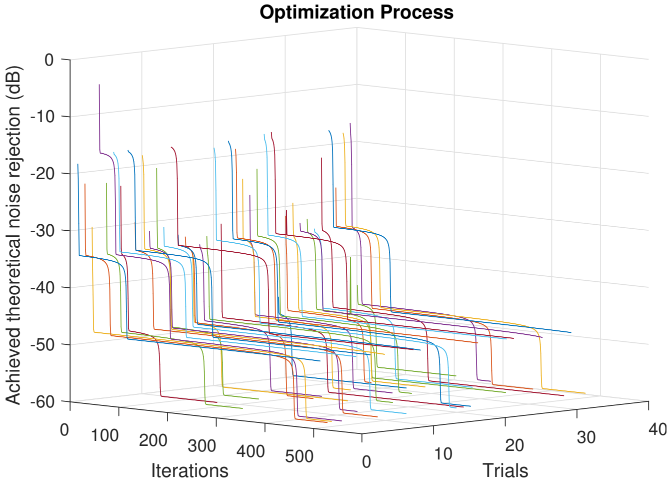

Figure 4.1: An example of the dependence of the iterative optimization scheme on initial conditions.

relaxation [52] to find an approximate convex starting point. This involves solving the optimization problem derived from that in Equation 4.26 but with an additional matrix variable and positive semidefinite constraint to try to force $\mathring{A} = aa^T$:

$$\begin{gathered}\min \operatorname{Tr}\{W\} \text{s.t.}\\ C_i(a, \lambda, \cdot) \geq 0 \quad \forall i\\ W = \begin{bmatrix} \mathring{A} & a \\ a^T & 1 \end{bmatrix} \geq 0\\ \mathring{A} = \mathring{A}^T.\end{gathered}$$

Figure 4.1 shows the an example of the achieved GKYP $\mathscr{H}_\infty$ norm in the sensitivity function signal band after 500 iterations of CvxIter from Algorithm 1 with initial condition poles randomly placed in the unit circle. It can be seen that the performance objective is minimized in "stages" corresponding to a pole-zero pair being optimized, increasing the apparent order of the system. If the algorithm terminates at a sub-optimal level, there often exists a pole-zero cancellation in the loop filter.

Chapter 5

Design Examples

The optimization framework developed in Chapter 4 may now be used to produce loop filters using various criteria from Chapter 3. In Sections 5.1 to 5.4, a 5th-order DT modulator with a 32 times OSR is designed for EEG recording applications with several different stability criteria. This application demands a LP design to capture EEG signals from the delta band (below 4 Hz) to the gamma band (up to 100 Hz) inclusive. Therefore, the modulator will be clocked at 6.4 kHz to avoid aliasing. In Section 5.5, a 3rd-order CT modulator with a 32 times OSR is shown to demonstrate the method applied to CT designs. This example is intended for audio applications, i.e., LP signals with Nyquist frequency 44.1 kHz.

5.1 Design Using $\mathscr{H}_\infty$ Stability Criterion

The $\mathscr{H}_\infty$ design procedure is done by solving the optimization problem in Equation 3.23 for performance while the constraint in Equation 3.7 promotes a stable design. Lemma 4.1.1 is used for the former while Lemma 4.1.1 combined with the auxiliary conditions in Equation 4.6 for the latter, which implicitly forces the NTF to be stable. For a Lee criterion of $\gamma_\infty = 1.5$, the optimization problem converges to the loop filter transfer function:

$$H_1(z) = \frac{0.799\left(z^2 - 1.59z + 0.657\right)\left(z^2 - 1.92z + 0.966\right)}{(z - 0.954)\left(z^2 - 1.95z + 0.953\right)\left(z^2 - 1.99z + 0.994\right)}.$$

The sensitivity function of this filter can be seen in Figure 5.1. Note that the

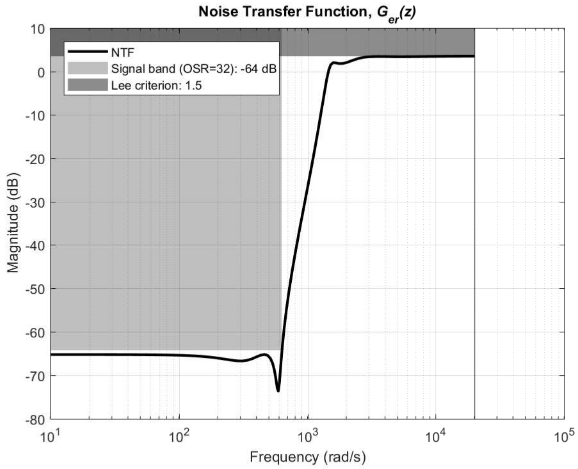

Figure 5.1: The sensitivity function of the design in Example 5.1. The dark shaded area represents the stability constraint and the light shaded area represents the achieved noise attenuation performance.

Lee criterion for stability is satisfied across all frequencies and the peak gain in the signal band has been minimized to −64 dB by the GKYP lemma. This compares favourably (in the $\mathscr{H}_\infty$ sense) to the design produced with the toolbox in [5, Appx. B], which has peak gain in the signal band of −55 dB[1].

Like most high-order designs using the Lee criterion, stability is conditional on input amplitude. A simulation of this can be seen in Figure 5.2, also performed with the Delta Sigma Toolbox. A peak signal-to-quantization-noise ratio (SQNR) of 86 dB at an input amplitude of 0.71 was achieved with a maximum stable input amplitude (MSIA) of 0.71 and a minimum resolvable input amplitude of −91 dB full scale (FS). The comparable toolbox design achieved a very similar peak SQNR but with a slightly better MSIA and minimum resolvable input amplitdue of 0.76 and −96 dB FS, respectively. The trade-off between stability and performance as the Lee criterion goal is changed is shown in Figure 5.3. For Lee criterion targets below 2.1, a feasible design is not found. An abrupt onset of instability is observed for designs with Lee criterion above 2.1.

[1] The Delta Sigma Toolbox command `synthesizeNTF(5, 32, 1, 1.5, 0)` was used to produce the transfer function used in this comparison.

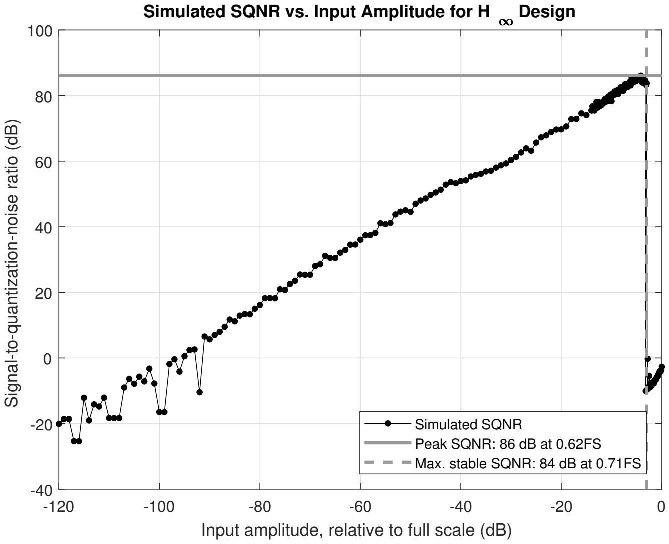

Figure 5.2: An SQNR plot of 247 simulations of the design in Example 5.1 to an input sinusoid of frequency 3.46×10^4 rad s^{-1} and varying amplitude to investigate its conditional stabiilty.

5.2 Design Using Root Locus Stability Criterion

The root locus design technique is done by solving the optimization problem in Equation 3.23 for performance while the constraint in Equation 3.12 is enforced for robustness against the quantizer gain. Similar to the previous example, Lemma 4.1.1 is applied to the $r \rightarrow e$ sensitivity channel and Lemma 4.1.1 along with conditions in Equation 4.6 to the $w \rightarrow z$ robustness channel. While a sufficient condition for stability would be that the root locus remains in the stable region for all positive quantizer gains K, this produces a very conservative design. Instead, the quantizer gain robustness criterion can be used to enhance the stable input range of Design 5.1. Instablility in sigma delta modulators is often associated with low quantizer gains. To improve stability, the lower bound of the quantizer gain, k_l, may be changed and the optimization problem solved. Thus k_l is a parameter that trades off performance and stability, which is shown in Figure 5.4. With some trial-and-error, $k_l = 0.1$ results in a modulator that is full-scale stable under simulation. The solver converges to the loop transfer function:

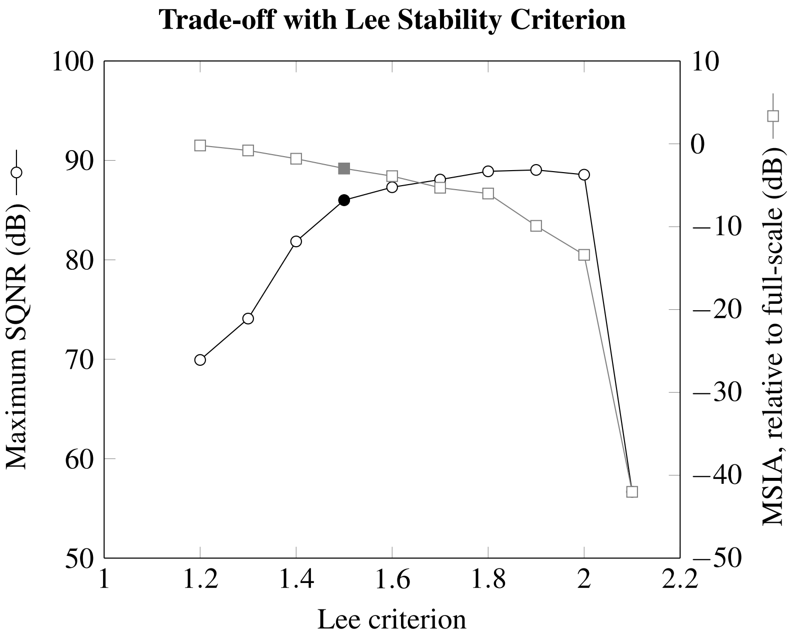

Figure 5.3: The performance (maximum simulated SQNR) and stability (simulated MSIA) achieved with the modulator design from Section 5.1 for different Lee criterion goals. The shaded markers indicate the selected design Lee criterion value.

$$H_1(z) = \frac{1.80\,(z-0.806)\,(z-0.641)\left(z^2-1.93z+0.949\right)}{(z-0.607)\,(z^2-1.94z+0.943)\,(z^2-1.98z+0.990)}.$$

The robustness and sensitivity channels are shown in Figure 5.5. The $\mathscr{H}_\infty$ norm of the $G_{zw}(z)$ transfer function is less than 1 for all frequencies, showing that the system is stable for all norm-bounded quantizer gains in the range $[k_l, k_h] = [0.1, \infty)$. The root locus is shown in Figure 5.6 confirms that this is the case. A simulation like done previously shows that system is empirically stable for input amplitudes up to full scale. As expected when stability is increased, the empirical peak SQNR is reduced to 66 dB with the minimum resolvable input amplitude at −52 dB FS.

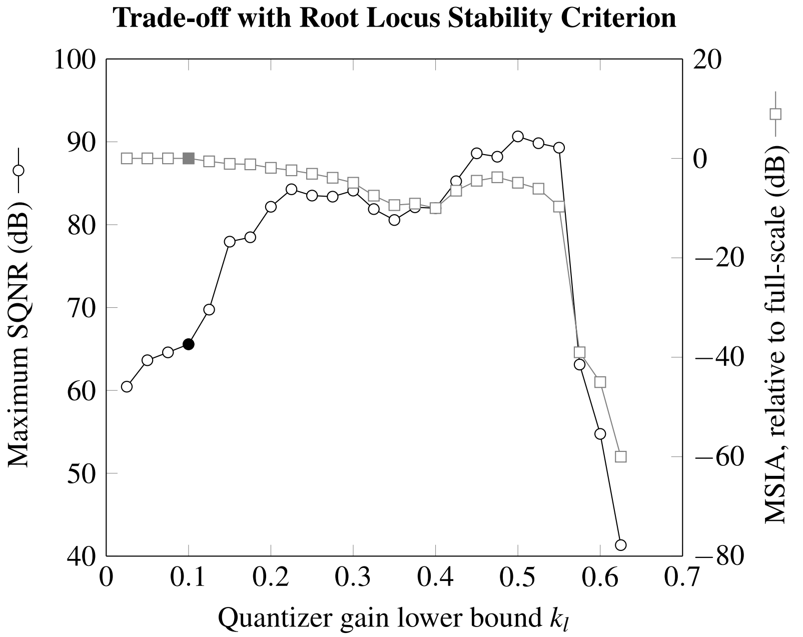

Figure 5.4: The performance (maximum simulated SQNR) and stability (simulated MSIA) achieved with the modulator design from Section 5.2 for different quantizer gain robustness goals. The shaded markers indicate the selected design k_l value. The root locus is entirely within the unit circle for $k_l \leq 0.075$.

5.3 Design Using $\mathscr{H}_2$ Stability Criterion

The $\mathscr{H}_2$ design technique is done by again solving the optimization problem in Equation 3.23 for performance while maintaining the stability constraint from Equation 3.15. The former uses Lemma 4.1.1 while the latter uses Theorem 4.2.1. In this example, the goal is to design a modulator for the same specifications as that from Section 5.1 but with slightly increased stability. One advantage of the $\mathscr{H}_2$ criterion is that there is a more systematic way to target a specific MSIA. For this example, the quantizer input signal is modelled with the Gaussian PDF. For a target MSIA of 0.90, the criterion is satisfied if $||G_{er}(z)||_2^2 < 1.29$. Using this constraint along with the performance optimization, the solver converges. Even after many iterations, the LF contains a pole-zero cancellation indicating that the optimization

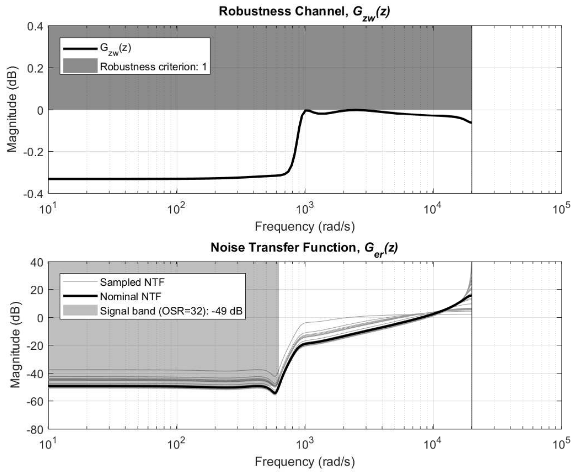

Figure 5.5: Upper: a frequency response plot of the robustness channel for the design in Example 5.2. Lower: the nominal sensitivity function of the same design along with sensitivity functions for randomly sampled quantizer gains with the achieved noise attenuation performance shaded.

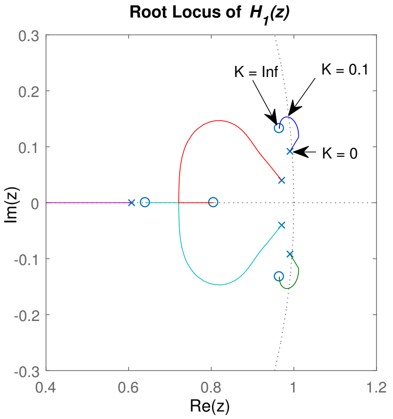

Figure 5.6: A subset of the complex plane showing the root locus of the filter from Example 5.2 across quantizer gains.

scheme was not able to find a feasible 5th-order design. After simplification, the loop filter transfer function is:

$$H_1(z) = \frac{0.536\,(z-0.690)\left(z^2-1.94z+0.967\right)}{(z^2-1.97z+0.968)\,(z^2-1.99z+0.995)}.$$

Computing the NTF gain for the optimum value of $K = 0.877$, we obtain $||G_{er}(z)||_2 = 1.28$, indicating that the 2-norm constraint was satisfied. Because this stability criteria is only an approximation, the empirical MSIA is 0.84. The peak SQNR was found to be 78 dB at an input amplitude of 0.73. The minimum resolvable input amplitude was found to be −77 dB FS. These measurements can be seen in Figure **??**.

5.4 Design Using ℓ_1 Stability Criterion

5.5 Continuous-Time Design

5.6 Summary of Design Examples

Chapter 6

Conclusions

Bibliography

[1] B. Blankertz, G. Dornhege, M. Krauledat, K. R. Müller, and G. Curio, “The non-invasive Berlin Brain-Computer Interface: Fast acquisition of effective performance in untrained subjects,” *Neuroimage*, vol. 37, no. 2, pp. 539–550, 2007. → pages x, 3

[2] L. Risbo, *Sigma Delta Modulators - Stability Analysis and Optimization*. Doctor of philosophy, Technical University of Denmark, 1994. → pages xi, 6, 16, 23, 24

[3] J. M. De La Rosa, “Sigma-delta modulators: Tutorial overview, design guide, and state-of-the-art survey,” *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 58, no. 1, pp. 1–21, 2011. → pages 2, 3

[4] M. Ortmanns and F. Gerfers, *Continuous-Time Sigma-Delta A/D Conversion*. 2005. → page 5

[5] R. Schreier and G. C. Temes, *Understanding Delta-Sigma Data Converters*, vol. 53. Wiley, 1997. → pages 5, 6, 7, 25, 27, 40, 43

[6] S. Hein and A. Zakhor, “On the Stability of Sigma Delta Modulators,” *IEEE Trans. Signal Process.*, vol. 41, no. 7, pp. 2322–2348, 1993. → page 6

[7] N. Wong and T.-s. Ng, “Fast detection of instability in sigma-delta modulators based on unstable embedded limit cycles,” *IEEE Trans. Circuits Syst. II*, vol. 51, no. 8, pp. 442–449, 2004. → page 6

[8] N. S. Sooch, “Gain Scaling of Oversampled Analog-to-Digital Converters,” 1989. → page 6

[9] S. M. Moussavi and B. H. Leung, “High-Order Single-Stage Single-Bit Oversampling A/D Converter Stabilized with Local Feedback Loops,” *IEEE Trans. Circuits Syst.*, vol. 41, no. 1, pp. 19–25, 1994. → page 6

[10] F. O. Eynde, G. M. Yin, and W. Sansen, “A CMOS Fourth-order 14b 500k-sample/s Sigma-delta ADC Converter,” 1991. → page 6

[11] J. Kenney and L. Carley, “CLANS: a high-level synthesis tool for high resolution data converters,” in *IEEE Int. Conf. Comput. Des. Dig. Tech. Pap.*, (Pittsburgh), pp. 496–499, 1988. → page 7

[12] A. Oberoi, *A Convex Optimization Approach to the Design of Multiobjective Discrete Time Systems*. Master of science, Rochester Institute of Technology, 2004. → pages 7, 8, 36

[13] S. Ohno and M. Rizwan Tariq, “Optimization of Noise Shaping Filter for Quantizer with Error Feedback,” *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 64, no. 4, pp. 918–930, 2017. → page 7

[14] M. M. Osqui and A. Megretski, “Semidefinite Programming in Analysis and Optimization of Performance of Sigma-Delta Modulators for Low Frequencies,” in *Proc. 2007 Am. Control Conf.*, no. 6, pp. 3582–3587, 2007. → pages 7, 8

[15] M. Nagahara and Y. Yamamoto, “Frequency Domain Min-Max Optimization of Noise-Shaping Delta-Sigma Modulators,” *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 1–12, 2012. → pages 7, 8, 28, 36

[16] M. R. Tariq and S. Ohno, “Unified LMI-based design of ΔΣ modulators,” *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, p. 29, 2016. → pages 7, 8, 36

[17] M. R. Tariq, S. Ohno, and M. Nagahara, “Synthesis of IIR error feedback filters for ΔΣ modulators using approximation,” in *2016 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA 2016*, (Jeju), pp. 7–11, 2017. → page 8

[18] M. S. Derpich, E. I. Silva, D. E. Quevedo, and G. C. Goodwin, “On optimal perfect reconstruction feedback quantizers,” *IEEE Trans. Signal Process.*, vol. 56, pp. 3871–3890, aug 2008. → page 8

[19] S. Callegari and F. Bizzarri, “Optimal design of the noise transfer function of ΔΣ modulators: IIR strategies, FIR strategies, FIR strategies with preassigned poles,” *Signal Processing*, vol. 114, pp. 117–130, 2015. → page 8

[20] X. Li, C. Yu, and H. Gao, “Design of delta-sigma modulators via generalized Kalman-Yakubovich-Popov lemma,” *Automatica*, vol. 50, no. 10, pp. 2700–2708, 2014. → pages 8, 38, 39, 40, 55

[21] H. K. Khalil, *Nonlinear Systems*. Upper Saddle River, NJ: Prentice Hall, 3 ed., 2002. → pages 17, 18

[22] Y. Z. Tsypkin, “Relay control systems,” 1984. → page 19

[23] K. C. H. Chao, S. Nadeem, W. L. Lee, and C. G. Sodini, “A Higher Order Topology for Interpolative Modulators for Oversampling A/D Converters,” *IEEE Trans. Circuits Syst.*, vol. 37, no. 3, pp. 309–318, 1990. → page 19

[24] Y. Zhao and V. Gupta, “A Bode-Like Integral for Discrete Linear Time-Periodic Systems,” *IEEE Trans. Automat. Contr.*, vol. 60, no. 9, pp. 2494–2499, 2015. → page 19

[25] R. Schreier, “An Empirical Study of High-Order Single-Bit Delta-Sigma Modulators,” *IEEE Trans. Circuits Syst. II Analog Digit. Signal Process.*, vol. 40, no. 8, pp. 461–466, 1993. → page 20

[26] M. Neitola, “Lee’s Rule Extended,” *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 64, no. 4, pp. 382–386, 2017. → page 20

[27] J. H. Taylor, “Describing functions,” *Electr. Eng. Encycl.*, no. April, pp. 1–35, 1999. → page 20

[28] J. A. E. P. Engelen, *Stability analysis and design of bandpass sigma delta modulators*. 1999. → page 22

[29] J. A. Van Engelen, R. J. Van De Plassche, E. Stikvoort, and A. G. Venes, “Sixth-order continuous-time bandpass sigma-delta modulator for digital radio IF,” *IEEE J. Solid-State Circuits*, vol. 34, no. 12, pp. 1753–1764, 1999. → page 22

[30] C.-c. Yang, K.-d. Chen, W.-C. Wang, and T.-h. Kuo, “Transfer function design of stable high-order sigma-delta modulators with root locus inside unit circle,” in *Proceedings. IEEE Asia-Pacific Conf. ASIC*, pp. 5–8, 2002. → page 22

[31] T.-H. Kuo, C.-C. Yang, K.-D. Chen, and W. C. Wang, “Design Method for High-Order Sigma Delta Modulator Stabilized by Departure Angles Designed to Keep Root Loci in Unit Circle,” *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 53, no. 10, pp. 1083–1087, 2006.

[32] K. Kang, *Simulation, and Overload and Stability Analysis of Continuous Time Sigma Delta Modulator*. PhD thesis, University of Nevada, 2014. → page 22

[33] R. M. Gray, *Source Coding Theory*. Stanford, CA: Kluwer Academic Publishers, 1 ed., 1990. → page 22

[34] S. Ardalan and J. Paulos, "An analysis of nonlinear behavior in delta - sigma modulators," *IEEE Trans. Circuits Syst.*, vol. 34, no. 6, pp. 593–603, 1987. → page 23

[35] D. Anastassiou, "Error Diffusion Coding for A/D Conversion," *IEEE Trans. Circuits Syst.*, vol. 36, no. 9, pp. 1175–1186, 1989. → page 26

[36] M. Yagyu and A. Nishihara, "Stability Analysis of 1-Bit Σ Modulators by Covering State Vector Transition with Hyper Cube for Specified Input Peak Amplitudes and Auto-Correlations," *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, vol. E87-A, no. 8, pp. 1855–1862, 2004. → page 27

[37] P. Steiner and W. Yang, "A Framework For Analysis of High-Order Sigma-Delta Modulators," *IEEE Trans. Circuits Syst. II Analog Digit. Signal Process.*, vol. 44, no. 1, pp. 1–10, 1997. → page 28

[38] P. Steiner and W. Yang, "Stability of High Order Sigma-Delta Modulators," *IEEE Int. Symp. Circuits Syst.*, vol. 3, no. 2, pp. 52–55, 1996. → page 28

[39] V. Mladenov, "Stability Analysis and Limit Cycles of High Order Sigma-Delta Modulators," in *Sel. Top. Nonlinear Dyn. Theor. Electr. Eng.* (K. Kyamakya, W. A. Halang, W. Mathis, J. C. Chedjou, and Z. Li, eds.), vol. 483, ch. 17, pp. 305–327, Berlin: Springer, 2013. → page 28

[40] T. Iwasaki and S. Hara, "Generalized KYP Lemma: Unified Frequency Domain Inequalities with Design Applications," *IEEE Trans. Autom. Control*, vol. 50, no. 1, pp. 41–59, 2005. → page 31

[41] T. Iwasaki and S. Hara, "Generalized KYP Lemma: Unified Characterization of Frequency Domain Inequalities with Applications to System Design." 2003. → page 32

[42] C. W. Scherer, P. Gahinet, and M. Chilali, "Multiobjective output-feedback control via LMI optimization," *IEEE Trans. Automat. Contr.*, vol. 42, no. 7, pp. 896–911, 1997. → page 34

[43] I. Masubuchi, A. Ohara, and N. Suda, "LMI-based controller synthesis: a unified formulation and solution," *Robust Nonlinear Control*, vol. 8, no. 9, pp. 669–686, 1998. → page 34

[44] S. Venkatesh and M. Dahleh, "Does star norm capture L_1 norm?," in *Proc. Am. Control Conf. (ACC), 1995*, (Seattle, WA), pp. 944–945, 1995. → page 35

[45] J. Bu and M. Sznaier, "Linear matrix inequality approach to synthesizing low-order suboptimal mixed l1/Hp controllers," *Automatica*, vol. 36, no. 7, pp. 957–963, 2000. → page 35

[46] A. Oberoi and J. C. Cockburn, "A simplified LMI approach to l1 Controller Design," in *Proc. 2005 Am. Control Conf.*, (Portland), pp. 1788–1792, 2005. → page 35

[47] S. L. Shishkin, "Optimization under non-convex Quadratic Matrix Inequality constraints with application to design of optimal sparse controller," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 10754–10759, 2017. → pages 36, 39

[48] M. C. De Oliveira, J. C. Geromel, and J. Bernussou, "Extended H2 and H norm characterizations and controller parameterizations for discrete-time systems," *Int. J. Control*, vol. 75, no. 9, pp. 666–679, 2002. → page 36

[49] J. Löfberg, "YALMIP: A Toolbox for Modeling and Optimization in MATLAB," in *Proc. CACSD Conf.*, (Taipei, Taiwan), 2004. → page 39

[50] P. Gahinet and A. J. Laub, *LMI Control Toolbox For Use with MATLAB*. 1 ed., 1995. → page 39

[51] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Comput. J.*, vol. 7, pp. 308–313, 1965. → page 39

[52] S. Boyd and L. Vandenberghe, "Semidefinite Programming Relaxations of Non-Convex Problems in Control and Combinatorial Optimization," *Commun. Comput. Control Signal Process. A Tribut. to Thomas Kailath*, pp. 279–288, 1997. → page 41

Appendix A

Derivation of Matrix Inequalities with One Quadratic Term

A.1 Derivation of GKYP Inequality with Arbitrary $\mathscr{D}$

Theorem A.1.1. *Equation 4.1 from Section 4.1 is equivalent to the following:*

$$\begin{bmatrix} -\Xi_{11} + aa^T & -\Xi_{12} + a & -\mathscr{C}_q^T - a\mathscr{D}_{qp}^T \\ -\Xi_{12}^T + a^T & -\Xi_{22} + 1 & -\mathscr{D}_{qp}^T \\ -\mathscr{C}_q - a^T \mathscr{D}_{qp} & -\mathscr{D}_{qp} & \gamma_\infty \end{bmatrix} \geq 0 \tag{A.1}$$

where (A.1) contains just one nonlinear term in variable a, and:

$$\begin{bmatrix} \Xi_{11} & \Xi_{12} \\ \Xi_{12}^T & \Xi_{22} \end{bmatrix} = \begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathscr{A} & \mathscr{B}_p \\ I & 0 \end{bmatrix}^T \left(\Phi \oplus P_\gamma + \Psi \oplus Q_\gamma\right) \begin{bmatrix} \mathscr{A} & \mathscr{B}_p \\ I & 0 \end{bmatrix} \begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix}^T$$

$$P_\gamma = \gamma_\infty^{-1} P \qquad\qquad Q_\gamma = \gamma_\infty^{-1} Q. \tag{A.2}$$

Proof. Starting from Equation 4.1, the procedure mentioned in Section 4.4.2 is followed to eliminate non-convex products in the first term of the LMI [20, Th. 1]:

$$-\begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \mathscr{A} & \mathscr{B}_p \\ I & 0 \end{bmatrix}^T f(\Phi,\Psi,P,Q)\begin{bmatrix} \mathscr{A} & \mathscr{B}_p \\ I & 0 \end{bmatrix}\begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix}^T + \ldots \geq 0. \tag{A.3}$$

Let the notation Ξ_{ij} be used for the linear part:

$$-\begin{bmatrix} \Xi_{11} & \Xi_{12} \\ \Xi_{12}^T & \Xi_{22} \end{bmatrix} + \ldots \geq 0. \tag{A.4}$$

Equation A.4 may undergo a congruent transformation by $\gamma_\infty^{-\frac{1}{2}} I$ introducing a commutable factor of γ_∞^{-1} to every element. For the first summation term, the factor is absorbed into Q and P with the redefinition from Equation A.9 yielding:

$$-\begin{bmatrix} \Xi_{11} & \Xi_{12} \\ \Xi_{12}^T & \Xi_{22} \end{bmatrix} - \begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \mathscr{C}_q & \mathscr{D}_{qp} \\ 0 & I \end{bmatrix}^T \begin{bmatrix} \gamma_\infty^{-1} & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} \mathscr{C}_y & \mathscr{D}_{qp} \\ 0 & I \end{bmatrix}\begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix}^T \geq 0. \tag{A.5}$$

Multiplying the inner factors in the second term of Equation A.5 leads to:

$$-\begin{bmatrix} \Xi_{11} & \Xi_{12} \\ \Xi_{12}^T & \Xi_{22} \end{bmatrix}^T - \begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \gamma_\infty^{-1}\mathscr{C}_q^T\mathscr{C}_q & \gamma_\infty^{-1}\mathscr{C}_q^T\mathscr{D}_{qp} \\ \gamma_\infty^{-1}\mathscr{D}_{qp}^T\mathscr{C}_q & \gamma_\infty^{-1}\mathscr{D}_{qp}^T\mathscr{D}_{qp} - 1 \end{bmatrix}\begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix}^T \geq 0$$

which can be expanded into:

$$-\begin{bmatrix} \Xi_{11} & \Xi_{12} \\ \Xi_{12}^T & \Xi_{22} \end{bmatrix} - \begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix}\begin{bmatrix} I & a\mathscr{D}_{qp}^T \\ 0 & \mathscr{D}_{qp}^T \end{bmatrix}\begin{bmatrix} \mathscr{C}_q^T \\ 1 \end{bmatrix}\gamma_\infty^{-1}\begin{bmatrix} \mathscr{C}_q^T \\ 1 \end{bmatrix}^T\begin{bmatrix} I & a\mathscr{D}_{qp}^T \\ 0 & \mathscr{D}_{qp}^T \end{bmatrix}^T\begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix}^T + \\ + \begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix}^T \geq 0. \tag{A.6}$$

The 3 outer factors multiplied with γ_∞^{-1} in the middle term of Equation A.6 are then combined together and the last summation term is also multiplied through, resulting in the following:

$$-\begin{bmatrix}\Xi_{11} & \Xi_{12}\\ \Xi_{12}^T & \Xi_{22}\end{bmatrix}-\begin{bmatrix}\mathscr{C}_q^T+a\mathscr{D}_{qp}^T\\ \mathscr{D}_{qp}^T\end{bmatrix}\gamma_\infty^{-1}\begin{bmatrix}\mathscr{C}_q^T+a\mathscr{D}_{qp}^T\\ \mathscr{D}_{qp}^T\end{bmatrix}^T+\begin{bmatrix}aa^T & a\\ a^T & 1\end{bmatrix}\geq 0. \tag{A.7}$$

The last summation term of Equation A.7 is then added with the linear part Ξ. Because $\gamma_\infty > 0 \leftrightarrow \gamma_\infty^{-1} > 0$, a Schur complement taken around γ_∞ allows Equation A.7 to be written as the single matrix inequality shown in Equation A.1.

□

A.2 Derivation of $\mathscr{H}_2$ and ℓ_1 Inequalities

Theorem A.2.1. *Equation 4.9 from Section 4.2 and Equation 4.16 from Section 4.3 are equivalent to the following:*

$$\begin{bmatrix}-\Xi_{11}+aa^T & -\Xi_{12}+a\\ -\Xi_{12}^T+a^T & -\Xi_{22}+1\end{bmatrix}\geq 0, \tag{A.8}$$

where Equation A.8 contains just one nonlinear term in variable a, and:

$$\begin{bmatrix}\Xi_{11} & \Xi_{12}\\ \Xi_{12}^T & \Xi_{22}\end{bmatrix}=\begin{bmatrix}I & a\\ 0 & 1\end{bmatrix}\begin{bmatrix}\mathscr{A} & \mathscr{B}_p\\ I & 0\end{bmatrix}^T f\left(\Phi,P_\gamma,\alpha\right)\begin{bmatrix}\mathscr{A} & \mathscr{B}_p\\ I & 0\end{bmatrix}\begin{bmatrix}I & a\\ 0 & 1\end{bmatrix}^T$$

$$f\left(\Phi,P_\gamma,\alpha\right)=\begin{cases}\Phi\oplus P_\gamma & \textit{for the } \mathscr{H}_2 \textit{ case}\\ \left(\Phi+\begin{bmatrix}0 & 0\\ 0 & \alpha\end{bmatrix}\right)\oplus P_\gamma & \textit{for the } \ell_1 \textit{ case}\end{cases} \tag{A.9}$$

$$P_\gamma=\gamma_\infty^{-1}P. \tag{A.10}$$

Proof. Starting from either Equation 4.9 or Equation 4.16, the procedure mentioned in Section 4.4.3 is followed to eliminate non-convex products in the first term of the LMI independent of $f\left(\Phi,P_\gamma,\alpha\right)$. The second summation term is the same in both LMIs and simplifies to Equation 4.25. Combining these, the matrix inequality from Equation A.8 is produced. □