**Noname manuscript No.**
(will be inserted by the editor)

# An Optimization Framework for the Design of Noise Shaping Loop Filters with Improved Stability Properties

**Brett C. Hannigan · Christian L. Petersen · A. Martin Mallinson · Guy A. Dumont**

Received: date / Accepted: date

**Abstract** A framework using semidefinite programming is proposed to enable the design of sigma delta modulator loop filters at the transfer function level. Both continuous-time and discrete-time, low-pass and band-pass designs are supported. For performance, we use the recently popularized Generalized Kalman-Yakubovič-Popov (GKYP) lemma to place constraints on the $\mathcal{H}_\infty$ norm of the noise transfer function (NTF) in the frequency band of interest. We expand the approach to incorporate common stability criteria in the form of $\mathcal{H}_2$ and $\ell_1$ norm NTF constraints. Furthering the discussion of stability, we

B. C. Hannigan
School of Biomedical Engineering
The University of British Columbia
251-2222 Health Sciences Mall
Vancouver, B.C., Canada V6T 1Z3
E-mail: bch@alumni.ubc.ca
ORCiD: 0000-0001-5778-1686

C. L. Petersen
ESS Technology Inc.
601-1726 Dolphin Ave.
Kelowna, B.C., Canada V1Y 9R9
E-mail: christian.petersen@esstech.com
ORCiD: 0000-0002-4840-275X

A. M. Mallinson
SiliconIntervention Inc.
403-460 Doyle Ave.
Kelowna, B.C., Canada V1Y 0C2
E-mail: martin.mallinson@siliconintervention.com

G. A. Dumont
Department of Electrical and Computer Engineering,
The University of British Columbia
3023-2332 Main Mall
Vancouver, B.C., Canada V6T 1Z4
E-mail: guyd@ece.ubc.ca
ORCiD: 0000-0003-2048-4391

introduce techniques from control systems to improve the robustness of the feedback system over a range of quantizer gains. The performance-stability trade-off is examined using this framework and motivated by simulation results.

**Keywords** sigma delta modulation · semidefinite programming · noise transfer function · generalized KYP lemma · analog/digital conversion

## 1 Introduction

Sigma delta modulators ($\Sigma\Delta$Ms) are nonlinear feedback systems containing a noise shaping filter and coarse quantizer element. Applied as an A/D converter, the systems operate on an oversampled input and produce a discrete-time, sampled value output. The feedback loop contains a noise shaping loop filter that pushes the error introduced by quantization out of the signal band, where it can be removed by a digital decimation filter outside the loop. The sigma delta architecture is widely used to digitize signals with moderate frequency content because of high resolution and reliance on less expensive digital circuitry rather than precision analog components. However, the presence of a nonlinearity in the feedback system makes analysis difficult and higher order systems are prone to instability.

The design of the loop filter transfer function may be done in many ways. Often, a linearized model is used, where the nonlinear quantizer is replaced by a fixed gain and an additive "quantization noise" signal. A loop filter of just one or two pure integrators is provably stable for dc inputs with magnitude less than one [29]. For higher orders, a common approach is to design a prototype noise transfer function (NTF), equivalent to the sensitivity function $S(\lambda)$ of the linearized model. The popular Delta Sigma Toolbox for MATLAB [29] uses the Chebyshev type II filter as an NTF prototype where the stop-band attenuation is related to the performance and the peak out-of-band gain is a proxy for instability.

Optimization techniques have been used in place of prototype filters to generate suitable noise transfer functions. For example, the CLANS approach assumes the quantization error can be represented as white noise, then uses nonlinear optimization to minimize the integral of this signal in the pass-band. [11]. Genetic algorithms have been used to design continuous-time sigma delta modulators with a combination of linear approximations and simulations [19]. Using the linear matrix inequality (LMI) methods from $\mathcal{H}_\infty$ control, one can define the quantizer as a very simple feedthrough plant and augment it with weighting filters with desirable noise rejection properties. The loop filter is then designed as an optimal controller for performance and stability [21]. However, the system is bound to the order of the augmented plant and this method relies on the designer to select the weighting filters.

The Generalized Kalman-Yakubovič-Popov (GKYP) lemma provides a way to optimize noise rejection over a finite frequency interval, eliminating the need for weighting filters. Unfortunately, the problem becomes non-convex if both poles and zeros are to be optimized. As a way around this, the poles may be fixed [24] or a finite impulse response (FIR) loop filter may be assumed [20,31], which is a sub-optimal choice [6]. Alternatively, iterative methods have been shown to provide a workaround for $\mathcal{H}_\infty$ minimization [15]. GKYP optimization can also be carried out to limit the reduction in performance resulting from quantization of the filter coefficients in the hardware implementation of the loop filter [4].

In Section 2 of this paper, we outline the structure of a general $\Sigma\Delta$M and introduce several stability criteria with varying levels of robustness. In Section 3, we introduce a semidefinite programming (SDP) framework used to optimize performance of the modulator under these stability criteria. In Section 4, simulation results are shown to compare modulator performance under different stability conditions. This paper is concerned with sigma delta A/D converters with single bit quantization, but would be easily generalizable to D/A designs and those with multi-bit quantization. Most of the paper is focused on the discrete-time ($\lambda = z$) systems used in switched capacitor designs but the framework also works with continuous-time ($\lambda = s$) systems with the caveat that many of the stability criteria are no longer valid. The frequency range of interest is from dc to the clock frequency driving the sample-and-hold block, located at the input for discrete-time modulators and in the loop for continuous-time modulators. The signal band is restricted to a small fraction of the system sampling frequency and expressed as the oversampling ratio (OSR). Using the proposed method, discrete-time modulators can be designed to a given OSR then scaled to any necessary sampling frequency. Continuous-time designs require the inclusion of constrains around the signal band as seen in Section 4.5.

## 2 Background

### 2.1 General Sigma Delta Modulator Model

In its general case, the block diagram of the sigma delta A/D converter is shown in Figure 1. The two input, one output loop filter $H(\lambda)$ may be manipulated into a conventional, negative feedback loop filter $H_1(\lambda)$ and an input pre-filter $H_0(\lambda)$ operating on input signal $r$ without loss of generality. The pre-filter may be used to shape the signal transfer function (STF) to unity, and can be neglected in this analysis to focus on the design of $H_1(\lambda)$. The nonlinear quantizer has been modelled by a variable gain $K$ and the addition of fictitious quantization noise $d$ that is summed with the output of the filter to produce output signal $y$, which is passed to the digital decimation filters.

For optimization, we are interested in placing constraints on the closed loop sensitivity function $S(\lambda)$ and possibly on the system robustness to the uncertain quantizer gain. For the former, we introduce a performance channel $e$ at the feedback error, which is equivalent to the NTF shown in Figure 1. For the latter, we model $K$ as a multiplicative uncertainty and extract the

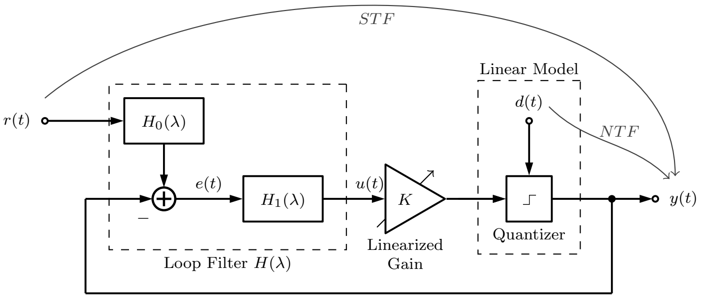

**Fig. 1** The general linearized $\Sigma\Delta$M block diagram with variable gain and additive quantization noise signal (see Section 2.1).

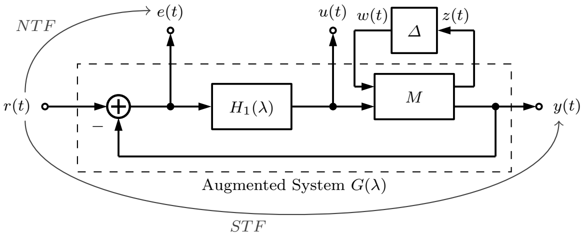

**Fig. 2** The augmented system is derived from Figure 1 by setting $H_0(\lambda) = 1$, taking the LFT of the uncertain gain, extracting the signals of interest, and writing the closed-loop equations.

norm-bounded uncertain block $\Delta$ using the upper linear fractional transform (LFT), as is common in robust control problems. The input $w$ and output $z$ encompass a robustness channel and matrix $M_{2\times 2}$ is a constant gain block.

From this model, an augmented state-space model is derived. Let the loop filter $H_1(\lambda)$ be an order $n$ strictly proper rational transfer function of the form shown in (1), which has an equivalent state-space representation (2). Naturally, the state-space feedthrough matrix $D_H$ must be zero to impose closed-loop realizability.

$$\begin{aligned} H_1(\lambda) &= \frac{b_{n-1}\lambda^{n-1} + b_{n-2}\lambda^{n-2} + \ldots + b_1\lambda + b_0}{\lambda^n + a_{n-1}\lambda^{n-1} + a_{n-2}\lambda^{n-2} + \ldots + a_1 z\lambda + a_0} \qquad (1) \\ &= C_H(\lambda I - A_H)^{-1}B_H \qquad (2) \end{aligned}$$

Taking $w(t)$, $r(t)$ as inputs and $z(t)$, $e(t)$, $u(t)$, $y(t)$ as outputs, the augmented system $G(\lambda)$ from Figure 2 may be shown in matrix form in (3), where $m_{ij}$ is element $(i, j)$ of gain matrix $M$. The notation $G_{qp}(\lambda)$ indicates the transfer function of $G(\lambda)$ from some input channel $p(t)$ to some output channel $q(t)$. The state-space matrices of $G$ are denoted in (4).

$$G : \begin{bmatrix} \lambda \mathbf{x}(t) \\ z(t) \\ e(t) \\ u(t) \\ y(t) \end{bmatrix} = \left[\begin{array}{c|cc} A_H - m_{22} B_H C_H & -m_{21} B_H & B_H \\ \hline m_{12} C_H & m_{11} & 0 \\ -m_{22} C_H & -m_{21} & 1 \\ C_H & 0 & 0 \\ m_{22} C_H & m_{21} & 0 \end{array}\right] \begin{bmatrix} \mathbf{x}(t) \\ w(t) \\ r(t) \end{bmatrix} \tag{3}$$

$$= \left[\begin{array}{c|cc} \mathcal{A} & \mathcal{B}_w & \mathcal{B}_r \\ \hline \mathcal{C}_z & \mathcal{D}_{zw} & \mathcal{D}_{zr} \\ \mathcal{C}_e & \mathcal{D}_{ew} & \mathcal{D}_{er} \\ \mathcal{C}_u & \mathcal{D}_{uw} & \mathcal{D}_{ur} \\ \mathcal{C}_y & \mathcal{D}_{yw} & \mathcal{D}_{yr} \end{array}\right] \begin{bmatrix} \mathbf{x}(t) \\ w(t) \\ r(t) \end{bmatrix} \tag{4}$$

## 2.2 Performance Goal and Stability Criteria

Modulator design is done by solving a multiobjective optimization problem with a single performance goal and one or more of the four stability criteria discussed in this section. The criteria range from heuristics to sufficient conditions. We introduce these criteria and show how they can be interpreted as norm bounds on the augmented system (4), which may be solved using the framework in Section 3.

### *2.2.1 Performance Goal*

The common performance goal in the optimization process is the $\mathcal{H}_\infty$ norm of the noise transfer function. We specify a frequency range of interest $[\omega_l, \omega_h]$ centred at the signal band with width inversely proportional to the OSR of the modulator. Using the GKYP LMI expression, the $\mathcal{H}_\infty$ norm of the NTF in the signal band is minimized, either below a target value (feasibility problem) or as low as possible (optimization problem). With reference to the state space system (3), the GKYP constraint is placed on the $r \to e$ channel as shown in Problem 1.

**Problem 1** *Given frequency range of the signal band $[\omega_l, \omega_h]$, find the following for stable $H_1(\lambda)$:*

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

### *2.2.2 $\mathcal{H}_\infty$ Stability Criterion*

Lee's rule is a heuristic predictor of stability which states that a modulator is likely to be stable if the NTF peak out-of-band gain, or $\mathcal{H}_\infty$ norm of the sensitivity function, does not exceed a benchmark value [5]. The criterion is not necessary nor sufficient for stability and must be verified with extensive simulations. Generally, $||S(z)||_\infty = 2$ is used, but this has been found to be conservative for low-order and insufficient for high-order designs [28]. Lee's rule is valid for discrete-time designs only, but has been used extensively and is easily included as part of an optimization problem. In this framework, the Lee's rule heuristic may be applied with an $\mathcal{H}_\infty$ constraint on the $r \to e$ channel as shown in Problem 2 by using the GKYP lemma with an infinite frequency interval.

**Problem 2** *For given positive $\gamma_\infty$, find $H_1(\lambda)$ such that:*

$$||G_{er}(\lambda)||_\infty < \gamma_\infty.$$

*2.2.3 Root Locus Stability*

With single-bit quantization, the instantaneous quantizer gain $K$ is in the interval $[k_0, k_1] = [1/||u||_\infty, \infty]$ with nominal value $k_0$. One method to design stable $\Sigma\Delta$M loop filters is to position the poles and zeros such that the root locus remains in the stable region of the complex plane when sweeping through this interval [32,13,10]. Using our optimization framework, the nonlinear quantizer gain is modelled as a multiplicative parametric uncertainty. Using the definition of the upper LFT [34, Def. 10.1], uncertain gain $K$ is split into fixed certain gain matrix $M$ and norm-bounded uncertain part $\Delta$:

$$\begin{aligned} K &= \mathcal{F}_U\{M, \Delta\} \quad ||\Delta||_\infty < 1 \\ &= m_{22} + m_{21}\Delta\left(1 - m_{11}\Delta\right)^{-1} m_{12} \end{aligned}$$

To find the entries of matrix $M$ for a range of gains $K \in [k_l, k_h]$, we use the fact that $\mathcal{F}_U\{M,1\} = k_h$, $\mathcal{F}_U\{M,0\} = k_0$, and $\mathcal{F}_U\{M,-1\} = k_l$ that follows from the normalized nature of $\Delta$. Equation 5 shows $M$ found by setting $m_{21} = 1$ and solving the system of equations.

$$M = \begin{bmatrix} \frac{k_h - 2k_0 + k_l}{k_h - k_l} & \frac{-2(k_0 - k_h)(k_0 - k_l)}{k_h - k_l} \\ 1 & k_0 \end{bmatrix} \tag{5}$$

The root locus stability criterion can be used in the optimization framework by constraining the $\mathcal{H}_\infty$ norm to unity for the $z \to w$ robustness channel and minimizing the performance goal like in previous sections.

**Problem 3** *Given $M$ from (5) with gain $k_l < K < k_h$, find stable $H_1(\lambda)$ such that:*

$$||G_{zw}(\lambda)||_\infty < 1.$$

*2.2.4 $\mathcal{H}_2$ Stability Criterion*

A statistical look at the $\Sigma\Delta$M loop with a single-bit quantizer shows that if the probability density function (PDF) of $u$ at the quantizer input is known, the quantizer gain $K$ is no longer undefined, the quantization noise $d$ is uncorrelated to the input $r$, and the stability of the modulator can be evaluated

[25]. With this analysis, the stability is dependent on the PDF and the power gain from $d$ to $y$, which is equal to $||S(\lambda)||_2^2$. In reality, the PDF depends on the input signal $r$, but may be assumed to be a standard type such as uniform, triangular, or Gaussian. The Gaussian PDF has been shown to be a close approximation for high-order modulators but is more conservative than the others [25].

To employ the stability criterion, the $\mathcal{H}_2$ LMI is used to constrain the 2-norm of the $r \rightarrow e$ channel to a value dependent on the desired maximum stable input amplitude $||u||_\infty$ and the choice of PDF, as in Problem 4. The $\mathcal{H}_2$ criterion is only applicable to discrete-time designs because continuous-time sensitivity functions have infinite $\mathcal{H}_2$ norm.

**Problem 4** *For given positive $\gamma_2$, solve the following for stable $H_1(z)$:*

$$||G_{er}(z)||_2 < \gamma_2.$$

### *2.2.5 Improved $\ell_1$ Stability Criterion*

The bounded-input bounded-output $\ell_1$ stability criterion is a sufficient criterion for stability. The $\ell_1$ norm of the loop filter, $||H(\lambda)||_1$, denotes its maximal peak-to-peak gain with a worst-case quantization noise signal. Assuming the quantized feedback is bounded $y \in [-1, 1]$, a limit to $||H(\lambda)||_1$ can be derived that ensures stability for a class of $\ell_\infty$-bounded input signals [2]. This bound is extremely conservative, but may be improved to some degree by taking advantage of the fact that the single-bit quantizer is invariant to any choice of positive gain $K$ that precedes the quantization operation. The improved $\ell_1$ criterion states that the modulator is guaranteed stable for inputs $r$ where $\min_K ||S(\lambda)||_1 \leq 3 - ||r||_\infty$ [25]. A time domain interpretation of this stability criterion is a bound on the sum of impulse response coefficients of the loop filter. Problem 5 shows the $\ell_1$ norm optimization objective.

**Problem 5** *For given positive $\gamma_1$, solve the following for stable $H_1(\lambda)$:*

$$||G_{er}(\lambda)||_1 < \gamma_1.$$

## 3 Optimization Framework

The optimization framework unifies the expression for the GKYP, $\mathcal{H}_2$, and $\ell_1$ norm LMIs for the augmented system (3) and allows it to be solved despite the infinite impulse response filter design problem being non-convex.

### 3.1 GKYP Lemma

The Generalized Kalman-Yakubovič-Popov lemma is a semidefinite expression that allows $\mathcal{H}_\infty$ minimization in only a specific finite frequency interval, such as that used in solving Problem 1.

**Lemma 3.1 (GKYP lemma [8])** *Given state-space matrices $\mathcal{A} \in \mathbb{R}^{n\times n}$, $\mathcal{B}_p \in \mathbb{R}^{n\times 1}$, $\mathcal{C}_q \in \mathbb{R}^{1\times n}$, $\mathcal{D}_{qp} \in \mathbb{R}^{1\times 1}$ 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:*

$$||G_{qp}(\lambda)||_\infty^2 < \gamma_\infty^2 \quad \omega_l \leq \lambda \leq \omega_h$$

*holds if and only if $Q \geq 0$ and quadratic matrix inequality (QMI):*

$$-\begin{bmatrix}\mathcal{A} & \mathcal{B}_p \\ I & 0\end{bmatrix}^T (\Phi \oplus P + \Psi \oplus Q) \begin{bmatrix}\mathcal{A} & \mathcal{B}_p \\ I & 0\end{bmatrix} + \\ -\begin{bmatrix}\mathcal{C}_q & \mathcal{D}_{pq} \\ 0 & I\end{bmatrix}^T \begin{bmatrix}1 & 0 \\ 0 & -\gamma_\infty^2\end{bmatrix} \begin{bmatrix}\mathcal{C}_q & \mathcal{D}_{pq} \\ 0 & I\end{bmatrix} \geq 0 \quad (6)$$

*is satisfied, where $\oplus$ denotes the Kronecker product. For the continuous-time case:*

$$\Phi = \begin{bmatrix}0 & 1 \\ 1 & 0\end{bmatrix} \qquad \Psi = \begin{bmatrix}-1 & j\omega_c \\ -j\omega_c & -\omega_1\omega_h\end{bmatrix} \quad (7)$$

*while for the discrete-time case:*

$$\Phi = \begin{bmatrix}1 & 0 \\ 0 & -1\end{bmatrix} \qquad \Psi = \begin{bmatrix}0 & e^{j\omega_c} \\ e^{-j\omega_c} & -2\cos w_0\end{bmatrix} \quad (8)$$

*where[1]:*

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

For $\mathcal{H}_\infty$ minimization across all frequencies, i.e. in Problem 3, Lemma 3.1 is modified by fixing $Q$ and adding an additional non-negative definiteness constraint:

$$Q = \mathbf{0} \qquad P \geq 0. \quad (9)$$

Although the strict inequality form of (6) is typically used in control problems (e.g. [33]), the non-strict positive-real LMI is presented here because it allows poles to occupy the unit circle (discrete-time) or imaginary axis (continuous-time) [9], as is often seen in heuristic-based $\Sigma\Delta M$ designs. To use the nonstrict inequality, the system must also be controllable, a guarantee of which follows in Section 3.4.

[1] The high-pass case of Lemma 3.1 where $\omega_h = \infty$ (continuous-time) or $\omega_h = \pi$ (discrete-time) is a special case and not shown here, because only low-pass or band-pass modulators are commonly used. For more information on high-pass GKYP design, see [8].

### 3.2 $\mathcal{H}_2$ Semidefinite Expression

The $\mathcal{H}_2$ norm can be minimized between 2 channels by solving a pair of inequalities with some similarities to Lemma 3.1.

**Theorem 3.2** *Given state-space matrices $\mathcal{A} \in \mathbb{R}^{n\times n}$, $\mathcal{B}_p \in \mathbb{R}^{n\times 1}$, $\mathcal{C}_q \in \mathbb{R}^{1\times n}$, $\mathcal{D}_{qp} \in \mathbb{R}^{1\times 1}$ of system $G_{qp}(\lambda)$, symmetric matrix variable $P \in \mathbb{R}^{n\times n}$ and $\Phi$ from (7) or (8), the $\mathcal{H}_2$ condition:*

$$||G_{qp}(\lambda)||_2^2 < \gamma_2^2$$

*holds if and only if the following QMI and LMI are satisfied:*

$$-\begin{bmatrix}\mathcal{A} & \mathcal{B}_p\\ I & 0\end{bmatrix}^T (\Phi \oplus P)\begin{bmatrix}\mathcal{A} & \mathcal{B}_p\\ I & 0\end{bmatrix} + \begin{bmatrix}\mathbf{0} & 0\\ 0 & 1\end{bmatrix} > 0 \tag{10}$$

$$\begin{bmatrix}\gamma_2^2 & \mathcal{C}_q & \mathcal{D}_{qp}\\ \mathcal{C}_q^T & P & 0\\ \mathcal{D}_{qp}^T & 0 & 1\end{bmatrix} > 0. \tag{11}$$

*Proof* Simplifying (10) by multiplying outer factors and summing yields:

$$-\begin{bmatrix}P\mathcal{A} + \mathcal{A}^T P & P\mathcal{B}_p\\ \mathcal{B}_p^T P & -1\end{bmatrix} > 0 \tag{12}$$

for continuous-time designs. Assuming $\mathcal{D}_{qp} = 0$ as is necessary for the continuous-time case, (11 )simplifies to:

$$\begin{bmatrix}\gamma_2^2 & \mathcal{C}_q\\ \mathcal{C}_q^T & P\end{bmatrix} > 0. \tag{13}$$

Equations 12 and 13 comprise the well-known $\mathcal{H}_2$ QMI for continuous-time systems [27,18]. For the discrete-time case, the simplification of (10) along the same lines results in:

$$\begin{bmatrix}-\mathcal{A}^T P\mathcal{A} + P & -\mathcal{A}^T P\mathcal{B}_p\\ -\mathcal{B}_p^T P\mathcal{A} & -\mathcal{B}_p^T P\mathcal{B}_p + 1\end{bmatrix} > 0 \tag{14}$$

which can be manipulated into the form:

$$\begin{bmatrix}P & 0\\ 0 & 1\end{bmatrix} - \begin{bmatrix}\mathcal{A}^T\\ \mathcal{B}_p^T\end{bmatrix} P \begin{bmatrix}\mathcal{A} & \mathcal{B}_p\end{bmatrix} > 0. \tag{15}$$

By Schur complement around $P^{-1}$, this becomes:

$$\begin{bmatrix}P^{-1} & \mathcal{A} & \mathcal{B}_p\\ \mathcal{A}^T & P & 0\\ \mathcal{B}_p^T & 0 & 1\end{bmatrix} > 0. \tag{16}$$

Finally, after a congruent transformation of (16) by $\mathrm{diag}\,(P, I, 1)$, combining it with (11) matches the well-known $\mathcal{H}_2$ QMI for discrete-time systems [18].

### 3.3 $\ell_1$ Semidefinite Expression

The computation of the $\ell_1$ norm is made tractable by instead minimizing the $\star$-norm, an upper bound on the $\ell_1$ norm. The $\star$-norm minimization is a set of 1 LMI, 1 QMI, and 1 bilinear matrix inequality (BMI) with a scalar parameter that enters nonlinearly. The equations can be solved by running a one-dimensional constrained minimization problem.

**Theorem 3.3** *Given state-space matrices $\mathcal{A} \in \mathbb{R}^{n\times n}$, $\mathcal{B}_p \in \mathbb{R}^{n\times 1}$, $\mathcal{C}_q \in \mathbb{R}^{1\times n}$, $\mathcal{D}_{qp} \in \mathbb{R}^{1\times 1}$ of system $G_{qp}(\lambda)$, symmetric matrix variable $P \in \mathbb{R}^{n\times n}$, auxiliary scalar variables $\mu > 0$, $\nu > 0$, and $\alpha \in (0, 1)$, and $\Phi$ from (7) or (8), the $\star$-norm condition:*

$$||G_{qp}(\lambda)||_\star^2 < \gamma_\star^2$$

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

$$-\begin{bmatrix} \mathcal{A} & \mathcal{B}_p \\ I & 0 \end{bmatrix}^T \left( \left( \Phi + \begin{bmatrix} 0 & 0 \\ 0 & \alpha \end{bmatrix} \right) \oplus P \right) \begin{bmatrix} \mathcal{A} & \mathcal{B}_p \\ I & 0 \end{bmatrix} + \\ + \begin{bmatrix} \mathbf{0} & 0 \\ 0 & 1 \end{bmatrix} > 0 \tag{17}$$

$$\begin{bmatrix} \alpha P & 0 & \mathcal{C}_q \\ 0 & \mu - 1 & \mathcal{D}_{qp} \\ \mathcal{C}_q^T & \mathcal{D}_{qp}^T & \nu \end{bmatrix} > 0 \tag{18}$$

*are satisfied for some $\alpha$, and the following LMI is also satisfied:*

$$\begin{bmatrix} \gamma_\star^2 & \mu & \nu \\ \mu & 1 & 0 \\ \nu & 0 & 1 \end{bmatrix} > 0. \tag{19}$$

*Proof* The proof proceeds in a similar way to that of Theorem 3.2 by transforming (17) as was done with (10) in Proof 3.2. Then, combined with (18) and (19), the equations are in the form of the $\star$-norm semidefinite program reported in literature [3, 22].

### 3.4 Convexification

Equations 6, 10, and 17 are non-convex because there are products between the state-space matrices and the optimization variable $P$. The number of product terms can be reduced by assuming the state space system (2) is in controllable canonical form (CCF) as shown below:

$$\lambda \mathbf{x} = \overbrace{\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix}}^{A_H} \mathbf{x} + \overbrace{\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}}^{B_H} e \tag{20}$$

$$u = \underbrace{\begin{bmatrix} b_0 & b_1 & b_2 & \cdots & b_{n-1} \end{bmatrix}}_{C_H} \mathbf{x} \tag{21}$$

where $a_H \equiv \begin{bmatrix} a_0 \; a_1 \; a_2 \cdots a_{n-1} \end{bmatrix}^T$ and $b_H \equiv \begin{bmatrix} b_0 \; b_1 \; b_2 \cdots b_{n-1} \end{bmatrix}^T$ for brevity. Note that when the system (2) is in CCF, sub-systems $G_{er}(\lambda)$ and $G_{zw}(\lambda)$ (3) 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 $T\mathcal{B}_w$ is unity. Having the system in CCF also permits the use of the non-strict GKYP lemma because the system is controllable by definition. 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 Problems 1–5.

### *3.4.1 Change of Variables*

Let $a \equiv a_H + m_{22} b_H$, the negative transpose of the lower row of $\mathcal{A}$, and define $b \equiv a_H$ for system $G_{er}(\lambda)$ and $b$ be that shown in (22) depending on the subsystem $G_{qp}(\lambda)$. The semidefinite program is redefined in terms of these variables to simplify nomenclature.

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

### *3.4.2 Sensitivity Shaping*

Addressing Problem 1, the authors of [15, Th. 1] have shown that a congruent transformation of Equation 6 by the matrix:

$$\begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix} \tag{23}$$

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 (6). 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 [15] is only applicable to shaping the sensitivity function $G_{er}(\lambda)$ because it assumes $\mathcal{D} = 1$. A full derivation that is valid for any $\mathcal{D}$, such as that encountered when solving Problem 3, is given in Appendix A.1.

### 3.4.3 $\mathcal{H}_2$, $\ell_1$ Optimization

The congruent transformation procedure from Section 3.4.2 does not depend on the centre expression (that may be a function of any of $\Phi$, $\Psi$, $P$, $Q$) so it is applicable to both $\mathcal{H}_2$ (10) and $\ell_1$ (17) cases, which have the same outer factors. This procedure restores linearity to the first summation term, while the second term of both is the same. Equation 24 shows this second term with congruent transformation (23) applied:

$$\begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{0} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix}^T = \begin{bmatrix} aa^T & a \\ a^T & 1 \end{bmatrix} \tag{24}$$

It is seen that, like (6), 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.

### 3.4.4 Iterative Procedure

The solving of a quadratically constrained LMI is a difficult problem. Several methods of solving (28) and (35) were attempted but had poor results. These included directly using a general non-convex solver, using a rank-constrained LMI solver [23], and using Shor's relaxation to linearize the problem [14]. Instead, we use the iterative method from [30] which, for problems with simple non-convexities, appears to be similar to the method used in [15] but with an extra parameter to guarantee finite convergence. The iterative LMI problems generated with this method were solved numerically using the YALMIP Toolbox for MATLAB [17] with the LMILAB solver [7].

## 4 Design Examples

The following design examples are intended to show how the optimization framework can be used with various stability criteria to design $\Sigma\Delta$Ms. In Examples 4.1 to 4.4, we designed discrete-time fourth- or fifth-order loop filters with an oversampling ratio of 32 intended for audio applications, i.e. low pass signals with Nyquist frequency 44.1 kHz. This architecture of modulator has been used for digital audio [26, 1]. In Example 4.5, we designed a continuous-time modulator for the same application.

### 4.1 $\mathcal{H}_\infty$ Design

The $\mathcal{H}_\infty$ design procedure is done by solving Problem 1 while enforcing stability with Problem 2. We used Lemma 3.1 for the former and also for the latter, along with conditions in Equation 9, which implicitly forces the NTF to be stable. In this example, the Lee criterion of $\gamma_\infty = 1.5$ was chosen. The optimization problem converged to the loop filter transfer function:

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

The sensitivity function of this filter can be seen in Figure 3. Note that the Lee criterion for stability is satisfied across all frequencies and the peak gain in the signal band has been minimized to $-64\,\mathrm{dB}$ by the GKYP lemma. This compared favourably (in the $\mathcal{H}_\infty$ sense) to the design produced with the toolbox[2] in [29], which has peak gain in the signal band of $-55\,\mathrm{dB}$.

Like most high-order designs using the Lee criterion, stability is conditional on input amplitude. A simulation of this can be seen in Figure 4, also performed with the Delta Sigma Toolbox. A peak signal-to-quantization-noise ratio (SQNR) of 86 dB at an input amplitude of 0.62 was achieved with a maximum stable input amplitude (MSIA) of 0.71 and a minimum resolvable input amplitude of $-91\,\mathrm{dB}$ full scale (FS). The comparable toolbox design achieved a very similar peak SQNR but with a slightly better MSIA and minimum resolvable input amplitude of 0.76 and $-96\,\mathrm{dB}$ FS, respectively.

### 4.2 Robust Root Locus Design

The root locus design technique is equivalent to solving Problems 1 and 3 simultaneously. Similar to the previous example, we applied Lemma 3.1 to the sensitivity channel $r \to e$ for performance and Lemma 3.1 along with conditions in Equation 9 to the robustness channel $w \to z$ for stability. 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, we investigated how using the quantizer gain robustness criterion can be used to enhance the stable input range of Design 4.1. Instability in $\Sigma\Delta$Ms is often associated with low quantizer gains. To improve stability, we modulate the lower bound of the quantizer gain, $k_l$, and solve the optimization problem. Thus, $k_l$ is a parameter that trades off performance and stability. We designed a fourth order modulator in order to compare it with the root locus design method described in [16]. With some trial-and-error, $k_l = 0.08$ was the lowest lower bound on the quantizer gain that maintained empirical full-scale stability. The solver converged to the loop transfer function:

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

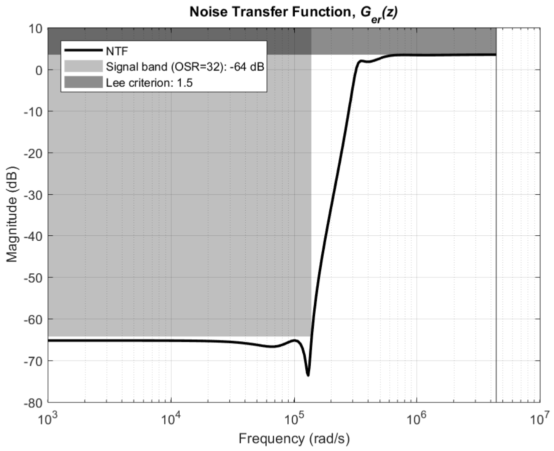

**Fig. 3** The sensitivity function of the design in Example 4.1. The dark shaded area represents the stability constraint and the light shaded area represents the achieved noise attenuation performance.

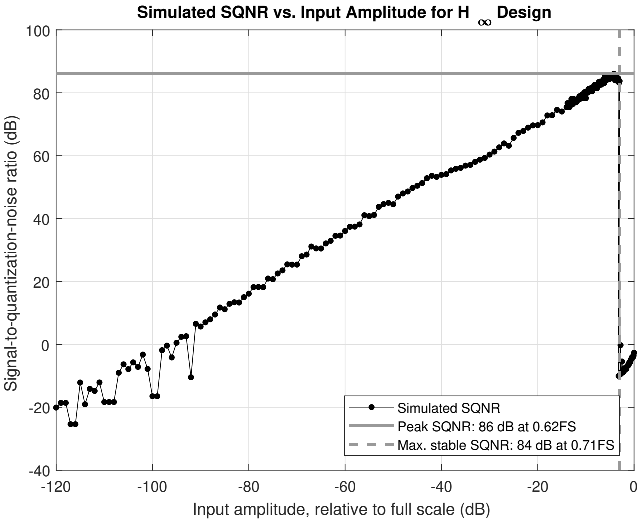

**Fig. 4** An SQNR plot of 247 simulations of the design in Example 4.1 to an input sinusoid of frequency $3.46 \times 10^4\,\mathrm{rad\,s^{-1}}$ and varying amplitude to investigate its conditional stability.

$$H_1(z) = \frac{2.16\,(z - 0.852)\left(z^2 - 1.93z + 0.947\right)}{(z^2 - 1.96z + 0.958)\,(z^2 - 1.98z + 0.993)}$$

The robustness and sensitivity channels are shown in Figure 5 and the root locus is shown in Figure 6. A simulation like done previously confirmed that system is stable for input amplitudes up to full scale. As expected when order

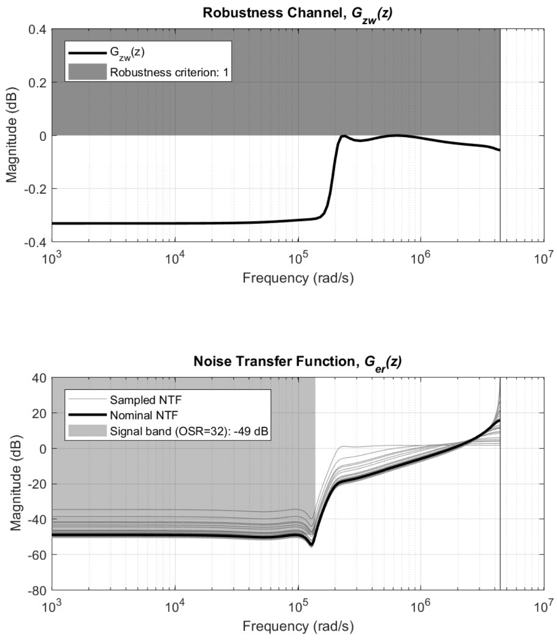

**Fig. 5** Upper: a frequency response plot of the robustness channel for the design in Example 4.2. The $\mathcal{H}_\infty$ norm of the transfer function is below 1 (shaded region) for all frequencies, showing that the system is stable for all norm-bounded quantizer gains in the range $[0.085, \infty)$. 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.

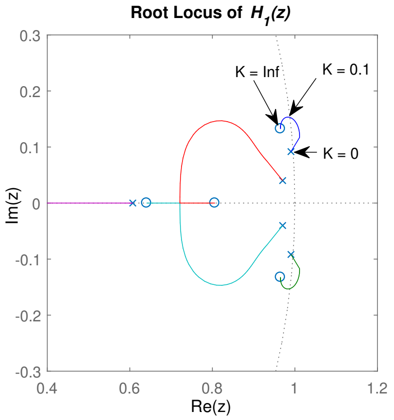

**Fig. 6** A subset of the complex plane showing the root locus of the filter from Example 4.2 across quantizer gains.

is decreased and stability is increased, the empirical peak SQNR is reduced, to 63 dB with the minimum resolvable input amplitude at −53 dB FS. In comparison, a similar fourth-order, 32 OSR sigma delta modulator designed using the procedure described in [16] with the root locus stability criterion achieves a peak SQNR of 52 dB at the margin of full-scale input stability.

### 4.3 $\mathcal{H}_2$ Design

The $\mathcal{H}_2$ design technique is done by solving Problems 1 and 4. The former is for performance and uses Lemma 3.1 while the latter favours stability and uses Theorem 3.2. In this example, we designed a modulator for the same specifications as Example 4.1. One advantage of the $\mathcal{H}_2$ criterion is that there is a more systematic way to target a specific MSIA using the Gaussian PDF [25]. For a target MSIA of 0.46, the criterion was satisfied when $||G_{er}(z)||_2^2 < 2.246$. Using this constraint along with the performance optimization, the solver converged to the loop filter transfer function:

$$H_1(z) = \frac{(z-0.932)\left(z^2-1.96z+0.959\right)\left(z^2-1.99z+0.995\right)}{(z-0.374)\left(z^2-1.92z+0.967\right)\left(z^2-1.66z+0.797\right)}.$$

Computing the NTF gain for the optimum value of $K = 0.881$, we obtain $||G_{er}(z)||_2^2 = 2.16$, indicating that the 2-norm constraint was satisfied. Because this stability criteria is only an approximation (in this case, pessimistic), the empirical MSIA was considerably higher at 0.63. The peak SQNR was found to be 89 dB at an input amplitude of 0.56 FS. The minimum resolvable input amplitude was −98 dB FS. These measurements can be seen in Figure 7. These results compare closely to the NTF design method in [12, Fig. 5], which achieved around 90 dB peak SQNR and a very similar stability threshold for a design with equivalent specifications.

### 4.4 Guaranteed Stable $\ell_1$ Design

In this example, we produced a modulator design that is mathematically guaranteed to be stable for a range of input amplitudes by solving Problems 1 and 5. As discussed in Section 3.3, the optimization target was an upper bound on the $\ell_1$ norm, so there was some trial-and-error to find the $\star$-norm upper bound that produces the desired $\ell_1$ norm bound. In this case, running the optimization problem with a $\star$-norm constraint of 4 resulted in $||S(z)||_1 = 2.36$ and a loop filter as follows:

$$H_1(z) = \frac{0.444\left(z^2-1.78z+0.795\right)}{(z-0.944)\left(z^2-1.945z+0.955\right)}.$$

This design is guaranteed to be stable for input amplitudes up to 0.64 FS, but empirical stability is seen for the entire input range. The cost for

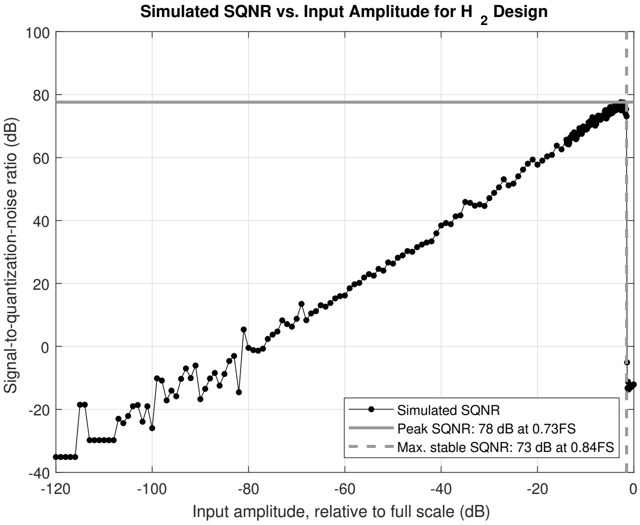

**Fig. 7** An SQNR plot of 121 simulations of the design in Example 4.3 to an input sinusoid of frequency $3.46 \times 10^4\,\mathrm{rad\,s^{-1}}$ and varying amplitude to investigate its conditional stability.

this stability was a decrease in performance: a peak SQNR of 62 dB and a minimum resolvable input amplitude of −40 dB FS. Note that similar to the case in Example 4.3, there are now 2 pole-zero cancellations in the final loop filter resulting in an order 3 transfer function. The NTF magnitude for this design can been seen in Figure 8 as well as its impulse response, the time-domain equivalent to the $\ell_1$ norm.

## 4.5 A Continuous-Time Design

As a proof of concept, we designed a 3rd order continuous-time loop filter to the same specifications as the discrete-time examples above. In the absence of stability criteria that may be directly applied to continuous-time designs, we specified the following (somewhat arbitrary) constraints:

$$\min_{\omega\in[0,2\pi\cdot 4.41\times 10^5]} ||G_{er}(j\omega)||_\infty \quad \text{s.t.} \tag{25}$$

$$||G_{er}(j\omega)||_\infty \leq 4\,\mathrm{dB} \quad \forall\omega \tag{26}$$

$$||G_{yr}(j\omega)||_\infty \leq -10\,\mathrm{dB} \quad \omega \in [2\pi\cdot 7.056\times 10^5, \infty) \tag{27}$$

The constraint in (25) uses Problem 1 to minimize the in-band noise of the sensitivity function. The constraint in Equation 26 favours stability using Problem 1. Unlike discrete-time designs, there is no influence of the quantizer sampling frequency captured in the first two constraints. Thus, we introduced

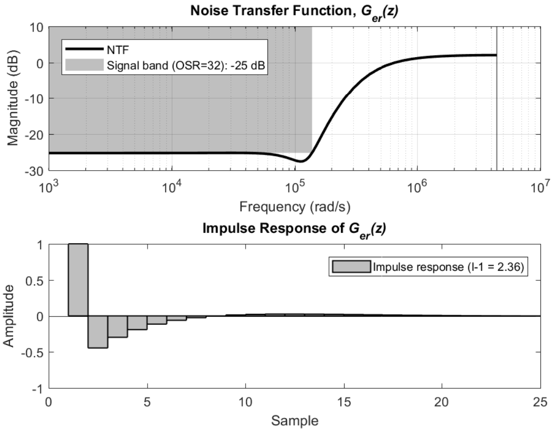

**Fig. 8** Upper: The sensitivity function of the $\ell_1$ norm design from Section 4.4 shows the signal band rejection has been diminished in order to provide mathematically guaranteed stability. Lower: the impulse response of the NTF has been minimized by the $\ell_1$ norm criterion.

the constraint in Equation 27 to force high roll-off by reducing the complementary sensitivity function (STF) outside the signal band.

The optimization process converged on the loop filter transfer function:

$$\frac{1.11 \times 10^8 \left(s^2 + 3.25 \times 10^6 + 1.178 \times 10^{13}\right)}{\left(s + 2.37 \times 10^8\right)\left(s^2 + 3.80 \times 10^4 + 3.94 \times 10^{10}\right)},$$

for which the sensitivity and complementary sensitivity functions are shown in Figure 9 and the spectrum of the simulated quantizer output is shown in Figure 10.

## 5 Conclusion

In this paper, we introduce an optimization framework using the GKYP lemma for IIR loop filter design that is compatible with any stability constraints that can be defined in terms of $\mathcal{H}_\infty$, $\mathcal{H}_2$, and $\ell_1$ norms. We allow the robustness against the linearized quantizer gain to be controlled by forming a closed-loop system and extracting the uncertainty via LFT. Building on the work of [15], we expand the GKYP lemma expression to support arbitrary state-space systems. Finally, the procedure can be used to design continuous-time $\Sigma\Delta$Ms. The design examples showcase how this procedure may be used with a wide range of stability criteria and is competitive with existing methods. The $\mathcal{H}_2$ design criteria produced a fourth-order loop filter with 89 dB SQNR and stability under inputs up to 0.63 FS in simulation. This method of design is a good balance because it retains performance while allowing stability to

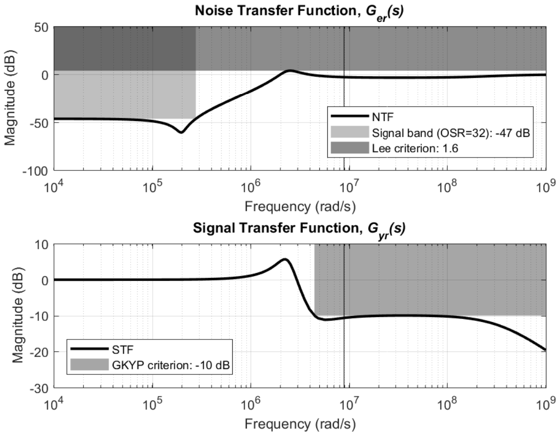

**Fig. 9** Upper: the sensitivity function of the continuous-time design from Section 4.5. Lower: the complementary sensitivity function with a GKYP constraint to enforce sharp roll-off.

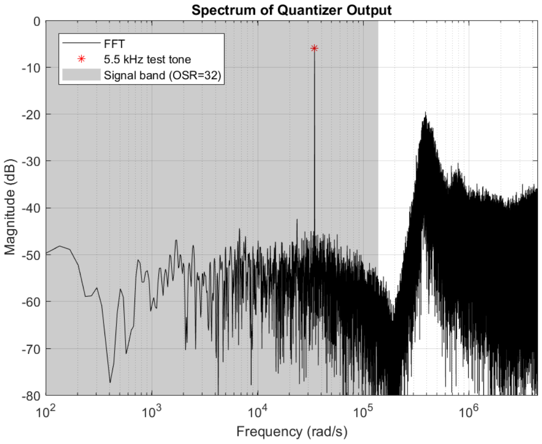

**Fig. 10** The 14 000-point FFT of the quantizer output of the continuous-time design from Section 4.5 shows about 40 dB of noise shaping to a $3.46 \times 10^4\,\mathrm{rad\,s^{-1}}$ input sinusoid with amplitude 0.5 FS.

be maintained in an intuitive way. While not a guarantee of stability like the $\ell_1$ criterion, it relies on reasonable assumptions and is suited for high order, single-bit designs. The root locus method shares many of these properties, however, the $\mathcal{H}_2$ stability criterion had a more broad and well-behaved range of design targets that yielded respectable loop filters. Areas of future interest include a closer look into the termination criteria of the iterative procedure used to solve the semidefinite program, formalizing how each stability criteria affects the resulting filter, using alternative performance criteria, and a better way of defining constraints for continuous-time designs.

**Acknowledgements** This work was supported by ESS Technology, Inc. through the Mitacs Accelerate program.

## References

1. Adams, R.W., Ferguson, P.F.J., Vincellette, S., Ganesan, A., Volpe, T., Libert, B.: Theory and Practical Implementation of a 5th-Order Sigma-Delta A/D Converter. In: Audio Engineering Society Convention 90 (1991)
2. Anastassiou, D.: Error Diffusion Coding for A/D Conversion. IEEE Transactions on Circuits and Systems **36**(9), 1175–1186 (1989). DOI 10.1109/31.34663
3. Bu, J., Sznaier, M.: Linear matrix inequality approach to synthesizing low-order suboptimal mixed l1/Hp controllers. Automatica **36**(7), 957–963 (2000). DOI 10.1016/S0005-1098(00)00005-4
4. Callegari, S., Bizzarri, F., Brambilla, A.: Optimal Coefficient Quantization in Optimal-NTF $\Delta\Sigma$ Modulators. IEEE Transactions on Circuits and Systems II: Express Briefs **65**(5), 542–546 (2018). DOI 10.1109/TCSII.2018.2821368
5. Chao, K.C.H., Nadeem, S., Lee, W.L., Sodini, C.G.: A Higher Order Topology for Interpolative Modulators for Oversampling A/D Converters. IEEE Transactions on Circuits and Systems **37**(3), 309–318 (1990). DOI 10.1109/31.52724
6. Derpich, M.S., Silva, E.I., Quevedo, D.E., Goodwin, G.C.: On optimal perfect reconstruction feedback quantizers. IEEE Transactions on Signal Processing **56**(8 II), 3871–3890 (2008). DOI 10.1109/TSP.2008.925577. URL http://ieeexplore.ieee.org/document/4522532/
7. Gahinet, P., Nemirovski, A., Laub, A.J., Chilali, M.: LMI Control Toolbox For Use with MATLAB, 1 edn. The Mathworks, Inc. (1995). DOI 10.1109/CDC.1994.411440. URL http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:LMI+Control+Toolbox+For+Use+with+MATLAB#1
8. Iwasaki, T., Hara, S.: Generalized KYP Lemma: Unified Frequency Domain Inequalities with Design Applications. IEEE Trans. Autom. Control **50**(1), 41–59 (2005)
9. Iwasaki, T., Hara, S., Yamauchi, H.: Dynamical system design from a control perspective: Finite frequency positive-realness approach. IEEE Transactions on Automatic Control **48**(8), 1337–1354 (2003). DOI 10.1109/TAC.2003.815013
10. Kang, K.: Simulation, and Overload and Stability Analysis of Continuous Time Sigma Delta Modulator. Ph.D. thesis, University of Nevada (2014)
11. Kenney, J., Carley, L.: CLANS: a high-level synthesis tool for high resolution data converters. In: IEEE International Conference on Computer-Aided Design (ICCAD-89) Digest of Technical Papers, pp. 496–499. Pittsburgh (1988). DOI 10.1109/ICCAD.1988.122557. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=122557
12. Kidambi, S.: Design of Noise Transfer Functions for Delta-Sigma Modulators Using the Least-pth Norm. IEEE Transactions on Circuits and Systems II: Express Briefs **66**(4), 707–711 (2019). DOI 10.1109/TCSII.2018.2880925
13. Kuo, T.H., Yang, C.C., Chen, K.D., Wang, W.C.: Design Method for High-Order Sigma Delta Modulator Stabilized by Departure Angles Designed to Keep Root Loci in Unit Circle. IEEE Transactions on Circuits and Systems II: Express Briefs **53**(10), 1083–1087 (2006). DOI 10.1109/TCSII.2006.882219
14. Lasserre, J.: Convergent LMI relaxations for nonconvex quadratic programs. Proceedings of the 39th IEEE Conference on Decision and Control (Cat. No.00CH37187) **5**(2), 5041–5046 (2000). DOI 10.1109/CDC.2001.914738. URL http://ieeexplore.ieee.org/document/914738/
15. Li, X., Yu, C., Gao, H.: Design of delta-sigma modulators via generalized Kalman-Yakubovich-Popov lemma. Automatica **50**(10), 2700–2708 (2014). DOI 10.1016/j.automatica.2014.09.002

16. Liu, J.M., Chien, S.H., Kuo, T.H.: Optimal design for delta-sigma modulators with root loci inside unit circle. IEEE Transactions on Circuits and Systems II: Express Briefs **59**(2), 83–87 (2012). DOI 10.1109/TCSII.2011.2180096
17. Löfberg, J.: YALMIP: A Toolbox for Modeling and Optimization in MATLAB (2004)
18. Masubuchi, I., Ohara, A., Suda, N.: LMI-based controller synthesis: a unified formulation and solution. Robust and Nonlinear Control **8**(9), 669–686 (1998). DOI 10.1002/(SICI)1099-1239(19980715)8:8¡669::AID-RNC337¿3.0.CO;2-W
19. de Melo, J.L.A., Pereira, N., Leitao, P.V., Paulino, N., Goes, J.: A Systematic Design Methodology for Optimization of Sigma-Delta Modulators Based on an Evolutionary Algorithm. IEEE Transactions on Circuits and Systems I: Regular Papers **66**(9), 3544–3556 (2019). DOI 10.1109/tcsi.2019.2925292
20. Nagahara, M., Yamamoto, Y.: Frequency Domain Min-Max Optimization of Noise-Shaping Delta-Sigma Modulators. IEEE Transactions on Signal Processing **60**(6), 1–12 (2012). DOI 10.1109/TSP.2012.2188522
21. Oberoi, A.: A Convex Optimization Approach to the Design of Multiobjective Discrete Time Systems. Master of science, Rochester Institute of Technology (2004)
22. Oberoi, A., Cockburn, J.C.: A simplified LMI approach to l1 Controller Design. In: Proceedings of the 2005 American Control Conference, pp. 1788–1792. Portland (2005)
23. Orsi, R., Helmke, U., Moore, J.B.: A Newton – Like Method for Solving Rank Constrained Linear Matrix Inequalities. Automatica **42**(11), 1875–1882 (2006). DOI 10.1016/j.automatica.2006.05.026. URL http://linkinghub.elsevier.com/retrieve/pii/S0005109806002391
24. Osqui, M.M., Megretski, A.: Semidefinite Programming in Analysis and Optimization of Performance of Sigma-Delta Modulators for Low Frequencies. In: Proceedings of the 2007 American Control Conference, 6, pp. 3582–3587 (2007)
25. Risbo, L.: Sigma Delta Modulators - Stability Analysis and Optimization. Doctor of philosophy, Technical University of Denmark (1994). URL http://orbit.dtu.dk/files/5274022/Binder1.pdf
26. Ritoniemi, T., Karema, T., Tenhunen, H.: A fifth order sigma-delta modulator for audio A/D-converter. In: 1991 International Conference on Analogue to Digital and Digital to Analogue Conversion, pp. 153–158. IET, Swansea (1991). DOI 10.1109/VLSIC.1991.760064. URL https://ieeexplore.ieee.org/abstract/document/151991
27. Scherer, C.W., Gahinet, P., Chilali, M.: Multiobjective output-feedback control via LMI optimization. IEEE Transactions on Automatic Control **42**(7), 896–911 (1997). DOI 10.1109/9.599969
28. Schreier, R.: An Empirical Study of High-Order Single-Bit Delta-Sigma Modulators. IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing **40**(8), 461–466 (1993). DOI 10.1109/82.242348
29. Schreier, R., Temes, G.C.: Understanding Delta-Sigma Data Converters, vol. 53. Wiley (1997). DOI 10.1109/9780470546772. URL https://cds.cern.ch/record/733538
30. Shishkin, S.L.: Optimization under non-convex Quadratic Matrix Inequality constraints with application to design of optimal sparse controller. IFAC-PapersOnLine **50**(1), 10754–10759 (2017). DOI 10.1016/j.ifacol.2017.08.2276. URL https://doi.org/10.1016/j.ifacol.2017.08.2276
31. Tariq, M.R., Ohno, S.: Unified LMI-based design of $\Delta\Sigma$ modulators. EURASIP Journal on Advances in Signal Processing **2016**(1), 29 (2016). DOI 10.1186/s13634-016-0326-2. URL https://asp-eurasipjournals.springeropen.com/articles/10.1186/s13634-016-0326-2
32. Yang, C.c., Chen, K.d., Wang, W.C., Kuo, T.h.: Transfer function design of stable high-order sigma-delta modulators with root locus inside unit circle. In: Proceedings. IEEE Asia-Pacific Conference on ASIC, pp. 5–8 (2002). DOI 10.1109/APASIC.2002.1031518. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1031518
33. Zhang, D., Shi, P., Yu, L.: Containment Control of Linear Multiagent Systems with Aperiodic Sampling and Measurement Size Reduction. IEEE Transactions on Neural Networks and Learning Systems **29**(10), 5020–5029 (2018). DOI 10.1109/TNNLS.2017.2784365

34. Zhou, K., Doyle, J.C., Glover, K.: Robust and Optimal Control, vol. 40. Prentice Hall, Englewood Cliffs (1996). DOI 10.1016/0967-0661(96)83721-X. URL http://www.ulb.tu-darmstadt.de/tocs/109091736.pdf

# A Derivation of Matrix Inequalities with One Non-Convex Term

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

**Theorem A.1** *Equation 6 from Section 3.1 is equivalent to the following:*

$$\begin{bmatrix} -\varXi_{11} + aa^T & -\varXi_{12} + a & -\mathcal{C}_q^T - a\mathcal{D}_{qp}^T \\ -\varXi_{12}^T + a^T & -\varXi_{22} + 1 & -\mathcal{D}_{qp}^T \\ -\mathcal{C}_q - a^T\mathcal{D}_{qp} & -\mathcal{D}_{qp} & \gamma_\infty \end{bmatrix} \geq 0 \tag{28}$$

*where (28) contains just one nonlinear term in variable a, and:*

$$\begin{bmatrix} \varXi_{11} & \varXi_{12} \\ \varXi_{12}^T & \varXi_{22} \end{bmatrix} = \begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathcal{A} & \mathcal{B}_p \\ I & 0 \end{bmatrix}^T (\varPhi \oplus P_\gamma + \varPsi \oplus Q_\gamma) \begin{bmatrix} \mathcal{A} & \mathcal{B}_p \\ I & 0 \end{bmatrix} \begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix}^T$$

$$P_\gamma = \gamma_\infty^{-1} P \qquad\qquad Q_\gamma = \gamma_\infty^{-1} Q. \tag{29}$$

*Proof* Starting from (6), we follow the procedure mentioned in Section 3.4.2 to eliminate non-convex products in the first term of the QMI [15, Th. 1]:

$$-\begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathcal{A} & \mathcal{B}_p \\ I & 0 \end{bmatrix}^T f(\varPhi, \varPsi, P, Q) \begin{bmatrix} \mathcal{A} & \mathcal{B}_p \\ I & 0 \end{bmatrix} \begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix}^T + \ldots \geq 0 \tag{30}$$

and introduce the notation $\varXi_{ij}$ for this linear part:

$$-\begin{bmatrix} \varXi_{11} & \varXi_{12} \\ \varXi_{12}^T & \varXi_{22} \end{bmatrix} + \ldots \geq 0. \tag{31}$$

Equation 31 may undergo a congruent transformation by $\gamma_\infty^{-\frac{1}{2}} I$ introducing a commutable factor of $\gamma_\infty^{-1}$ to every element. For the first summation term, we absorb the factor with redefinition (36) yielding:

$$-\begin{bmatrix} \varXi_{11} & \varXi_{12} \\ \varXi_{12}^T & \varXi_{22} \end{bmatrix} - \begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathcal{C}_q & \mathcal{D}_{qp} \\ 0 & I \end{bmatrix}^T \begin{bmatrix} \gamma_\infty^{-1} & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \mathcal{C}_q & \mathcal{D}_{qp} \\ 0 & I \end{bmatrix} \begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix}^T \geq 0. \tag{32}$$

Multiplying the inner factors in the second term of (32) leads to:

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

which can be expanded into:

$$\begin{aligned} &-\begin{bmatrix} \varXi_{11} & \varXi_{12} \\ \varXi_{12}^T & \varXi_{22} \end{bmatrix} - \\ &\qquad \begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix} \begin{bmatrix} I & a\mathcal{D}_{qp}^T \\ 0 & \mathcal{D}_{qp}^T \end{bmatrix} \begin{bmatrix} \mathcal{C}_q^T \\ 1 \end{bmatrix} \gamma_\infty^{-1} \begin{bmatrix} \mathcal{C}_q^T \\ 1 \end{bmatrix}^T \begin{bmatrix} I & a\mathcal{D}_{qp}^T \\ 0 & \mathcal{D}_{qp}^T \end{bmatrix}^T \begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix}^T \\ &\qquad\qquad + \begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} I & a \\ 0 & 1 \end{bmatrix}^T \geq 0. \end{aligned} \tag{33}$$

The 3 outer factors multiplied with $\gamma_\infty^{-1}$ in the middle term of (33) are then combined together and the last summation term is also multiplied through, resulting in the following:

$$-\begin{bmatrix}\Xi_{11} & \Xi_{12}\\ \Xi_{12}^T & \Xi_{22}\end{bmatrix}-\begin{bmatrix}\mathcal{C}_q^T+a\mathcal{D}_{qp}^T\\ \mathcal{D}_{qp}^T\end{bmatrix}\gamma_\infty^{-1}\begin{bmatrix}\mathcal{C}_q^T+a\mathcal{D}_{qp}^T\\ \mathcal{D}_{qp}^T\end{bmatrix}^T+\begin{bmatrix}aa^T & a\\ a^T & 1\end{bmatrix}\geq 0. \tag{34}$$

The last summation term of (34) is then added with the linear part $\Xi$. Because $\gamma_\infty > 0 \leftrightarrow \gamma_\infty^{-1} > 0$, a Schur complement taken around $\gamma_\infty$ allows (34) to be written as the single matrix inequality (28).

## A.2 Derivation of $\mathcal{H}_2$ and $\ell_1$ Inequalities

**Theorem A.2** *Equations 10 from Section 3.2 and (17) from Section 3.3 are equivalent to the following:*

$$\begin{bmatrix}-\Xi_{11}+aa^T & -\Xi_{12}+a\\ -\Xi_{12}^T+a^T & -\Xi_{22}+1\end{bmatrix}>0 \tag{35}$$

*where (35) contains just one nonlinear term in variable a, and:*

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

$$f\left(\Phi,P_\gamma,\alpha\right)=\begin{cases}\Phi\oplus P_\gamma & \textit{for the } \mathcal{H}_2 \textit{ case}\\ \left(\Phi+\begin{bmatrix}0 & 0\\ 0 & \alpha\end{bmatrix}\right)\oplus P_\gamma & \textit{for the } \ell_1 \textit{ case}\end{cases} \tag{36}$$

$$P_\gamma=\gamma_\infty^{-1}P. \tag{37}$$

*Proof* Starting from either (10) or (17), we follow the procedure mentioned in Section 3.4.3 to eliminate non-convex products in the first term of the QMI independent of $f\left(\Phi,P_\gamma,\alpha\right)$. The second summation term is the same in both QMIs and simplifies as shown in (24). Combining these results in the matrix inequality (35).