Min–Max Design of Error Feedback Quantizers Without Overloading

Shuichi Ohno, *Senior Member, IEEE*, Yuma Ishihara, and Masaaki Nagahara, *Senior Member, IEEE*

***Abstract*— In this paper, we design a no-overloading error feedback quantizer based on a ΔΣ modulator, composed of an error feedback filter and a static quantizer. To guarantee no-overloading in the quantizer, we impose an l_∞ norm constraint on the feedback signal in the quantizer. Then, for a prescribed l_∞ norm constraint on the error at the system output induced by the quantizer, we design the error feedback filter that requires the minimum number of bits that achieves the constraint. Next, for a fixed number of bits for the quantizer, we investigate the achievable minimum l_∞ norm of the error at the system output with the no-overloading quantizer. Numerical examples are provided to validate our analysis and synthesis.**

***Index Terms*— Quantization, overloading, delta-sigma modulator, linear matrix inequalities.**

I. Introduction

QUANTIZATION is fundamental in digital processing. If a sufficient number of bits can be assigned to the quantizer, its errors and overloading may be negligible. However, there are still some applications where a sufficient number of bits cannot be utilized. For example, to transmit signals over rate-limited digital communication channels, the continuous-valued (or even discrete-valued) signals have to be quantized into low-resolution signals. But, it is often the case that communication rates are limited due to physical constraints especially when wireless communication is used. When only a small number of bits can be assigned to represent the signals, quantization errors may cause serious degradation.

An error feedback quantizer is more efficient than the conventional uniform quantization. It consists of a static uniform quantizer and a feedback filter, where the quantization error of a static uniform quantizer is filtered by an error feedback filter and then it is fed back to the input to the static uniform quantizer (see Fig. 4 in Section II).

Error feedback quantizers have been used to reduce quantization error in the coefficients of digital filters [1]–[3]. On the other hand, $\Delta\Sigma$ modulators also employ the error feedback mechanism, and are often utilized in practice to convert real numbers into fixed-point numbers [4]. For control systems, a variant of $\Delta\Sigma$ modulator has been studied in [5], which is called a *dynamic quantizer*. The parameters in the dynamic quantizer can be obtained by linear programming (LP) [6] and by convex optimization [7]. To avoid overloading, [6] proposes to limit the l_∞ norm of the feedback signals. However, the dynamic quantizer only supports a smaller set of error feedback filters than conventional ΔΣ modulators and hence the optimal performance cannot be guaranteed [8].

Recently, H_∞ optimal design of error feedback filters has been proposed based on the generalized Kalman-Yakubovich-Popov lemma [9], [10]. Also, a post filter connected to the ΔΣ modulator is incorporated into the design of the error feedback filter [11] and the weighted noise spectrum is also exploited [12]. In [13], the optimal error feedback filter has been synthesized such that it minimizes the variance of the quantization error subject to the constraint on the variance of the input to the static uniform quantizer. However, the constraint on the variance does not necessarily guarantee no-overloading in the quantizer. In practical systems, an overloading may cause instability followed by a serious effect. To assure that no overloading occurs, we should take into account the maximum absolute value, i.e., the l_∞ norm of the input to the quantizer.

In some applications, particularly in control, the worst largest absolute value of the error is often important, since the system may be physically broken due the the error which is not within an allowable range. Thus, this paper develops a quantizer with error feedback that needs a small number of bits required for quantization to achieve the requirement on the worst-case error in the output connected to the quantizer, while keeping no-overloading in the quantizer. We regulate the l_∞ norm of the feedback signals in the quantizer to assure no-overloading.

First, we consider finite impulse response (FIR) feedback filters since the l_∞ norm of an FIR filter can be exactly and directly evaluated. We formulate the design of the optimal FIR feedback filter as LP, which can be readily solved numerically. The minimum number of bits assigned to the quantizer is determined with the optimized feedback filters. Then, we deal with infinite impulse response (IIR) feedback filters to reduce the l_∞ norm of the error at the output under the constraint on the l_∞ norm of the feedback signals. For the design of IIR filters, an upper bound of the l_∞ norm, which is not tight, is utilized, since the exact l_∞ norm of an IIR filter is not easily evaluated.

Next, for a given number of bits for the quantizer, that is, for a fixed data rate, we investigate the achievable minimum

Manuscript received May 1, 2017; revised August 21, 2017; accepted September 21, 2017. Date of publication November 16, 2017; date of current version March 9, 2018. This work was supported by JSPS KAKENHI under Grant JP16K06356. This paper was recommended by Associate Editor G. Russo. *(Corresponding author: Shuichi Ohno.)*

S. Ohno and Y. Ishihara are with Hiroshima University, Higashihiroshima 739-8527, Japan (e-mail: o.shuichi@ieee.org).

M. Nagahara is with the Institute of Environmental Science and Technology, The University of Kitakyushu, Fukuoka 808-0135, Japan.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSI.2017.2758801

l_∞ norm of the error at the system output induced by the quantizer with a no-overloading quantizer. The minimum data rate to keep the state of a closed-loop system in a bounded region with state feedback control has been provided in [14]. Stabilizability and observability under a communication constraint has been studied in [15] for discrete-time, linear and time-invariant (LTI) systems. Then, an LQG control with minimum directed information has been developed in [16]. Although these information theoretical analyses give valuable insights into control under limited data rates, the system may not always work well in practice, since the minimum data rate is not a constant rate for each time slot but an averaged rate over time. On the other hand, our quantizer guarantees the stability for a fixed bit rate. The error due to the quantization is evaluated by finding the relationship between l_∞ norm of the feedback signal and the l_∞ norm of the error in the output, which can be obtained by solving convex optimization problems.

Finally, numerical examples are provided to validate our analysis and synthesis.

This paper is organized as follows: Systems and quantization are reviewed in Section II. Relevance to circuits and systems is shortly discussed in Section III. Then, quantizers are synthesized in Section IV based on the l_∞ norm of the effect of the quantization error and the output of the error feedback filter. Section V presents numerical results on our synthesis and Section VI concludes this paper.

Notation: $\mathbb{Z}$, $\mathbb{R}$, and $\mathbb{R}_+$ stand for the set of real numbers, integers, and non-negative real numbers, respectively. The z transform of a sequence (or a vector) $h = \{h_k\}_{k=0}^{\infty}$ is denoted as $H[\mathrm{z}] = \sum_{k=0}^{\infty} h_k \mathrm{z}^{-k}$. The output sequence y of an linear and time-invariant (LTI) system $H[\mathrm{z}]$ with the input sequence x (i.e. $y = h * x$ where $*$ denotes the convolution) is expressed as $y = H[\mathrm{z}]x$. The l_∞ signal space is defined as the set of all vectors $x = \{x_k\}_{k=0}^{\infty}$ with real components x_k such that $\|x\|_\infty := \max_k |x_k| < +\infty$. The norm of $H[\mathrm{z}]$ induced by the l_∞ norm of the input and output signals is defined as [17]

$$\|H[\mathrm{z}]\| = \sup_{x \neq 0} \frac{\|H[\mathrm{z}]x\|_\infty}{\|x\|_\infty}$$

for $x \in l_\infty$. If $H[\mathrm{z}]$ is an single-input and single-output system, the norm is equivalent to the l_1 norm of the impulse response of the system, that is,

$$\|H[\mathrm{z}]\| = \sum_{k=0}^{\infty} |h_k|.$$

II. Error Feedback Quantizer

To see the effect of quantization errors at the output of the system connected to a quantizer, let us consider quantization in a feedback control system depicted in Fig. 1, in which the plant is assumed to be linear and time-invariant (LTI), and the signals y and u are functions of time in general. Based on the observation signal y from the plant, the controller generates the control input u to the plant.

The observation signal y and the control input u are assumed to be transmitted through digital communication channels. If y and u are real-valued signals, quantization is required to convert them into discrete-valued signals before transmission as illustrated in Fig. 2. Note that even if y and u are discrete-valued digital signals, they may have to be rounded off if the capacities of the communication channels are limited.

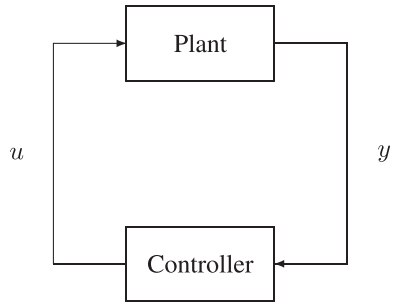

Fig. 1. Feedback control system.

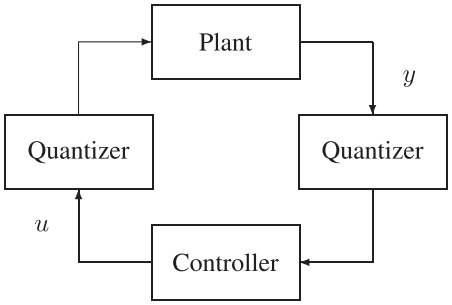

Fig. 2. Control system with quantization.

The difference between the input and the output of the quantizer is called the quantization error. There are two quantization errors; one is the quantization error denoted by e_c for the control signal u and the other is the quantization error e for the observation signal y. With these quantization errors, the control system in Fig. 2 can be modeled by an additive-noise control system shown in Fig. 3.

Controllers are often connected to plants through wired networks. On the other hand, the observation signal are collected by sensors, which may be connected through wireless networks. Thus, we here focus on the quantization error e of the observation signal, assuming that there is no quantization error at the controller. We assume that each sensor observes a scalar-valued signal to be quantized and works independently of the other sensors. Since we consider the independent quantization of each entries of y, we assume y to be a scalar-valued signal to a particular quantizer for simplicity of presentation. We also assume that the plant is a single-input and single-output (SISO) system. Note that, most of our results may be applied to the quantization error at the controller and the multiple-input and multiple-output (MIMO) systems. We assume the reachability and the observability of the plant, without which the plant cannot be stabilized in general.

In quantization, real numbers are mapped into their binary representation. Fixed-point representation and floating-point representation are available for quantization. In this paper, we take fixed-point representation into account since it is often adopted in embedded systems.

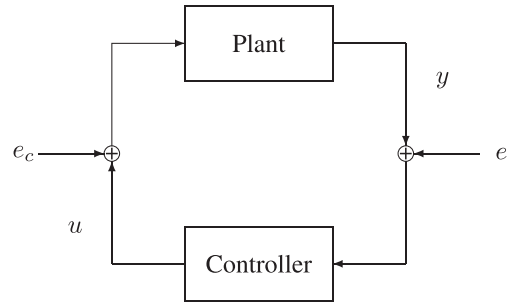

Fig. 3. Control system and quantization error signals.

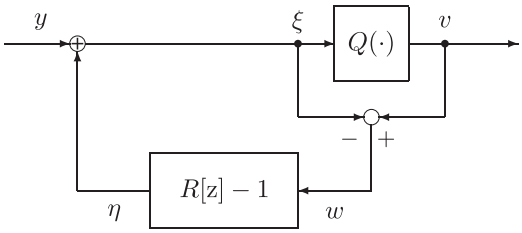

Fig. 4. Error feedback quantizer.

Let us take a static uniform quantizer for example. The static uniform quantizer can be described by two parameters, the quantization interval $d \in \mathbb{R}_+$ and the saturation level $L \in \mathbb{R}_+$. For simplicity, we assume that L is an integer multiple of d. For the static quantizer, let us consider a mid-rise quantizer[1] $Q(\xi)$ expressed as

$$Q(\xi) = \begin{cases} \left(i + \frac{1}{2}\right) d, & |\xi| \le L + \frac{d}{2} \\ & \text{and } \xi \in [id, (i+1)d), \quad i \in \mathbb{Z} \\ L, & \xi > L + \frac{d}{2} \\ -L, & \xi < -L - \frac{d}{2} \end{cases} \tag{1}$$

The overloading is the saturation due to the fixed number of bits to represent the quantized values in binary. For the mid-rise quantizer, the overloading occurs if $|\xi| > L + \frac{d}{2}$.

The static uniform quantizer is often utilized in practice but its errors and effects of the overloading are significant unless a sufficient number of bits is assigned to the quantizer. To mitigate these influences, we adopt a quantizer with an error feedback filter.

Fig. 4 illustrates a block diagram of our quantizer. The quantization error, or the round-off error, of the static uniform quantizer $Q(\cdot)$ is defined as

$$w = v - \xi \tag{2}$$

where ξ and v are the input and the output vectors of the static uniform quantizer, respectively. Note that the round-off error w of the static quantizer is different from the quantization error defined as

$$e = v - y. \tag{3}$$

The round-off error signal w is filtered by the error feedback filter $R[\mathrm{z}] - 1$ and then it is fed back to the input to the static uniform quantizer. The error feedback filter $R[\mathrm{z}] - 1$ has to be strictly proper, that is, $R[\infty] = 1$. The error feedback quantizer in Fig. 4 is also known as a $\Delta\Sigma$ modulator, which is an efficient analog to digital (A/D) converter with feedback from the output of a static uniform quantizer to shape the quantization noise [4].

[1]Similar results can be obtained for mid-tread quantizers with slight modifications.

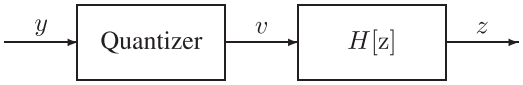

Fig. 5. Quantizer and system.

The input signal ξ to the static quantizer can be expressed from Fig. 4 as

$$\xi = y + (R[\mathrm{z}] - 1)w. \tag{4}$$

From (2) and (4), the quantization error e is given by

$$e = v - y = R[\mathrm{z}]w. \tag{5}$$

Thus, the output signal of the quantizer can be expressed as

$$v = y + e = y + R[\mathrm{z}]w. \tag{6}$$

Let the signal of interest in Fig. 3 be z and the transfer function from the output v of the quantizer to z be $H[\mathrm{z}]$. Fig. 5 illustrates an equivalent system from y to z. Since the plant is assumed to be reachable and observable, there exists a controller that stabilizes the control system when there is no quantization error. With this controller, $H[\mathrm{z}]$ is stable. Thus, without loss of generality, we can assume that $H[\mathrm{z}]$ is stable.

Since e also goes through $H[\mathrm{z}]$, the error signal in z that comes from the quantization error e can be expressed as

$$\epsilon = H[\mathrm{z}]e = H[\mathrm{z}]R[\mathrm{z}]w. \tag{7}$$

Unless $H[\mathrm{z}]R[\mathrm{z}] = 0$, we cannot assure that $\epsilon_k \to 0$ as $k \to \infty$ due to the unpredictable round-off errors. Thus, we cannot guarantee the exponential stability of the feedback system, even if $H[\mathrm{z}]$ is stable. All we can do is to mitigate the effect of the quantization given by (7). Thus, our goal is to design an error feedback quantizer so that the maximum absolute value of ϵ is not greater than a prescribed threshold γ_ϵ, which can be expressed as

$$\max_k |\epsilon_k| \le \gamma_\epsilon \tag{8}$$

or equivalently as

$$\|\epsilon\|_\infty \le \gamma_\epsilon. \tag{9}$$

III. Relevance of the Results for Circuits and Systems

Error feedback quantizers have been used in many fields such as circuits and systems, signal processing, communications, image processing, and control engineering.

Quantization with error feedback has been developed to reduce quantization error in the coefficients of digital filters [1]–[3], where $H[\mathrm{z}]$ in (7) is the filter to be realized and $R[\mathrm{z}]$ is called an error spectrum shaping filter. The optimal finite impulse response (FIR) filter that minimizes the mean squared error (MSE) has been designed in [2]. Error spectrum

shaping has also been utilized for two-dimensional recursive digital filters [18].

$\Delta\Sigma$ modulators are often used as A/D or D/A converters, since they exhibit better performance than conventional converters at the expense of a relatively low cost. To develop single-bit RF transmitters, bandpath $\Delta\Sigma$ modulators can be utilized [19]. A linearized model of the $\Delta\Sigma$ modulator can be described by the same structure with the quantizer having error feedback. The filter $R[z]$ in a $\Delta\Sigma$ modulator is called a noise shaping filter or a noise transfer function [4].

There are a lot of works on $\Delta\Sigma$ modulators; see e.g. [4] and the references therein. Design methodologies for $\Delta\Sigma$ modulators are well documented in [20]. Once the architecture is fixed, the first step is the design of the noise shaping filter $R[z]$, which is the main topic of this paper.

The noise shaping filter has been designed in [9] and [10] to reduce e in (5) (in place of ϵ). However, it is not necessarily optimal for the reduction of ϵ, since the system $H[z]$ whose input is generated by the $\Delta\Sigma$ modulator is not taken into account.

Recently, the importance of the design based on the system $H[z]$ connected to the $\Delta\Sigma$ modulator is recognized. The knowledge about the system $H[z]$ is incorporated to design an error feedback filter [11]. The optimal FIR filter that minimizes the variance of ϵ is developed in [12] where $H[z]$ in (7) is considered as the weighting function. In [8], the noise shaping infinite impulse response (IIR) filter is designed to minimize the variance of ϵ under the constraint on the variance of the feedback signal.

If the quantization error of the uniform quantizer can be approximated as a white uniformly distributed random sequence, the variances of η and ϵ can be characterized by H_2 system norms. To see the performance of $\Delta\Sigma$ modulators, the relation between the variances is analyzed in [21]. However, the constraint on the variances of the feedback signal η cannot always assure no-overloading. Thus, this paper considers the l_∞ norm of the error ϵ at the output of the system, which can be more important than the variance in some applications that use $\Delta\Sigma$ modulators as A/D or D/A converters.

To stabilize the $\Delta\Sigma$ modulator, Lee criterion [22] is often utilized, which constraints the H_∞ norm of $R[z]$ such that

$$\|R[z]\|_\infty = \max_\omega |R[e^{j\omega}]| < \gamma \tag{10}$$

for some $\gamma > 0$. However, Lee criterion does not necessarily guarantee no-overloading of the $\Delta\Sigma$ modulator. To strictly stabilize the $\Delta\Sigma$ modulator, we have to evaluate the l_∞ norm of the feedback signal η.

For these reasons, this paper deals with the l_∞ norms of signals and designs a no-overloading error feedback quantizer that satisfies the l_∞ constraint on the error ϵ at the output of the system connected to a $\Delta\Sigma$ modulator.

IV. Synthesis of Error Feedback Quantizer

Unless overloading occurs, the round-off error w is bounded such as

$$\|w\|_\infty \le \frac{d}{2}. \tag{11}$$

Otherwise, the signal z of interest cannot be bounded in general, since w may be unbounded due to overloading. Then, the overloading complicates the system behavior, since $\|\epsilon\|_\infty$ depends on it. To design the quantizer and the system connected to the quantizer independently, it is better to avoid the overloading in the static uniform quantizer.

If there is no overloading, then y is bounded, since the system is stable and the error ϵ is bounded with a stable $R[z]$. Without loss of generality, we can assume that the observation signal has the symmetric magnitude limitation described as

$$\|y\|_\infty \le L_y. \tag{12}$$

Let us adopt the static uniform quantizer characterized by (1) in our error feedback quantizer. From our definitions, if the feedback signal η meets

$$\|y+\eta\|_\infty \le L + \frac{d}{2}, \tag{13}$$

then no overloading happens at the static quantizer.

It follows from the triangle inequality $\|y+\eta\|_\infty \le \|y\|_\infty + \|\eta\|_\infty$ and $\|\eta\|_\infty \le \|R[z]-1\|(d/2)$ that if one sets

$$L_y + \|R[z]-1\|\frac{d}{2} \le L + \frac{d}{2} \tag{14}$$

then no-overloading in the static uniform quantizer is assured. In other words, the l_∞ norm of the feedback signal should be equal to or less than $L + d/2 - L_y$.

For the binary representation of the observation signals, we have to determine its accuracy and range, i.e., the quantization interval d and the saturation level L for the uniform quantizer. If we assign b bits to represent the observation signal, we have

$$2L = (2^b - 1)d. \tag{15}$$

From (14) and (15), we summarize the above discussion as a proposition:

Proposition 1: There is no overloading in the error feedback quantizer composed of an error feedback filter $R[z]-1$ and a mid-rise quantizer if

$$L_y + \|R[z]-1\|\frac{d}{2} \le 2^{b-1}d \tag{16}$$

where b is the number of bits assigned to the mid-rise quantizer, d and L_y denote its quantization interval and saturation level respectively.

It should be remarked that (16) is a sufficient condition for no-overloading. Since the triangle inequality is loose in general, the design based on (16) may be conservative.

Now, we would like to find the number of bits that assures no-overloading under the constraint (9), which is, from (7), achieved if

$$\|H[z]R[z]\|\frac{d}{2} \le \gamma_\epsilon. \tag{17}$$

To obtain the minimum b that satisfies (16), we set the quantization interval of our static uniform quantizer to be

$$d = \frac{2\gamma_\epsilon}{\|H[z]R[z]\|}. \tag{18}$$

Substituting (18) into (16) leads to

$$\frac{L_y}{\gamma_\epsilon}\|H[z]R[z]\| + \|R[z] - 1\| \le 2^b. \tag{19}$$

For given L_y and γ_ϵ, the left hand side of the inequality above can be evaluated with $\|H[z]R[z]\|$ and $\|R[z] - 1\|$, whose minimum can be obtained by solving the following optimization problem:

$$\min_{R[z]\in RH_\infty, \tilde{\gamma}_\epsilon, \tilde{\gamma}_\eta} \; c\tilde{\gamma}_\epsilon + \tilde{\gamma}_\eta \tag{20}$$

subject to $R[\infty] = 1$ and

$$\|H[z]R[z]\| \le \tilde{\gamma}_\epsilon \tag{21}$$

$$\|R[z] - 1\| \le \tilde{\gamma}_\eta \tag{22}$$

where RH_∞ is the set of stable proper rational functions with real coefficients and

$$c = \frac{L_y}{\gamma_\epsilon}. \tag{23}$$

It should be noted that the objective function is linear in $\tilde{\gamma}_\epsilon$ and $\tilde{\gamma}_\eta$.

The problem above can be solved if we restrict $R[z]$ to have a finite impulse response. On the other hand, the global optimal solution is not available for general IIR filters.

A. FIR Filter Design

If $R[z]$ is an FIR filter of order n, then the problem can be formulated as a linear programming (LP) and be numerically solved as follows.

To solve the problem, we express the composite system $H[z]R[z]$ as a state-space realization. We denote the state-space matrices of a state-space realization of $H[z]$ as (A_h, B_h, C_h, D_h), while the state-space matrices of a state-space realization of $R[z]$ as (A_r, B_r, C_r, D_r) with $D_r = 1$. Then, the state-space realization of $H[z]R[z]$ can be written as

$$x_{k+1} = \mathcal{A}x_k + \mathcal{B}w_k \tag{24}$$

$$\epsilon_k = \mathcal{C}x_k + \mathcal{D}w_k \tag{25}$$

where the state-space matrices for this are given as

$$\mathcal{A} = \begin{bmatrix} A_r & B_rC_h \\ \mathbf{0} & A_h \end{bmatrix} \tag{26}$$

$$\mathcal{B} = \begin{bmatrix} B_r \\ B_h \end{bmatrix} \tag{27}$$

$$\mathcal{C} = \begin{bmatrix} C_r & D_rC_h \end{bmatrix} \tag{28}$$

$$\mathcal{D} = D_h. \tag{29}$$

It is noted that the impulse response from w to ϵ can be expressed with the space-space matrices as

$$f_k = \begin{cases} \mathcal{D}, & k = 0 \\ \mathcal{C}\mathcal{A}^{k-1}\mathcal{B}, & k = 1, 2, 3, \cdots \end{cases} \tag{30}$$

A state-space realization $(A_r, B_r, C_r, 1)$ of the FIR filter $R[z] = 1 + \sum_{k=1}^{n} r_k z^{-k}$ is given by

$$A_r = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \vdots & & & \ddots & 1 \\ 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix}, \quad B_r = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{31}$$

$$C_r = \begin{bmatrix} r_n, & r_{n-1}, & \cdots & r_1 \end{bmatrix}, \quad D_r = 1. \tag{32}$$

Since A_r and B_r are constant, $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{D}$ are constant. Moreover, $\mathcal{A}$ is Schur, that is, all eigenvalues of $\mathcal{A}$ are strictly inside the unit circle, since $H[z]$ is stable.

For a sufficiently large integer m, we can approximate $\|H[z]R[z]\|$ such that

$$\|H[z]R[z]\| = |D_h| + \sum_{k=1}^{m} |\mathcal{C}\mathcal{A}^{k-1}\mathcal{B}|. \tag{33}$$

On the other hand, we have from $R[z] - 1 = \sum_{k=1}^{n} r_k z^{-k}$ that

$$\|R[z] - 1\| = \sum_{k=1}^{n} |r_k|. \tag{34}$$

Then, our problem can be expresses as the following minimization problem:

$$\min_{r_1,\ldots,r_n,\tilde{\gamma}_\epsilon,\tilde{\gamma}_\eta} \; c\tilde{\gamma}_\epsilon + \tilde{\gamma}_\eta \tag{35}$$

subject to

$$|D_h| + \sum_{k=1}^{m} |\mathcal{C}\mathcal{A}^{k-1}\mathcal{B}| \le \tilde{\gamma}_\epsilon \tag{36}$$

$$\sum_{k=1}^{n} |r_k| \le \tilde{\gamma}_\eta. \tag{37}$$

Note that the matrix $\mathcal{C}$ depends linearly on $r_1, \ldots, r_n$ as in (28) and (32). However, since the absolute values of the variables are not linear, we cannot easily solve the problem directly. To cast the problem into a solvable linear programming (LP), we introduce auxiliary variables.

Putting $\bar{f}_k = |\mathcal{C}\mathcal{A}^{k-1}\mathcal{B}|$ for $k = 1, \ldots, m$ in (36), we have

$$|D_h| + \sum_{k=1}^{m} \bar{f}_k \le \tilde{\gamma}_\epsilon. \tag{38}$$

It follows from $\bar{f}_k = |\mathcal{C}\mathcal{A}^{k-1}\mathcal{B}|$ that

$$-\bar{f}_k \le \mathcal{C}\mathcal{A}^{k-1}\mathcal{B} \le \bar{f}_k \quad \text{for } k = 1, \ldots, m \tag{39}$$

Similarly, with non-negative auxiliary variables $\bar{r}_k \ge 0$ for $k = 1, \ldots, n$, (37) is found to be equivalent to

$$\sum_{k=1}^{n} \bar{r}_k \le \tilde{\gamma}_\eta \tag{40}$$

$$-\bar{r}_k \le r_k \le \bar{r}_k \quad \text{for } k = 1, \ldots, n. \tag{41}$$

Then, our minimization problem is formulated as the following LP:

$$\min_{r_1,\ldots,r_n,\bar{f}_1,\ldots,\bar{f}_m,\bar{r}_1,\ldots,\bar{r}_n,\tilde{\gamma}_\epsilon,\tilde{\gamma}_\eta} \; c\tilde{\gamma}_\epsilon + \tilde{\gamma}_\eta \tag{42}$$

subject to (38), (39), (40), (41), and

$$\bar{f}_k \geq 0 \quad \text{for } k = 1, \ldots, m \tag{43}$$

$$\bar{r}_k \geq 0 \quad \text{for } k = 1, \ldots, n. \tag{44}$$

Our noise shaping filter $R[z]$ is designed by solving this LP numerically [23].

B. IIR Filter Design

Let us briefly introduce the IIR filter design, where the order of $R[z]$ is set to be equal to the order of $H[z]$. For the design of IIR filters, we re-express the state-space realization of $H[z]R[z]$ as

$$\mathcal{A} = \begin{bmatrix} A_h & B_h C_r \\ \mathbf{0} & A_r \end{bmatrix} \tag{45}$$

$$\mathcal{B} = \begin{bmatrix} B_h \\ B_r \end{bmatrix} \tag{46}$$

$$\mathcal{C} = \begin{bmatrix} C_h & D_h C_r \end{bmatrix}. \tag{47}$$

In [24], the following lemma has been provided by using the invariant set of a discrete-time system.

Lemma 1: Suppose the initial state x_0 is 0 and the input w is bounded as $\|w\|_\infty \leq 1$. Then, the state vectors $x_k, k = 1, 2, \ldots$ remain in the ellipsoid

$$\mathcal{E}(\mathcal{P}) = \{x : x^T \mathcal{P} x \leq 1\} \tag{48}$$

if and only if there exist a scalar $\alpha \in [0, 1 - \rho^2(\mathcal{A})]$ and a positive definite matrix $\mathcal{P}$ satisfying

$$\begin{bmatrix} (1-\alpha)\mathcal{P} & \mathbf{0} & \mathcal{A}^T\mathcal{P} \\ \mathbf{0} & \alpha & \mathcal{B}^T\mathcal{P} \\ \mathcal{P}\mathcal{A} & \mathcal{P}\mathcal{B} & \mathcal{P} \end{bmatrix} \succeq \mathbf{0}, \tag{49}$$

where $\rho(A)$ is the spectrum radius of A.

It follows from $\epsilon_k = \mathcal{C}x_k + D_h w_k$ that if $x_k \in \mathcal{E}(\mathcal{P})$, then

$$\sup_{x_k \in \mathcal{E}(P)} |\epsilon_k - D_h w_k|^2 = \sup_{x_k \in \mathcal{E}(P)} |\mathcal{C}x_k|^2 \tag{50}$$

$$= \sup_{\tilde{x}_k \in \mathcal{E}(I)} |\mathcal{C}\mathcal{P}^{-\frac{1}{2}}\tilde{x}_k|^2 \tag{51}$$

$$\leq \text{trace}\left(\mathcal{C}\mathcal{P}^{-1}\mathcal{C}^T\right) \tag{52}$$

where I is an identity matrix. From the triangle inequality for the absolute value, we have

$$\|\epsilon\|_\infty \leq |D_h|\frac{d}{2} + \left[\text{trace}\left(\mathcal{C}\mathcal{P}^{-1}\mathcal{C}^T\right)\right]^{\frac{1}{2}}. \tag{53}$$

On the other hand, with

$$\tilde{\mathcal{C}} = \begin{bmatrix} \mathbf{0} & C_r \end{bmatrix} \tag{54}$$

we can express η as

$$\eta_k = \tilde{\mathcal{C}}x_k \tag{55}$$

which leads to

$$\|\eta\|_\infty \leq \left[\text{trace}\left(\tilde{\mathcal{C}}\mathcal{P}^{-1}\tilde{\mathcal{C}}^T\right)\right]^{\frac{1}{2}}. \tag{56}$$

Unlike FIR filters, we cannot evaluate $\|\epsilon\|_\infty$ and $\|\eta\|_\infty$ directly. Instead, we consider the minimization using the right hand sides of (53) and (56), that is, the upper bounds of $\|\epsilon\|_\infty$ and $\|\eta\|_\infty$, such that:

$$\text{trace}\left(\mathcal{C}\mathcal{P}^{-1}\mathcal{C}^T\right) \leq \mu_\epsilon \tag{57}$$

$$\text{trace}\left(\tilde{\mathcal{C}}\mathcal{P}^{-1}\tilde{\mathcal{C}}^T\right) \leq \mu_\eta. \tag{58}$$

Note that the upper bound of our objective function $c\tilde{\gamma}_\epsilon + \tilde{\gamma}_\eta$ is given by $c\sqrt{\mu_\epsilon} + \sqrt{\mu_\eta}$, which is not convex in μ_ϵ and μ_η. Instead of directly solving the problem, let us consider the following problem:

$$\min_{R[z] \in RH_\infty, x_k \in \mathcal{E}(\mathcal{P}), \mu_\epsilon} \mu_\epsilon \tag{59}$$

subject to $R[\infty] = 1$, (57) and (58).

The condition $x_k \in \mathcal{E}(\mathcal{P})$ is described by (49), which is a bilinear matrix inequality (BMI) of the variables. On the other hand, by using the Schur complement, (57) and (58) can be expressed as linear matrix inequalities (LMIs):

$$\begin{bmatrix} \mathcal{P} & \mathcal{C}^T \\ \mathcal{C} & \mu_\epsilon \end{bmatrix} \succeq \mathbf{0} \tag{60}$$

$$\begin{bmatrix} \mathcal{P} & \tilde{\mathcal{C}}^T \\ \tilde{\mathcal{C}} & \mu_\eta \end{bmatrix} \succeq \mathbf{0}. \tag{61}$$

Since the BMI is not convex, we cannot yet solve (59). Fortunately, we can convert the BMI into an LMI and then, since the LMI is convex, we can solve the minimization problem (59) numerically as detailed in Appendix.

By solving the convex optimization problem (59) for different values for μ_η, we obtain the optimal IIR feedback filters for different constraints on the l_∞ norms of the feedback signals. With the designed feedback filters $R[z] - 1$, we can evaluate the pair $(\|R[z] - 1\|, \|H[z]R[z]\|)$ and find the optimal $R[z]$ that achieves the minimum of the left hand side of (19), which gives the minimum number for b.

C. Minimum l_∞ Norm for a Fixed Number of Bits

We have investigated the number of bits required for quantization to attain a prescribed requirement on the performance and the no-overloading quantization at the same time. Now, we would like to consider another problem to find the achievable minimum l_∞ norm of the error in the signal of interest with a no-overloading quantizer for a given number of bits.

Suppose that the number of bits assigned to the static quantizer is given and fixed. We would like to design the no-overloading error feedback quantizer that minimizes the l_∞ norm of the error ϵ. From (16), we obtain

$$L_y \leq \left(2^{b-1} - \frac{1}{2}\|R[z] - 1\|\right) d. \tag{62}$$

Since d must be positive, we have to meet $\|R[z] - 1\| < 2^b$. It follows from $\|\epsilon\|_\infty \leq \|H[z]R[z]\|d/2$ that $\|\epsilon\|_\infty$ is bounded with $d = L_y / \left(2^{b-1} - \|R[z] - 1\|/2\right)$ as

$$\|\epsilon\|_\infty \leq \frac{L_y \|H[z]R[z]\|}{2^b - \|R[z] - 1\|}. \tag{63}$$

It is obvious that for a fixed value of $\|R[z] - 1\|$, the upper bound for $\|\epsilon\|_\infty$ is minimized by the filter $R[z]$ that minimizes $\|H[z]R[z]\|$. Then, for a fixed upper bound $\tilde{\gamma}_\eta$ of

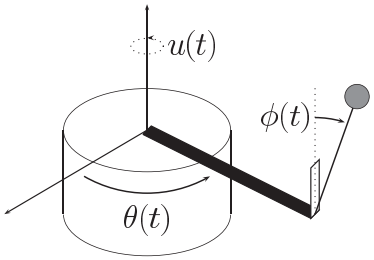

Fig. 6. Rotary inverted pendulum.

$\|R[z]-1\|$, the optimal filter can be found by solving the following optimization problem:

$$\min_{R[z]\in RH_\infty,\tilde{\gamma}_\epsilon} \tilde{\gamma}_\epsilon \tag{64}$$

subject to $R[\infty]=1$, (21), and (22).

For a fixed upper bound for $\|R[z]-1\|$, we have the value for $\|H[z]R[z]\|$ by solving the optimization problem. Then, by solving the problem for different values for $\|R[z]-1\|$, the relationship between $\|R[z]-1\|$ and $\|H[z]R[z]\|$ can be obtained. Finally, with the values for the pair ($\|R[z]-1\|$, $\|H[z]R[z]\|$), we can obtain the minimum of the right hand side of (63).

As in Sections IV-A and IV-B, the problem can be formulated as an LP for FIR $R[z]$, whereas the upper bounds of the norms have to be evaluated for IIR $R[z]$. The details are omitted, since the optimization problems can be solved similarly as described in IV-A and IV-B.

V. Numerical Results

We here consider the quantization for the rotary inverted pendulum (see e.g. [25]) depicted in Fig. 6 for our design example.

The pendulum connected at the end of the rotary arm is controlled by rotating the main body in the horizontal plane. The yaw angle of the arm is $\theta(t)$. The pendulum freely swings about a pitch angle $\phi(t)$ in the vertical plane to the arm. The torque $u(t)$ is applied to actuate the pendulum. If $\phi(t)=0$, then the pendulum is balanced in the inverted position.

We define the state of the rotary inverted pendulum as

$$x^T(t)=[\phi(t),\theta(t),\dot{\phi}(t),\dot{\theta}(t)]. \tag{65}$$

We periodically change the yaw angle, while keeping the stability of the rotary inverted pendulum. The target value of the yaw angle $\bar{\theta}(t)$ is

$$\bar{\theta}(t)=\begin{cases}\dfrac{\pi}{2} & (10k\le t<5+10k)\\ 0 & (5+10k\le t<10(k+1))\end{cases} \tag{66}$$

for $k=0,1,\ldots$. The initial values of the states are assumed to be zero.

We linealize the equations of motions about the upward equilibrium, that is $\phi(t)=0$, and derive the zero-order hold equivalent discrete-time model [26] with the sampling period $T_s=0.01$. Let A, B, C, D be the state-space matrices of the linearized and discretized system. Since the continuous-time system is strictly proper, we have $D=0$. The state-space matrices A and B of the discrete-time linearized system are given by

$$A=\begin{bmatrix}1.0056 & 0 & 0.0100 & 0.0001\\ -0.0003 & 1.0000 & -0.0000 & 0.0100\\ 1.1134 & 0 & 1.0056 & 0.0149\\ -0.0653 & 0 & -0.0003 & 0.9926\end{bmatrix} \tag{67}$$

$$B=\begin{bmatrix}-0.0004\\ 0.0002\\ -0.0864\\ 0.0431\end{bmatrix}. \tag{68}$$

Assuming that all of the state variables be available at the controller (i.e. C is the identity matrix), we adopt the state feedback control and determined its gain K by the linear quadratic regulator (LQR) technique to minimize

$$\sum_{k=0}^{\infty}\left(x_k^T Q_{lqr} x_k + r|u_k|^2\right) \tag{69}$$

where the weights are

$$Q_{lqr}=\mathrm{diag}[10,2,0.5,0],\quad r=0.05. \tag{70}$$

Our signal of interest is the discretized $\phi(t)$, which is expressed as $\phi_k=C_1\ x_k$ with

$$C_1=\begin{bmatrix}1 & 0 & 0 & 0\end{bmatrix}. \tag{71}$$

The transfer function from the lth entry of the quantization error to ϕ is found to be $C_1(zI-A-BK)^{-1}BK_l$ with $K=[K_1,K_2,K_3,K_4]=[57.2598, 6.0910, 6.2562, 3.4953]$.

Now let us design FIR and IIR filters of order 4 for the error feedback. We consider the quantization of ϕ, the first entry of the state variables, to mitigate the effect of the quantization on ϕ. The transfer function is given by

$$\begin{aligned}H[z]&=C_1(zI-A-BK)^{-1}BK_1\\ &=\frac{0.02475z^3-0.02482z^2-0.02463z+0.02469}{z^4-3.59z^3+4.808z^2-2.844z+0.626}\end{aligned} \tag{72}$$

whose zeros are -0.9975, 1, and 1.

The constraint on the maximum absolute value of ϵ is set to be $\gamma_\epsilon=0.05$ and L_y in (12) is set to be $\pi/2$.

With the designed optimal FIR feedback filter, the value of the objective function $c\|H[z]R[z]\|+\|R[z]-1\|$ is 5.2729. On the other hand, the value with the designed IIR feedback filter is 5.8831. The FIR filter exhibits a better performance than the IIR filter. This is due to the fact that the exact value of the l_∞ norm is evaluated for the design of the FIR filter, whereas only an upper bound can be used for the design of the IIR filter. In both cases, the required number of bits is 3 to satisfy the constraint (19), while the conventional static uniform quantizer requires 6 bits to meet the constraint on ϵ, since $(L_y/\gamma_\epsilon)\|H[z]\|=47.788<2^6$.

The signal-to-quantization-noise-ratio (SQNR) is a performance measure for quantization errors, SQNR is evaluated at the output of the quantizer and is defined as the ratio of the variance of the input y to the variance of the error e in Fig. 5. The uniform quantizer with 6 bits achieves a better SQNR than the feedback quantizer with 3 bits, since it utilizes a larger number of bits than the error feedback quantizer.

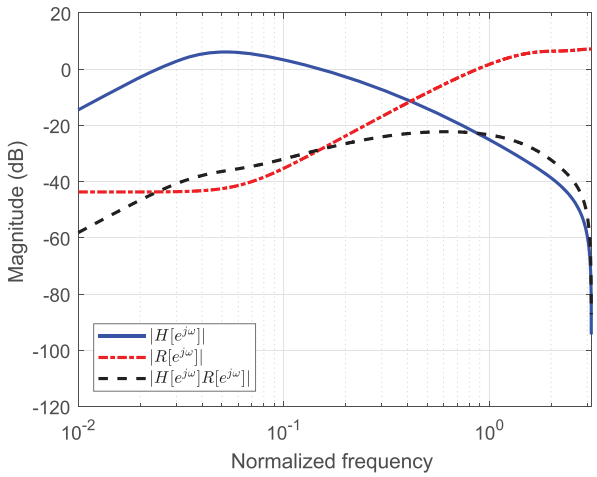

Fig. 7. Frequency responses of the system $H[z]$, the designed FIR filter $R[z]$, and the transfer function $H[z]R[z]$ from the quantization error to the system output.

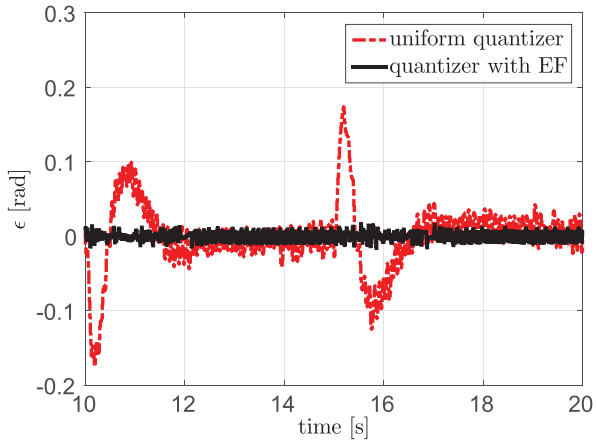

Fig. 8. Error of pitch angles of the pendulum controlled with the 3-bit quantizers having the optimized FIR error feedback filter (solid line) and the conventional uniform quantizers (dash-dotted line).

However, comparing the two quantizers having an identical number of bits, e.g. 3 bits, we can conclude that the uniform quantizer has a larger error ϵ at the output of the system $H[z]$.

Fig. 7 depicts frequency responses of the system $H[z]$, the designed FIR filter $R[z]$, and the transfer function $H[z]R[z]$ from the quantization error to the system output. The designed FIR filter has small/large gains at the pass-band/stopband of $H[z]$. Although $c\|H[z]R[z]\|+\|R[z]-1\|$ is minimized, the gain of $H[z]R[z]$ is small at every frequency.

Simulations are conducted with the designed optimal FIR error feedback filter. To clarify the difference, we only quantize the signal ϕ of interest.

Fig. 8 compares the error signal ϵ of the pendulum controlled with the 3-bit quantizers having the optimized FIR error feedback filter (solid line) and the conventional uniform quantizer (dash-dotted line) for $10 \leq t < 20$. The maximum absolute value of the error for our designed quantizer is less than 0.05, while the maximum absolute value of the error for the conventional uniform quantizers is about 0.18. Our designed quantizer satisfies the requirement on the error clearly outperforms the conventional uniform quantizer.

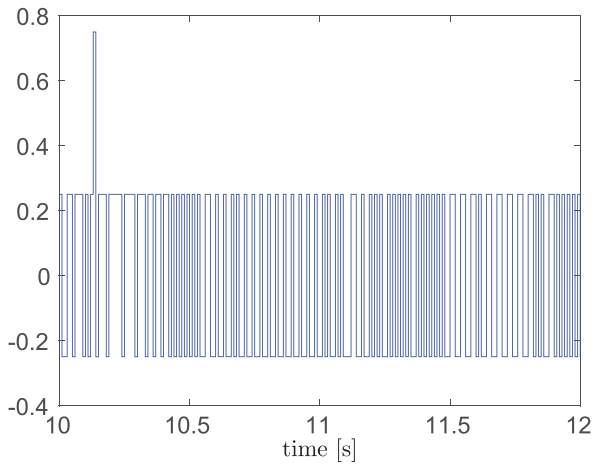

Fig. 9. Output of the designed quantizer for pitch angle.

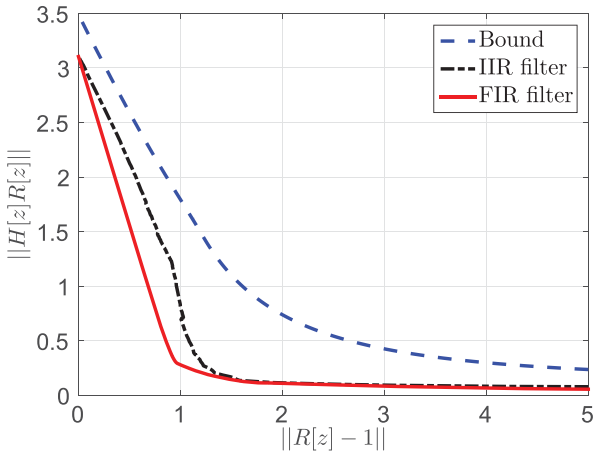

Fig. 10. Relation between $\|R[z]-1\|$ and $\|H[z]R[z]\|$ for pitch angle ϕ.

The output of our designed quantizer for $10 \leq t < 12$ is shown in Fig. 9. Only three values are taken, which implies that only 2 bits are required in practice, although our analysis suggests 3 bits. This is because we adopt the worst-case error for our performance measure. Indeed, it is well-known that the condition on the maximum of the absolute value of errors leads to conservative results.

Next, for a fixed number of bits for the quantizer, we evaluate the l_∞ norm of the error in the signal of interest with the designed no-overloading error feedback quantizer. We solve the optimization problems discussed in Section IV-C.

Fig. 10 depicts $\|H[z]R[z]\|$ as a function of $\|R[z]-1\|$. In the design of IIR filters, we minimize the upper bound and then the designed filter is not assured to be optimal. Here, $(\sqrt{\mu_\eta}, \sqrt{\mu_\epsilon})$ serves as an upper bound for $(\|R[z]-1\|, \|H[z]R[z]\|)$ of IIR filters. On the other hand, in the design of FIR filters, we minimize the objective function directly and the designed filter is optimal among FIR filters. This may be a reason why the designed FIR filters achieve smaller error norm than the designed IIR filers.

As $\|R[z]-1\|$ increases from zero, $\|H[z]R[z]\|$ decreases rapidly at first and then floors. It should be remarked that $\|R[z]-1\|=0$ implies that there is no error feedback filter, that is, the quantizer is just a static uniform quantizer.

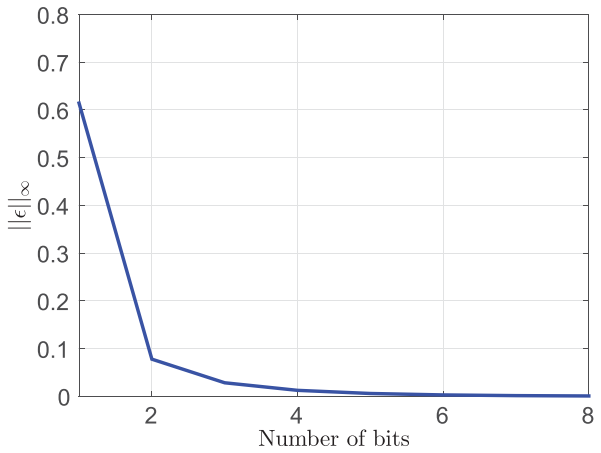

Fig. 11. $\|\epsilon\|_\infty$ for different numbers of bits.

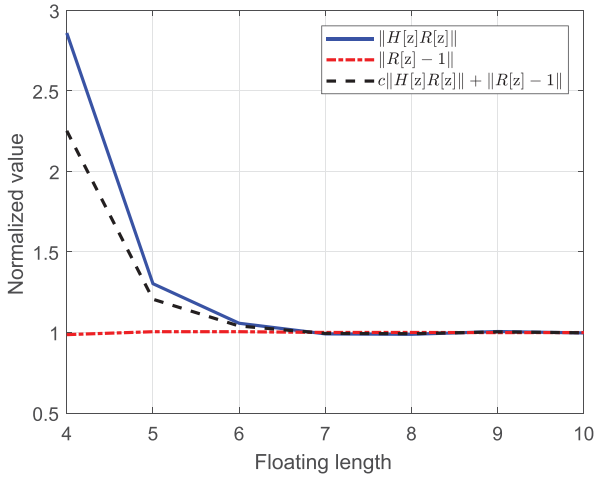

Fig. 12. Ratios between values by 16-bit fixed-point binary numbers with different floating length and floating-point numbers.

From the values of $(\|R[z]-1\|, \|H[z]R[z]\|)$ in Fig. 10, we compute the norm $\|\epsilon\|_\infty$ with (63) for different numbers of bits from 1 to 8, which is plotted in Fig. 11. This figure clarifies the relationship between the error norm and the number of bits assigned to the quantizer.

The norm of our quantizer decays exponentially with a rate faster than $1/2$. We may conclude that our quantizer is more efficient in the number of bits than the conventional quantizer without the error feedback filter, since its decay rate of the error norm with respect to the number of bits is given by $1/2$.

Finally, we evaluate our designed FIR filter in signed fixed-point binary number formats with a specified word length and fraction lengths. We fix the word length to be 16 and change the fraction length from 4 to 10. We compute the normalized value by dividing the value obtained with fixed-point numbers by the corresponding value with 64-bit floating-point numbers.

Fig. 12 illustrates the normalized values for $\|H[z]R[z]\|$, $\|R[z]-1\|$, and $c\|H[z]R[z]\|+\|R[z]-1\|$. In this example, we require 7 bits for the fraction of each number to achieve a comparable performance with the filter realized by 64-bit floating-point numbers. Since $R[z]$ is designed based on the system $H[z]$, the required precision is dependent on system $H[z]$. It should also be remarked that some of the existing techniques developed for the physical implementation of $\Delta\Sigma$ modulators may be applied to our feedback quantizer, since our feedback quantizer has the same structure with some of $\Delta\Sigma$ modulators.

VI. Conclusion

We have studied a feedback quantizer composed of a static quantizer and an error feedback filter. We have investigated the necessary number of bits required for quantization to attain the requirement on the system output, while keeping no-overloading in the quantizer. The number of bits assigned to the quantizer can be obtained by designing the error feedback filter that minimizes a constraint for no-overloading. The design of FIR filters has been formulated as linear programming by directly evaluating the l_∞ norm, whereas the design of IIR filters has been as a convex optimization problem by using upper bounds on the l_∞ norm. In our design example, if one assigns the same order for filters, the optimized FIR filter exhibits a better performance than the designed IIR filter. The efficiency of the designed quantizer has been demonstrated by simulation.

Appendix

Let us convert the non-convex BMI (49) to an LMI by using the change of variables proposed independently in [27] and [28].

Let the order of $R[z]$ be equal to the order n of the system $H[z]$. The set of $n \times n$ positive definite matrices is denoted as $PD(n)$. We define the following matrices $\{P_f, S_f, W_f, W_g, L\}$, where $P_f \in PD(n)$, $S_f \in PD(n)$, $W_f \in \mathbb{R}^{1\times n}$, $W_g \in \mathbb{R}^{n\times 1}$, $L \in \mathbb{R}^{n\times n}$, with P_f and P_g. Let us also define matrices from $\{P_f, S_f, W_f, W_g, L\}$ as

$$\mathcal{P}^{-1} = \begin{bmatrix} P_f & S_f \\ S_f & S_f \end{bmatrix} \tag{73}$$

$$U = \begin{bmatrix} P_f & I_n \\ S_f & \mathbf{0} \end{bmatrix} \tag{74}$$

$$P_g = (P_f - S_f)^{-1} \tag{75}$$

and the matrices $\{M_{\mathcal{A}}, M_{\mathcal{B}}, M_{\mathcal{C}}, M_{\mathcal{P}}\}$ as

$$M_{\mathcal{A}} = \begin{bmatrix} A_h P_f + B_h W_f & A_h \\ L & P_g A_h \end{bmatrix} \tag{76}$$

$$M_{\mathcal{B}} = \begin{bmatrix} B_h \\ W_g \end{bmatrix} \tag{77}$$

$$M_{\mathcal{C}} = \begin{bmatrix} C_h P_f + D_h W_f & C_h \end{bmatrix} \tag{78}$$

$$M_{\mathcal{P}} = \begin{bmatrix} P_f & I_n \\ I_n & P_g \end{bmatrix} \tag{79}$$

Direct computations show that if the matrices $\{A_r, B_r, C_r\}$ are

$$A_r = [B_h W_f - P_g^{-1}(L - P_g A_h P_f)]S_f^{-1} \tag{80}$$

$$B_r = B_h - P_g^{-1} W_g \tag{81}$$

$$C_r = W_f S_f^{-1} \tag{82}$$

then $\{\mathcal{A}, \mathcal{B}, \mathcal{C}\}$ satisfy

$$M_{\mathcal{A}} = U^T \mathcal{P} \mathcal{A} U \tag{83}$$

$$M_{\mathcal{B}} = U^T \mathcal{P} \mathcal{B} \tag{84}$$

$$M_{\mathcal{C}} = \mathcal{C} U \tag{85}$$

$$M_{\mathcal{P}} = U^T \mathcal{P} U. \tag{86}$$

Theorem 1 [28] proves that the BMI for the original variables $\{\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{P}\}$ is equivalent to the LMI for the new variables $\{M_{\mathcal{A}}, M_{\mathcal{B}}, M_{\mathcal{C}}, M_{\mathcal{P}}\}$ by replacing $\{\mathcal{P}\mathcal{A}, \mathcal{P}\mathcal{B}, \mathcal{C}, \mathcal{P}\}$ with $\{M_{\mathcal{A}}, M_{\mathcal{B}}, M_{\mathcal{C}}, M_{\mathcal{P}}\}$. Thus, (49) and (60) are converted into

$$\begin{bmatrix} (1-\alpha)M_{\mathcal{P}} & \mathbf{0} & M_{\mathcal{A}}^T \\ \mathbf{0} & \alpha & M_{\mathcal{B}}^T \\ M_{\mathcal{A}} & M_{\mathcal{B}} & M_{\mathcal{P}} \end{bmatrix} \succeq \mathbf{0} \tag{87}$$

$$\begin{bmatrix} M_{\mathcal{P}} & M_{\mathcal{C}}^T \\ M_{\mathcal{C}} & \mu_\epsilon \end{bmatrix} \succeq \mathbf{0}. \tag{88}$$

On the other hand, we have

$$\tilde{\mathcal{C}}U = \begin{bmatrix} C_r S_f & \mathbf{0} \end{bmatrix} = \begin{bmatrix} W_f & \mathbf{0} \end{bmatrix} := M_{\tilde{\mathcal{C}}}. \tag{89}$$

Premultiplying (61) by $\mathrm{diag}(U^T, I)$ and postmultiplying (61) by $\mathrm{diag}(U, I)$ results in

$$\begin{bmatrix} M_{\mathcal{P}} & M_{\tilde{\mathcal{C}}}^T \\ M_{\tilde{\mathcal{C}}} & \mu_\eta \end{bmatrix} \succeq \mathbf{0}. \tag{90}$$

Therefore the minimization problem

$$\min_{P_f, P_g, W_f, W_g, L, \mu_\epsilon} \ \mu_\epsilon \tag{91}$$

subject to (87), (88), and (90), gives the minimum of the minimization problem (59) for a given α.

For a fixed α, the minimization problem is a semidefinite program, which can be numerically solved by existing optimization packages, e.g., CVX [29], a package for specifying and solving convex programs. then, the minimum is given by a line search for $\alpha \in (0, \rho^2(A_h))$.

REFERENCES

[1] T. Thong and B. Liu, "Error spectrum shaping in narrow-band recursive filters," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 2, pp. 200–203, Apr. 1977.

[2] W. Higgins and D. Munson, "Noise reduction strategies for digital filters: Error spectrum shaping versus the optimal linear state-space formulation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-30, no. 6, pp. 963–973, Dec. 1982.

[3] T. I. Laakso and I. O. Hartimo, "Noise reduction in recursive digital filters using high-order error feedback," *IEEE Trans. Signal Process.*, vol. 40, no. 5, pp. 1096–1107, May 1992.

[4] R. Schreier and G. C. Temes, *Understanding Delta-Sigma Data Converters*. Hoboken, NJ, USA: Wiley, 2004.

[5] S. Azuma and T. Sugie, "Optimal dynamic quantizers for discrete-valued input control," *Automatica*, vol. 44, no. 2, pp. 396–406, Feb. 2008.

[6] S. I. Azuma and T. Sugie, "Synthesis of optimal dynamic quantizers for discrete-valued input control," *IEEE Trans. Autom. Control*, vol. 53, no. 9, pp. 2064–2075, Oct. 2008.

[7] K. Sawada and S. Shin, "Dynamic quantizer synthesis based on invariant set analysis for SISO systems with discrete-valued input," in *Proc. 19th Int. Symp. Math. Theory Netw. Syst.*, 2010, pp. 1385–1390.

[8] S. Ohno and M. R. Tariq, "Optimization of noise shaping filter for quantizer with error feedback," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 64, no. 4, pp. 918–930, Apr. 2017.

[9] 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. 2828–2839, Jun. 2012.

[10] X. Li, C. B. Yu, and H. Gao, "Design of delta–sigma modulators via generalized Kalman–Yakubovich–Popov lemma," *Automatica*, vol. 50, no. 10, pp. 2700–2708, 2014.

[11] S. Callegari and F. Bizzarri, "Output filter aware optimization of the noise shaping properties of $\Delta\Sigma$ modulators via semi-definite programming," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 60, no. 9, pp. 2352–2365, Sep. 2013.

[12] S. Callegari and F. Bizzarri, "Noise weighting in the design of $\Delta\Sigma$ modulators (with a psychoacoustic coder as an example)," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 60, no. 11, pp. 756–760, Nov. 2013.

[13] 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, no. 8, pp. 3871–3890, Aug. 2008.

[14] W. S. Wong and R. W. Brockett, "Systems with finite communication bandwidth constraints. II. Stabilization with limited information feedback," *IEEE Trans. Autom. Control*, vol. 44, no. 5, pp. 1049–1053, May 1999.

[15] S. Tatikonda and S. Mitter, "Control under communication constraints," *IEEE Trans. Autom. Control*, vol. 49, no. 7, pp. 1056–1068, Jul. 2004.

[16] T. Tanaka, P. M. Esfahani, and S. K. Mitter, "LQG control with minimum directed information: Semidefinite programming approach," *IEEE Trans. Autom. Control*, to be published.

[17] V. Chellaboina, M. Haddad, D. Bernstein, and D. Wilson, "Induced convolution operator norms for discrete-time linear systems," in *Proc. 38th IEEE Conf. Decision Control*, vol. 1. Dec. 1999, pp. 487–492.

[18] T. Hinamoto, S. Karino, N. Kuroda, and T. Kuma, "Error spectrum shaping in two-dimensional recursive digital filters," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 46, no. 10, pp. 1203–1215, Oct. 1999.

[19] D. Markert, X. Yu, H. Heimpel, and G. Fischer, "An all-digital, single-bit RF transmitter for massive MIMO," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 64, no. 3, pp. 696–704, Mar. 2017.

[20] J. M. de la Rosa, "Sigma-delta modulators: Tutorial overview, design guide, and state-of-the-art survey," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 58, no. 1, pp. 1–21, Jan. 2011.

[21] S. Ohno, T. Shiraki, M. R. Tariq, and M. Nagahara, "Mean squared error analysis of quantizers with error feedback," *IEEE Trans. Signal Process.*, vol. 65, no. 22, pp. 5970–5981, Nov. 2017.

[22] 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, Mar. 1990.

[23] H. A. Eiselt and C.-L. Sandblom, *Linear Programming and its Applications*, 1st ed. Springer, 2007.

[24] H. Shingin and Y. Ohta, "Optimal invariant sets for discrete-time systems: Approximation of reachable sets for bounded inputs," in *Proc. 10th IFAC/IFORS/IMACS/IFIP Symp. Large Scale Syst., Theory Appl. (LSS)*, 2004, pp. 401–406.

[25] N. J. Ploplys, P. A. Kawka, and A. G. Alleyne, "Closed-loop control over wireless networks," *IEEE Control Syst.*, vol. 24, no. 3, pp. 58–71, Jun. 2004.

[26] G. F. Franklin, J. D. Powell, and M. L. Workman, *Digital Control of Dynamic Systems*, 3rd ed. Menlo Park, CA, USA: Addison-Wesley, 1998.

[27] C. Scherer, P. Gahinet, and M. Chilali, "Multiobjective output-feedback control via LMI optimization," *IEEE Trans. Autom. Control*, vol. 42, no. 7, pp. 896–911, Jul. 1997.

[28] I. Masubuchi, A. Ohara, and N. Suda, "LMI-based controller synthesis: A unified formulation and solution," *Int. J. Robust Nonlinear Control*, vol. 8, no. 8, pp. 669–686, Jul. 1998.

[29] M. Grant and S. Boyd. *CVX: MATLAB Software for Disciplined Convex Programming, Version 2.0 Beta*. Sep. 2012. [Online]. Available: http://cvxr.com/cvx

Shuichi Ohno (M'95–SM'11) received the B.E., M.E., and Dr.Eng. degrees in applied mathematics and physics from Kyoto University, in 1990, 1992, and 1995, respectively. From 1995 to 1999, he was a Research Associate with the Department of Mathematics and Computer Science, Shimane University, Shimane, Japan, where he became an Assistant Professor. He spent 14 months in 2000 and 2001 with the University of Minnesota as a Visiting Researcher. Since 2010, he has been an Associate Professor with the Department of System Cybernetics, Hiroshima University. His current interests are in the areas of signal processing in control and communications, and adaptive signal processing.

Dr. Ohno is a member of SICE, IEICE, and ISCIE. He served as an Associated Editor of the IEEE SIGNAL PROCESSING LETTERS from 2001 to 2003.

Yuma Ishihara received the bachelor's degree in electrical engineering from Hiroshima University, Hiroshima, Japan, in 2016.

Masaaki Nagahara (S'00–M'03–SM'14) received the bachelor's degree in engineering from Kobe University in 1998, and the master's degree and the Doctoral degree in informatics from Kyoto University in 2000 and 2003, respectively.

He is currently a Full Professor with the Institute of Environmental Science and Technology, The University of Kitakyushu. He has been a Visiting Professor with IIT Bombay since 2017. His research interests include control theory, machine learning, and sparse modeling.

Dr. Nagahara is a member of SICE, ISCIE, IEICE, and JSAI. He received the Young Authors Award in 1999 and Best Paper Award in 2012 from SICE, the Transition to Practice Award from the IEEE Control Systems Society in 2012, the Best Tutorial Paper Award from the IEICE Communications Society in 2014, and the Best Book Authors Award from SICE in 2016.