Mean Squared Error Analysis of Quantizers With Error Feedback

Shuichi Ohno, *Senior Member, IEEE*, Teruyuki Shiraki, M. Rizwan Tariq, *Student Member, IEEE*, and Masaaki Nagahara, *Senior Member, IEEE*

***Abstract*—Quantization is a fundamental process in digital signal processing. $\Delta\Sigma$ modulators are often utilized for quantization, which can be easily implemented with static uniform quantizers and error feedback filters. In this paper, we analyze the mean squared quantization error of the quantizer with error feedback including the $\Delta\Sigma$ modulators. First, we study the quantizer with an ideal optimal error feedback filter that minimizes the mean squared error (MSE) of quantization. We show that the amplitude response of the optimal error feedback filter can be parameterized by one parameter. This parameterization enables us to find the optimal error feedback filter numerically. Second, the relationship between the number of bits used for the quantizer and the achievable MSE is clarified by using the optimal error feedback filter. This makes it possible to investigate the efficiency of the quantizer with the optimal error feedback filter in terms of MSE. Then, ideal optimal error feedback filters are approximated by practically implementable filters using the Yule-Walker method and the linear matrix inequality-based method. Numerical examples are provided for demonstrating our analysis and synthesis.**

***Index Terms*—Quantization, $\Delta\Sigma$ modulator, error feedback, MSE.**

I. Introduction

QUANTIZATION is a fundamental process in digital signal processing, wherein, a large set of input values are mapped onto a smaller set of output values. The simplest type of quantizer is the uniform quantizer that has fixed-length code words, i.e., a fixed number of bits per sample. However, the uniform quantizer is not efficient because it does not consider the statistics of the input and/or the information about the system connected to the quantizer. Additional information regarding the input and/or the connected system can be exploited to obtain good quantizers. Under the assumption that the quantization error is a white uniformly distributed random sequence, the Lloyd-Max quantizer is optimal among the quantizers having fixed-length code words in the sense that it minimizes the distortion of the quantization error [1, Ch.9]. However, the probability density function of the input to the quantizer, that is often unavailable in practice, is required for constructing the Lloyd-Max quantizer.

Quantization with error feedback is more efficient than the conventional uniform quantization. It includes a uniform quantizer and a feedback filter, where the filtered error of the uniform quantizer is fed back to it for mitigating the error introduced by quantization. Quantization with error feedback is used for reducing the effect of the quantized coefficients in fixed-point digital filters [2], [3]. Finite impulse response (FIR) error feedback filters have been proposed for recursive digital filters composed of cascaded second order sections in [4].

Various designs for the feedback filter have been proposed. Based on the generalized Kalman-Yakubovich-Popov (GKYP) lemma, an FIR error feedback filter has been designed to minimize the worst case gain in the signal passband using convex optimization [5], whereas an infinite impulse response (IIR) filter using an iterative algorithm [6]. Under the whiteness assumption for the error of the uniform quantizer, an optimal FIR feedback filter that minimizes the variance of the error owing to quantization has been proposed in [7]. On the other hand, IIR error feedback filters have been presented in [8] for minimizing the norm of the error in the signal of interest, introduced by the quantization.

Quantization with error feedback is also adopted in $\Delta\Sigma$ or $\Sigma\Delta$ modulators that are often utilized to convert real values into fixed-point numbers and vise versa [9]. $\Delta\Sigma$ modulators are widely used for several applications, e.g., audio signal processing [10], RF transmitter architectures [11], compressive sensing [12], and independent source separation [13].

It is known that when a $\Delta\Sigma$ modulator is used to quantize an analog signal into a digital signal, oversampling can effectively reduce the error introduced by quantization. However, oversampling increases the number of bits per time, if the same number of bits are assigned to each output of the quantizer. Whether oversampling is effective when the number of bits per time is fixed, continues to remain unclear. To answer this, we need to show the relationship between the achievable mean squared quantization error and the number of bits used for the quantization.

It has been found in [14] that for bandlimited signals, the variance of the distortion, i.e., the mean squared error (MSE) of a simple single-loop one-bit $\Delta\Sigma$ modulator decays at a rate of $O(\lambda^{-4})$, where λ is the oversampling ratio. In [15], it is proven

Manuscript received January 12, 2017; revised May 27, 2017 and July 31, 2017; accepted August 9, 2017. Date of publication August 29, 2017; date of current version September 19, 2017.The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Wee Peng Tay. This work was supported by the JSPS KAKENHI under Grant JP16K06356. *(Corresponding author: Shuichi Ohno.)*

S. Ohno, T. Shiraki, and M. R. Tariq are with the Hiroshima University, Higashi-Hiroshima 739-8527, Japan (e-mail: o.shuichi@ieee.org; teru.s.1991@gmail.com; rizwantariq16@yahoo.com).

M. Nagahara is with the Institute of Environmental Science and Technology, University of Kitakyushu, Kitakyushu 808-0135, Japan (e-mail: nagahara@ieee.org).

Digital Object Identifier 10.1109/TSP.2017.2745450

that for bandlimited bounded signals the squared maximum absolute value, i.e., the squared l_∞ norm of the distortion of a one-bit $\Delta\Sigma$ modulator can decrease at a rate of $O(\lambda^{-4})$; further, a family of one-bit $\Delta\Sigma$ modulators that attain this rate has been provided. In [16], optimal filters in this family are designed to improve the decay rate demonstrating that an exponential rate of $O(2^{-r\lambda})$ for $r \approx 0.102$ is achieved by the designed filter. On the other hand, for bandlimited stationary signals, the MSE of the optimal one-bit $\Delta\Sigma$ modulator that minimizes the MSE under a constraint on the variance of the input to the uniform quantizer, decreases exponentially at a rate of $O(2^{-r\lambda})$ for $r \approx 0.807$ [17]. This improvement becomes possible by exploiting the knowledge on the power spectral density function of the input and by using an additional pre-filter and post-filer with an infinite order. In this paper, we consider a more practical situation, wherein the spectrum of the input is unavailable, and investigate the achievable MSE of the conventional $\Delta\Sigma$ modulators without pre/post-filters.

The input to the static quantizer in a quantizer with an error feedback exhibits a larger amplitude than the input to a conventional uniform quantizer without an error feedback. To enable fair comparisons between quantizers with different input amplitudes, we utilize static uniform quantizers having an identical signal-to-quantization noise ratio (SQNR). Then, we study the variance of the error at the output of the system connected to the quantizer.

After formulating our problem as an optimization problem, we show that the amplitude response of the optimal error feedback filter that minimizes the MSE at the output can be parameterized by one parameter. Then, the optimal error feedback filter can be determined numerically by minimizing the MSE with respect to the parameter. Using optimal error feedback filter, we present the relationship between the achievable MSE and the number of bits used for the quantization. These analytical results are reported in the conference version [18] of this paper, which do not have any proofs of the results. Here, in addition to detailed proofs, further analysis and new synthesis of error feedback quantizers are provided. Unlike [17], we do not require the power spectral density function of the input, that is not always available in practice, nor the additional pre-filter and post-filer, which are dependent on the system connected to the quantizer.

Our MSE analysis of the quantizer with the optimal error feedback filter guarantees that if a fixed number of bits are assigned for each quantized signal, the optimal quantizer with an error feedback always outperforms the uniform quantizer. It also demonstrates the effect of oversampling on the MSE. If the number of bits per sample is fixed, oversampling improves the MSE. On the other hand, if the number of bits per time, i.e., bit-rate is fixed, oversampling degrades the MSE. Finally, we present two approximations for ideal optimal filters using the Yule-Walker method [19] and the linear matrix inequality (LMI)-based method for obtaining implementable error feedback filters. Our LMI-based method enables to design an error feedback filter having an identical order with the system connected to the quantizer. Numerical examples are provided to demonstrate our analysis and synthesis.

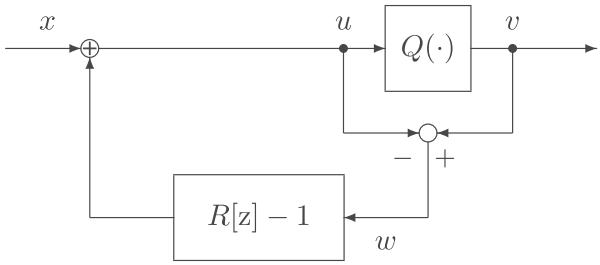


Fig. 1. Quantizer with an error feedback filter.

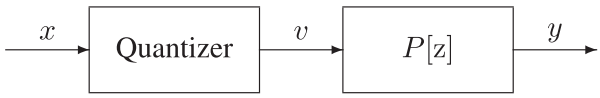


Fig. 2. Quantizer and system.

II. Quantizer with Error Feedback

Fig. 1 depicts our quantizer with the error feedback, where x is the input signal to the quantizer with the error feedback, v is its output signal, and $Q(\cdot)$ denotes a conventional static uniform quantizer. All the signals are assumed to be of discrete-time. We denote the z transform of a discrete-time signal, $f = \{f_k\}_{k=0}^{\infty}$, as $F[\mathrm{z}] = \sum_{k=0}^{\infty} f_k \mathrm{z}^{-k}$.

In Fig. 1, the signal $w = v - u$ is the quantization error signal of the static uniform quantizer that is filtered by $R[\mathrm{z}] - 1$ and fed back to x. The first coefficient of the impulse response of $R[\mathrm{z}]$ is assumed to be one, implying that $R[\mathrm{z}] - 1$ is strictly causal and hence, practically implementable. The minus one in $R[\mathrm{z}] - 1$ is only for the simplicity of presentation.

Quantization with error feedback has a simple structure that can be implemented at a relatively low cost. The linearized model of the $\Delta\Sigma$ modulator can be expressed by a quantizer with an error feedback filter.

The input signal u to the uniform quantizer is expressed in the z domain as $U[\mathrm{z}] = X[\mathrm{z}] + (R[\mathrm{z}] - 1)W[\mathrm{z}]$. On the other hand, the z transform of the output of the quantizer can be expressed as

$$V[\mathrm{z}] = X[\mathrm{z}] + R[\mathrm{z}]W[\mathrm{z}]. \tag{1}$$

As $R[\mathrm{z}]$ shapes the spectrum of the noise w, it is called a *noise shaping filter* or a *noise transfer function*.

We assume that the output of the quantizer passes through the system $P[\mathrm{z}]$ as depicted in Fig. 2. The z transform of the output y of $P[\mathrm{z}]$ can be expressed as $Y[\mathrm{z}] = P[\mathrm{z}]X[\mathrm{z}] + E[\mathrm{z}]$, where $E[\mathrm{z}]$ is the z transform of the error at the output introduced by the quantization and is given by

$$E[\mathrm{z}] = P[\mathrm{z}]R[\mathrm{z}]W[\mathrm{z}]. \tag{2}$$

The purpose of this paper is to clarify the achievable mean squared error (MSE) of the quantizer with an optimal error feedback, when the input spectrum cannot be used.

III. Optimal Error Feedback Filter for Quantization

First, let us review static uniform quantizers. Although most of our analysis holds true for the other types of static quantizers

under the same conditions, we consider the mid-rise quantizer as an example.

The mid-rise quantizer can be described by two parameters, the quantization interval $d(>0)$ and the saturation (or equivalently, overloading) level $L(>0)$. Its output for a scalar input ξ is expressed as

$$Q(\xi) = \begin{cases} \left(i+\frac{1}{2}\right)d, & \xi \in [id,(i+1)d) \\ & \text{for an integer } i \text{ and } |\xi| \le L+\frac{d}{2} \\ L, & \xi > L+\frac{d}{2} \\ -L, & \xi < -L-\frac{d}{2} \end{cases}. \quad (3)$$

The overload is the saturation owing to the fixed number of bits representing the binary-values. In the mid-rise quantizer, an overload occurs if $|\xi| > L + \frac{d}{2}$.

If we assign b bits to the mid-rise quantizer, where b is a positive integer, the number of quantization levels is 2^b that is related to the saturation level L of the input to the mid-rise quantizer and the quantization interval d through

$$2L = (2^b - 1)d. \quad (4)$$

For our analysis, we assume that a sufficient number of bits are assigned to the output of the uniform quantizer so that:

Assumption 1: There is no overloading in the uniform quantizer.

Suppose that the input x is bounded such that $\max_k |x_x| \le L_x$ for $L_x > 0$. Then, we have $\max_k |u_x| = \max_k |x_k + \sum_{l=1}^{\infty} r_l w_{k-l}| \le L_x + \sum_{l=1}^{\infty} |r_l| d/2$. If we set $L + d/2 \ge L_x + \sum_{l=1}^{\infty} |r_l| d/2$, then there is no overloading in the uniform quantizer. However, it should be noted that this setting is often too conservative and a smaller value for L may attain the no-overloading.

The input x to our quantizer is assumed to be a wide-sense stationary process with a zero mean and a variance σ_x^2. We also assume that the quantization error signal of the static uniform quantizer is a white noise and is uncorrelated with the input x [20].

Assumption 2: The quantization error signal w of the uniform quantizer is a white random signal with a zero mean and a variance σ_w^2 and uncorrelated with the input of the uniform quantizer.

For the uniform quantizer, it is known that Assumption 2 approximately holds true if there is no overloading, the quantization interval d is sufficiently small, and a sufficiently large number of bits is assigned [21]. On the other hand, Assumption 2 is not always satisfied for the quantization error of the uniform quantizer in the error feedback quantizer due to the feedback signal and the oversampling [22]. The feedback signal complicates the theoretical analysis. Except for some specific inputs and simple error feedback quantizers, there are few theoretical results on the property of the quantization error [23]. However, Assumption 2 is usually adopted [9], since it is still a good approximation for error feedback quantizers having sufficiently small quantization intervals and error feedback filters with long impulse responses. For example, it has been demonstrated in [17] that empirical results are well matched to the theoretical results under Assumption 2. Moreover, if a white thermal noise of the analog circuit is present at the input of the uniform quantizer or a dither, that adds a white noise, is used, Assumption 2 is asymptotically met under Assumption 1 [24], [25].

Under Assumptions 1 and 2, the signal-to-quantization-noise ratio (SQNR) of the static uniform quantizer is defined as

$$\gamma = \frac{\sigma_u^2}{\sigma_w^2} \quad (5)$$

where σ_u^2 is the variance of the input u.

In an error feedback quantizer, the range of the input to the static uniform quantizer depends on the feedback signal. To deal with error feedback quantizers having different ranges, let us fix SQNR of static uniform quantizers. The constraint on SQNR enables us to analyze and to fairly compare quantizers with different feedback filters. Now, let us evaluate error feedback quantizers with static uniform quantizers having an identical SQNR.

Let us denote the L_2 norm of a filter $H[\mathrm{z}]$ as $||H[\mathrm{z}]||$ that is defined as

$$||H[\mathrm{z}]|| = \left(\frac{1}{2\pi}\int_{-\pi}^{\pi} H^*[e^{j\omega}]H[e^{j\omega}]d\omega\right)^{\frac{1}{2}} \quad (6)$$

where c^* is the complex conjugate of c.

From Assumption 2, the variance of the input u to the uniform quantizer is expressed as

$$\sigma_u^2 = \sigma_x^2 + ||R[\mathrm{z}] - 1||^2 \sigma_w^2. \quad (7)$$

Then, under Assumption 2, the variance of the quantization error of the uniform quantizer is expressed from (5) as

$$\sigma_w^2 = \frac{\sigma_x^2}{\gamma - ||R[\mathrm{z}] - 1||^2} \quad (8)$$

that requires

$$\gamma - ||R[\mathrm{z}] - 1||^2 > 0. \quad (9)$$

This implies that the energy of the feedback signal has to be limited. As the first entry of the impulse response of $R[\mathrm{z}]$ is unity, we have $||R[\mathrm{z}] - 1||^2 + 1 = ||R[\mathrm{z}]||^2$ and then

$$\sigma_w^2 = \frac{\sigma_x^2}{\nu - ||R[\mathrm{z}]||^2} \quad (10)$$

where

$$\nu = \gamma + 1. \quad (11)$$

The variance of the quantization error at the output of the system is obtained from (2) by

$$||P[\mathrm{z}]R[\mathrm{z}]||^2 \sigma_w^2. \quad (12)$$

Substituting (10) in (12) results in

$$||P[\mathrm{z}]R[\mathrm{z}]||^2 \sigma_w^2 = \frac{||P[\mathrm{z}]R[\mathrm{z}]||^2}{\nu - ||R[\mathrm{z}]||^2}\sigma_x^2. \quad (13)$$

To observe the performance of our quantizer, we would like to obtain the optimal error feedback filter $R[\mathrm{z}]$ and the minimum of the mean squared error (MSE). For a given σ_x^2 and $P[\mathrm{z}]$, we minimize the MSE with respect to $R[\mathrm{z}]$. To stabilize the

quantizer, $R[z]$ must be stable. Then, as σ_x^2 in (13) is a constant, our problem can be formulated as the following minimization:

$$\min_{R[z]\in RH_\infty} \frac{||P[z]R[z]||^2}{\nu - ||R[z]||^2} \tag{14}$$

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

$$||R[z]||^2 < \nu \tag{15}$$

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

To enable theoretical analysis, we relax the stable proper rational function $R[z]$ to a function $r(\omega) \in L_2$ belonging to a more general class of functions that is piece-wise differentiable on $[-\pi, \pi]$, has at most a finite number of discontinuity points, and satisfies

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln r(\omega) d\omega = c_0 \tag{16}$$

for $0 \le c_0 < \infty$. We note that (16) is imposed by the stability of the original function $R[z]$.[1]

The L_2 norm of $q(\omega) \in L_2$ is defined as

$$||q(\omega)|| = \left(\frac{1}{2\pi}\int_{-\pi}^{\pi} q^*(\omega)q(\omega)d\omega\right)^{\frac{1}{2}}. \tag{17}$$

We denote the set of L_2 functions that satisfy (16) as $\mathcal{C}_0$. We also define a set $\mathcal{C}_1$ of L_2 functions as

$$\mathcal{C}_1 = \left\{r(\omega) : ||r(\omega)||^2 < \nu\right\}. \tag{18}$$

Then, we would like to determine $r(\omega) \in \mathcal{C}_0 \cap \mathcal{C}_1$ that minimizes

$$\frac{||p(\omega)r(\omega)||^2}{\nu - ||r(\omega)||^2} \tag{19}$$

where

$$p(\omega) = |P[e^{j\omega}]|. \tag{20}$$

Although we extend the class of functions, from Lemma 1 in [17], we can find a stable proper rational function $R[z]$ such that $|R[e^{j\omega}]|$ approximates $r(\omega)$ arbitrarily well in $[-\pi, \pi]$. Then, the stable proper rational function that approximates the solution for the minimization of (19) can be considered as an approximate solution for the original minimization problem.

Now, our problem is to find the optimal function that minimizes (19) such that

$$r_{opt}(\omega) = \arg\min_{r(\omega)\in\mathcal{C}_0\cap\mathcal{C}_1} \frac{||p(\omega)r(\omega)||^2}{\nu - ||r(\omega)||^2}. \tag{21}$$

For our analysis, let us introduce the notion of almost constant functions.

Definition 1: A function $\psi : [a, b] \to \mathbb{R}$ is said to be almost constant if and only if

$$\int_a^b \left|\psi(x) - \frac{1}{b-a}\int_a^b \psi(x)dx\right| \psi(x)dx = 0 \tag{22}$$

[1]Since $R[z]$ is stable, $R[z]$ is analytic outside of the unit circle, that is, $R[z^{-1}]$ is analytic on the unit circle as a function of z. Moreover, $R[z]|_{z=\infty} = 1$ implies $R[z^{-1}]|_{z=0} = 1$. Then, we can apply Jensen' formula [26] to obtain $\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln|R(e^{j\omega})|d\omega = c_0$.

The optimal solution for our problem cannot be expressed in a closed-form but can be characterized with one parameter as follows (see Appendix A for proof):

Theorem 1: Suppose that $p(\omega)$ is not almost constant. Then, for any $\gamma > 0$, the optimal function that minimizes (19) can be expressed using a parameter α as

$$r_\alpha(\omega) = \frac{\theta(\alpha)}{\sqrt{p^2(\omega) + \alpha}} \tag{23}$$

where

$$\theta(\alpha) = \exp\left(\frac{1}{4\pi}\int_{-\pi}^{\pi} \ln(p^2(\omega) + \alpha)d\omega\right). \tag{24}$$

If $p(\omega)$ is almost constant, then the optimal function is almost constant.

If we can utilize the pre-filter and post-filter as in [17], the minimum MSE results in [17]

$$\frac{\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} |P(e^{j\omega})R(e^{j\omega})\Omega_x(e^{j\omega})|d\omega\right)^2}{\nu - ||R[z]||^2} \tag{25}$$

where $\Omega_x(e^{j\omega})$ is the spectrum of the input x. Our MSE (13) has the L_2 norm in its numerator, whereas (25) has the squared L_1 norm in its numerator. This leads to similar yet different results by similar derivations. For example, if $\Omega_x(e^{j\omega}) = 1$, the optimal filter is expressed as [17][Theorem 1]

$$\frac{\tilde{\theta}(\alpha)}{\sqrt{p^2(\omega) + \alpha} + p(\omega)}$$

with

$$\tilde{\theta}(\alpha) = \exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln(\sqrt{p^2(\omega) + \alpha} + p(\omega))d\omega\right). \tag{26}$$

It has been shown in [27] that the optimal error feedback filter $R[z]$ that minimizes $||P[z]R[z]||^2\sigma_w^2$ without any constraint on the input to the static quantizer has an amplitude response proportional to $1/p(\omega)$. Theorem 1 reveals that the optimal error feedback filter under constraint (5) has a similar amplitude response as the optimal error feedback filter. More importantly, Theorem 1 assures that a quantizer with an error feedback outperforms a static uniform quantizer, except for the trivial case where $P[z]$ is almost constant.

To proceed further, we express our objective function by the parameter α as

$$\Phi(\alpha) = \frac{N(\alpha)}{\nu - C(\alpha)} \tag{27}$$

where

$$N(\alpha) = \frac{1}{2\pi}\int_{-\pi}^{\pi} p^2(\omega)r_\alpha^2(\omega)d\omega \tag{28}$$

and

$$C(\alpha) = ||r_\alpha||^2 = \frac{\theta^2(\alpha)}{2\pi}\int_{-\pi}^{\pi} \frac{1}{p^2(\omega) + \alpha}d\omega. \tag{29}$$

We have to determine the global minimizer of $\Phi(\alpha)$, i.e.,

$$\alpha_{opt} = \arg\min_\alpha \Phi(\alpha). \tag{30}$$

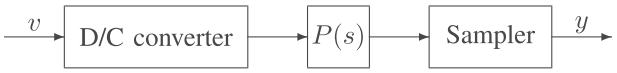


Fig. 3. D/C converter and sampling.

In Appendix B, we show that minimizing $\Phi(\alpha)$ with respect to α leads to the following theorem, enabling us to compute the minimizer numerically.

Theorem 2: For any $\gamma > 0$, the optimal α denoted by α_{opt} that minimizes $\Phi(\alpha)$ satisfies $\alpha_{opt} > 0$ and

$$\nu = \frac{\theta^2(\alpha_{opt})}{\alpha_{opt}}. \tag{31}$$

It can be easily discerned that

$$\frac{d}{d\alpha}\left(\frac{\theta^2(\alpha)}{\alpha}\right) = -\frac{\theta^2(\alpha)}{\alpha^2}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{p^2(\omega)}{p^2(\omega)+\alpha}d\omega\right) < 0. \tag{32}$$

On the other hand, we have from the definition (24) that $\theta^2(\alpha)/\alpha = \infty$ at $\alpha = 0$ and $\theta^2(\alpha)/\alpha \to 1$ as $\alpha \to \infty$. Since $\theta^2(\alpha)/\alpha$ is a monotonically decreasing function in α and $\nu > 1$, the solution of (31) always exists and unique. This proves that the optimization problem (30) can be solved uniquely for $\alpha > 0$.

In practice, for a given γ, we can obtain α_{opt} that satisfies (31) numerically by e.g., the bisection algorithm. Once the optimal α_{opt} is computed, the optimal function is given by

$$r_{opt}(\omega) = \frac{\theta(\alpha_{opt})}{\sqrt{p^2(\omega)+\alpha_{opt}}}. \tag{33}$$

IV. MSE Analysis of Optimal Error Feedback Quantizers

Let us suppose that the output of the quantizer is converted by a discrete-time-to-continuous-time (D/C) converter into a continuous-time signal and the continuous-time signal passes through the continuous-time system $P(s)$ as shown in Fig. 3. To evaluate MSE due to the quantization, we sample the continuous-time output signal of $P(s)$ to obtain a discrete-time signal y. It is noted that $P[z]$ in Fig. 2 corresponds to the discrete-time equivalent system from v to y in Fig. 3.

Based on the results of the previous section, we reveal the relationship between the sampling period and the achievable MSE of the optimal error feedback quantizer.

We assume that the continuous-time system $P(s)$ is bandlimited as follows:

Assumption 3: The continuous-time system $P(s)$, is bandlimited in $[-\pi/T_s, \pi/T_s]$ and $1/T_s$ is its Nyquist frequency.

Under Assumption 3, it suffices to sample the output of the continuous-time system $P(s)$ at the Nyquist rate to reconstruct the continuous-time output from its sampled discretized output. However, the oversampling is often adopted where error feedback quantizers are utilized.

Sampling with a sampling period T_s/λ when λ is a positive integer and $\lambda > 1$ is known as oversampling. The integer λ is called the *oversampling ratio* and is the sampling frequency divided by the Nyquist frequency. Then under Assumption 3, the sampled system with a sampling period T_s/λ satisfies

$$P_\lambda[e^{j\omega}] = P\left(\frac{\lambda\omega}{T_s}\right) \quad \text{for} \quad |\omega| \le \omega_c. \tag{34}$$

To obtain the relationship between the oversampling ratio and the achievable MSE of the optimal error feedback quantizer, we define

$$p_\lambda(\omega) = \begin{cases} p(\lambda\omega) & |\omega| \le \omega_c \\ 0 & \omega_c < |\omega| \le \pi \end{cases} \tag{35}$$

and consider the following minimization problem.

$$\min_{r(\omega)\in\mathcal{C}_0\cap\mathcal{C}_1} \frac{||p_\lambda(\omega)r(\omega)||^2}{\nu - ||r(\omega)||^2}. \tag{36}$$

This gives the minimum MSE, or equivalently, the distortion of the optimal quantizer.

Let us denote the minimum of (36) as $D(\nu,\lambda)$ that is a function in ν and λ. To designate the dependency of α_{opt} on ν and λ, we also denote α_{opt} as $\alpha_{opt}(\nu,\lambda)$. Substituting (33) in (36) and using (31) and (74), we find

$$D(\nu,\lambda) = \alpha_{opt}(\nu,\lambda). \tag{37}$$

Using (37), we prove in Appendix C that:

Theorem 3: Let the oversampling rates be λ and $\nu = \gamma + 1$, where γ is defined as (5). The MSE of the optimal quantizer with an error feedback is a function of ν and λ that satisfies

$$D(\nu,\lambda) = D(\nu^\lambda, 1). \tag{38}$$

Theorem 3 enables us to compare error feedback quantizers for different sampling rate.

Let b be the number of bits per sample in the error feedback quantizer with oversampling rate λ. We assume that the uniform quantizer has $N = 2^b$ quantization levels and an interval of d. The loading factor is defined as $L_f = L/\sigma_u = Nd/(2\sigma_u)$ [28] and is the ratio between L and the standard deviation of the input to the uniform quantizer. The loading factor regulates the frequency of the overloading. For example, if the input to the uniform quantizer is Gaussian, then the probability of the input exceeding the range is approximately 0.045, when the loading factor is four.

Let us find the number b' of bits per sample of the error feedback quantizer without oversampling whose MSE is identical with the MSE of the error feedback quantizer with oversampling rate λ.

If the quantization error of the uniform quantizer is uniformly distributed, under our Assumptions 1 and 2, γ is given by [29]

$$\gamma = \frac{3N^2}{L_f^2} = \frac{3\cdot 2^{2b}}{L_f^2}. \tag{39}$$

If we allow b and b' to take real values, there exist b and b' that satisfy

$$D\left(\frac{3\cdot 2^{2b}}{L_f^2}+1,\lambda\right) = D\left(\frac{3\cdot 2^{2b'}}{L_f^2}+1,1\right). \tag{40}$$

This means that the MSE of the optimal error feedback quantizer with b bits and oversampling ratio λ is equal to the MSE of

the optimal error feedback quantizer with b' bits without oversampling. The former consumes λb bits per unit time, whereas the latter does b' bits per unit time.

Substituting (39) in (38), we find that

$$D\left(\left[\frac{3\cdot 2^{2b}}{L_f^2}+1\right]^{\lambda},1\right)=D\left(\frac{3\cdot 2^{2b'}}{L_f^2}+1,1\right). \tag{41}$$

Thus, we have

$$\left[\frac{3\cdot 2^{2b}}{L_f^2}+1\right]^{\lambda}=\frac{3\cdot 2^{2b'}}{L_f^2}+1. \tag{42}$$

For $x>0$, $f(x)=\log_2(x+1)-\log_2(x)$ is monotonically decreasing and $f(1)=1$. There exits a positive $\delta\le 1$ such that $\log_2(x+1)=\log_2(x)+\delta$. Then, if $3\cdot 2^{2b}/L_f^2\ge 1$, we can express

$$\log_2\left(\frac{3\cdot 2^{2b}}{L_f^2}+1\right)^{\lambda}=\lambda\left(\log_2\frac{3\cdot 2^{2b}}{L_f^2}+\delta_b\right)=\lambda\left(2b+c+\delta_b\right) \tag{43}$$

$$\log_2\left(\frac{3\cdot 2^{2b'}}{L_f^2}+1\right)=\log_2\frac{3\cdot 2^{2b'}}{L_f^2}+\delta_{b'}=2b'+c+\delta_{b'} \tag{44}$$

where $0<\delta_{b'}<\delta_b<1$ and

$$c=\log\frac{3}{L_f^2}. \tag{45}$$

From (42), (43), and (44), we have

$$2b'+c+\delta_{b'}=\lambda\left(2b+c+\delta_b\right) \tag{46}$$

from which we finally obtain

$$\lambda\left[b+\frac{1}{2}(c+\delta_b)\right]-\frac{1}{2}(c+\delta_b)\le b'<\lambda\left[b+\frac{1}{2}(c+\delta_b)\right]-\frac{1}{2}c. \tag{47}$$

We note that For $L_f\ge 4$, $c=\log_2(3/L_f^2)$ is less than -1. For example, it is -1.208 for $L_f=4$. Then, for $L_f\ge 4$, $c+\delta_b<0$, as $\delta_b<1$.

If follow from (47) that for a fixed number b of bits per sample, if $3\cdot 2^{2b}/L_f^2\ge 1$, an increase in λ increases b', implying that oversampling improves the MSE. On the other hand, if we fix the number of bits per time, i.e., the bit-rate λb as a constant, we find from (47) that for $L_f\ge 4$, since $c+\delta_b<0$, an increase in the oversampling decreases the number b' of bits of the quantizer without oversampling that achieves the same MSE as the quantizer with oversampling.

As the static uniform quantizer cannot outperform the quantizer with an error feedback, we have

$$D(\nu,1)\le\frac{||P[\mathrm{z}]||^2}{\gamma}=\frac{||P[\mathrm{z}]||^2}{\nu-1}. \tag{48}$$

It follows from (38) and (48) that:

Theorem 4: The MSE of the optimal error feedback quantizer is upper bounded as

$$D(\nu,\lambda)\le\left(\frac{1}{\nu^{\lambda}-1}\right)||P[\mathrm{z}]||^2. \tag{49}$$

Theorem 4 shows that the MSE of the optimal error feedback quantizer decays at a rate of $O(\nu^{-\lambda})$. On the other hand, the decay rate of the optimal error feedback quantizer having pre/post-filters and designed with a knowledge of the input spectrum is $O(\nu^{-\lambda}/\lambda)$ [17, Theorem 6].[2] Thus, we can conclude that the scalar $1/\lambda$ in $O(\nu^{-\lambda}/\lambda)$ is the benefit we obtain from the availability of the input spectrum.

V. Synthesis of Error Feedback Filters for Quantization

We only know the amplitude response of the optimal error feedback filter from the results in Section II. In practice, we have to implement an error feedback filter with a stable rational transfer function. This necessitates the acquisition of an implementable filter approximating the optimal error feedback filter.

For approximating a given spectrum, the Yule-Walker method [19] is well-known, efficient, and is optimal in the least squares sense. If we permit the usage of a filter with a sufficiently high order, then the amplitude response of the approximated filter can be almost the same as the amplitude response of the ideal optimal filter. However, the head of the impulse response of the error feedback filter has to be unity and this is not assured by the Yule-Walker method in general. Although we may be able to modify the Yule-Walker method, we only normalize the approximated filter to have a unity head for its impulse response.

Let us develop another approximation to obtain an error feedback filter with a low order. Once the amplitude response of the optimal error feedback is obtained, we can compute its norm by using $r_{opt}(\omega)$ in (33). Then, we consider the following optimization problem:

$$\min_{R[\mathrm{z}]\in RH_\infty}||P[\mathrm{z}]R[\mathrm{z}]||^2 \tag{50}$$

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

$$||R[\mathrm{z}]||^2\le||r_{opt}(\omega)||^2. \tag{51}$$

We would like to determine the error feedback filter $R[\mathrm{z}]$ that minimizes the MSE under the norm constraint. If the amplitude response of the optimal filter for (21) can be expressed as a rational function, then we can find the error feedback filter that is close to the optimal error feedback filter. Even if this is not the case, we may expect the obtained $R[\mathrm{z}]$ to have a comparable MSE with the optimal error feedback filter.

As shown in Appendix D, the optimization problem is cast into a convex optimization problem that can be solved numerically and efficiently with a numerical solver such as the CVX

[2] In [17], the decay rate is given by $O(\nu^{-\lambda})$, as $p_\lambda(\omega)$ in (35) is scaled by $\sqrt{\lambda}$.

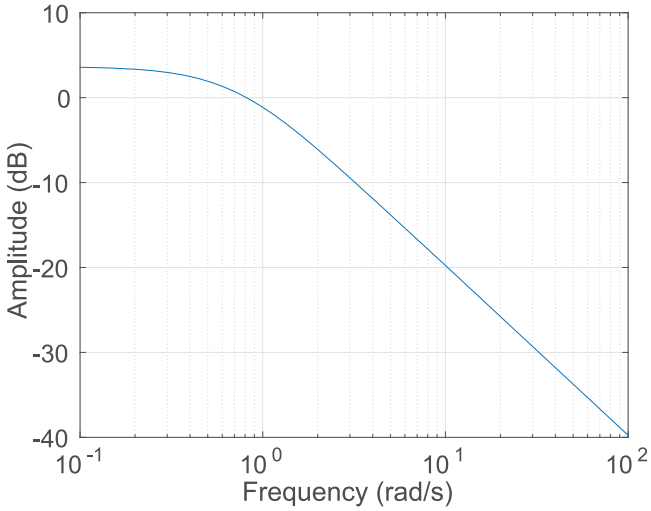


Fig. 4. Amplitude response of the continuous-time system $P(s)$.

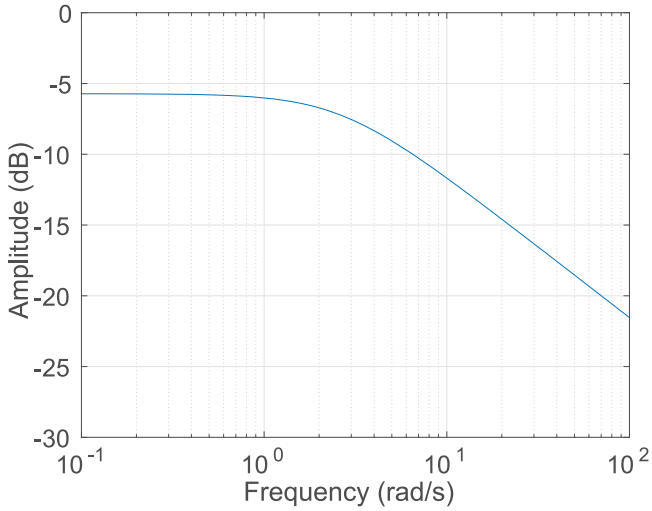


Fig. 5. Input spectrum.

[30]. In this case, the order of $R[z]$ could be set to be equal to the order of $P[z]$ because it is the minimum order that can achieve a minimum and a higher order for $R[z]$ does not reduce the minimum [31], [32].

VI. Numerical Examples

To validate our analysis and synthesis, we utilize a typical lowpass system and a lowpass input. For a lowpass system, we consider a continuous-time system of order four whose transfer function is

$$P(s) = \frac{1.029s^3 + 4.589s^2 + 7.146s + 3.882}{s^4 + 5.088s^3 + 9.789s^2 + 8.296s + 2.548}. \quad (52)$$

The amplitude response of this system is plotted in Fig. 4. We discretize this continuous-time system with a sampling period $T_s = 0.1$ to obtain the discrete-time system $P[z]$.

We model the continuous-time input signal as a stationary process with a zero mean and a spectrum given by

$$S(\omega) = c\left|\frac{1}{j\omega + 2.62}\right|^2 \quad (53)$$

where c is a constant. We set the value of c so that the sampled signal should have a unit variance. The spectrum is depicted in Fig. 5.

The loading factor is set to be four. For $b = 1, 2, \ldots, 8$, we obtain γ from (39). Then, for a given γ, we numerically find the optimal α from (24) and (31) that is the minimum MSE (c.f. (37)), replacing $p(\omega)$ by $p_\lambda(\omega)$ in (35).

For the oversampling ratio $\lambda = 1, 2, 3, 4$, Fig. 6 compares the MSEs of the optimal feedback quantizer, the optimal feedback quantizer with the pre/post-filters [17] (dotted curve), and the uniform quantizer (dashed curve), where $\circ$, $*$, and $\square$ correspond to the oversampling ratios $\lambda = 2$, $\lambda = 3$, and $\lambda = 4$, respectively.

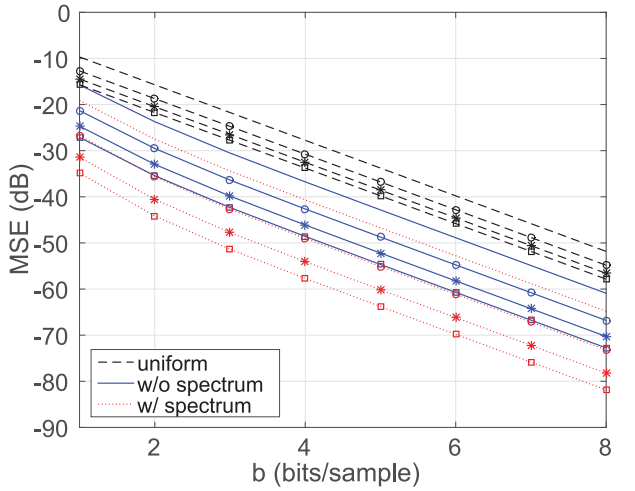


Fig. 6. MSEs of the optimal feedback quantizer without pre/post-filters, the optimal feedback quantizer with pre/post-filters [17] (dotted curve), and the uniform quantizer (dashed curve) with different oversampling rates λ, for a colored input, where $\circ$, $*$, and $\square$ correspond to the oversampling ratios $\lambda = 2$, $\lambda = 3$, and $\lambda = 4$, respectively.

The feedback quantizer has an approximately 10 dB gain against the uniform quantizer that is enabled by utilizing the feedback filter that is optimized based on the system $P[z]$. A further gain is obtained by exploiting the input spectrum for the quantizer having an optimized feedback filter and pre/post-filters. For all quantizers, as the oversampling ratio increases, the MSE decreases and the increment of the MSE gain decreases. It can be also observed that if the number of bits per time is fixed, oversampling degrades the MSE of the optimal error feedback quantizer without pre/post-filters as we have analyzed. This also holds true for the optimal error feedback quantizer with pre/post-filters and the uniform quantizer.

Fig. 7 shows the MSEs of the optimal feedback quantizer, the optimal feedback quantizer with the pre/post-filters and the

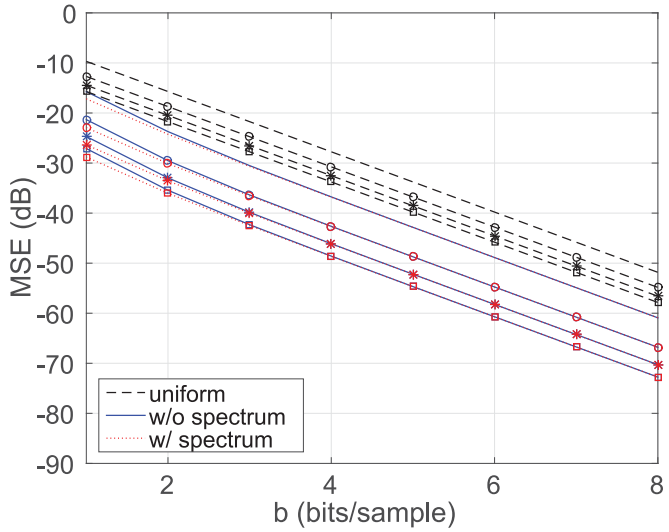


Fig. 7. MSEs of the optimal feedback quantizer without pre/post-filters, the optimal feedback quantizer with pre/post-filters [17] (dotted curve), and the uniform quantizer (dashed curve) with different oversampling rates, λ, for a white input, where $\circ$, $*$, and $\square$ correspond to the oversampling ratios $\lambda = 2$, $\lambda = 3$, and $\lambda = 4$, respectively.

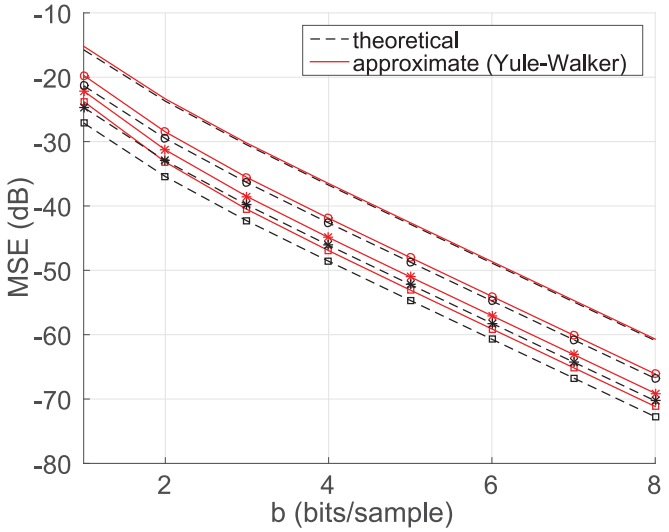


Fig. 8. MSEs of the feedback quantizers with ideal feedback filters and feedback quantizers with IIR feedback filters of order four approximated by the Yule-Walter method for different oversampling rates λ, where $\circ$, $*$, and $\square$ correspond to the oversampling ratios $\lambda = 2$, $\lambda = 3$, and $\lambda = 4$, respectively.

uniform quantizer for a white input signal. The optimal feedback quantizer and the optimal feedback quantizer with the pre/post-filters have a gain of more than 10 dB over the uniform quantizer. As the input has a flat spectrum, the optimal feedback quantizer has almost the same performance as the optimal feedback quantizer with the pre/post-filters. It should be noted that the latter requires additional pre/post-filters.

In Figs. 6 and 7, we have utilized ideal feedback filters both for the feedback quantizer and the feedback quantizer with the pre/post-filters, which cannot be implemented in practice. We approximate the ideal feedback filters for the optimal feedback quantizers using IIR filters of order four by the Yule-Walker method [19] with a normalization and by the LMI-based method discussed in Section V.

Fig. 8 illustrates the MSEs of the feedback quantizers with ideal optimal feedback filters and the feedback quantizers with feedback filters of order four approximated by the Yule-Walker method, whereas Fig. 9 presents the MSEs of the feedback quantizers with ideal feedback filters and the feedback quantizers with feedback filters of order four approximated by the LMI-based method. The approximation by the Yule-Walker method suffers a small loss, while the approximation by the LMI-based method has almost the same MSE as the ideal case.

If the order of the IIR filter is increased, a better performance can be expected for the Yule-Walker method. On the other hand, it is known that the minimum of (50) is attained by $P[\mathrm{z}]$ having the same order as $R[\mathrm{z}]$ [31], [32]. Therefore, if the order of $P[\mathrm{z}]$ is increased more than the order of $R[\mathrm{z}]$, the MSE does not improve. In this example, as the order of $P[\mathrm{z}]$ is four, an $R[\mathrm{z}]$ of order four is sufficient for the LMI-based method. The performance difference between the Yule-Walker method and the LMI-based method may be decreased by increasing the filter order for the Yule-Walker method.

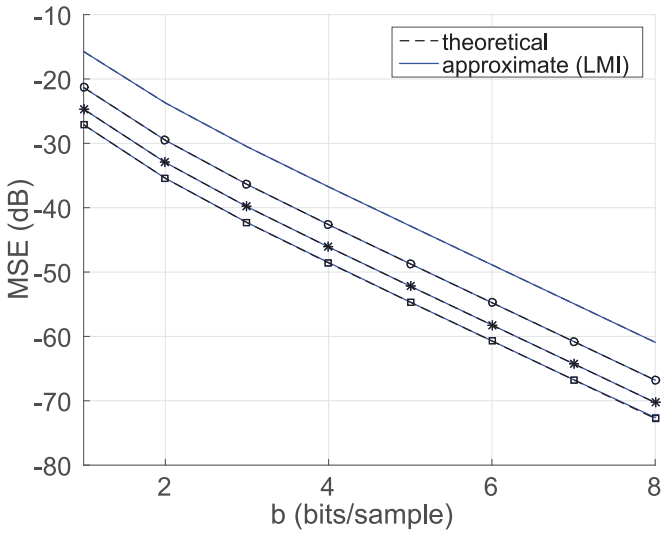


Fig. 9. MSEs of the feedback quantizers with optimal feedback filters and feedback quantizers with IIR feedback filters of order four approximated by the LMI-based method for different oversampling rates λ, where $\circ$, $*$, and $\square$ correspond to the oversampling ratios $\lambda = 2$, $\lambda = 3$, and $\lambda = 4$, respectively.

VII. Conclusion

We have presented the MSE analysis of quantizers with error feedback. We have shown that the amplitude response of the optimal error feedback filter that minimizes the MSE can be parameterized by one parameter and can be found numerically. With the optimal error feedback filter, the relationship between the number of bits used for the quantization and the achievable MSE has been clarified. We have also developed two designs for the IIR error feedback filters for approximating the ideal optimal error feedback filters. Numerical examples have been provided to demonstrate our analysis and synthesis.

APPENDIX A
PROOF OF THEOREM 1

Suppose that $r(\omega)$ is optimal. If $c_0 > 0$, then $r'(\omega) = r(\omega)e^{-c_0}$ gives a smaller value for (19) that contradicts the optimally of $r(\omega)$. Thus c_0 in (16) has to be zero.

Let us denote the norm of $r_{opt}(\omega)$ as c_{opt} and define the set of $r(\omega) \in \mathcal{C}_0$ having the same norm as $r_{opt}(\omega)$ by $\mathcal{C}_{opt}$. As $\mathcal{C}_0 \cap \mathcal{C}_{opt} \subset \mathcal{C}_0 \cap \mathcal{C}_1$, the minimization of (19) subject to $\mathcal{C}_0 \cap \mathcal{C}_1$ is equivalent to the minimization of $||p(\omega)r(\omega)||^2$ subject to

$$||r(\omega)||^2 = c_{opt} \tag{54}$$

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln r(\omega) d\omega = 0 \tag{55}$$

The Lagrangian of this problem is given by

$$L(r(\omega)) := p^2(\omega)r^2(\omega) + \mu_1 r^2(\omega) + \mu_2 \ln r(\omega) \tag{56}$$

where μ_1 and μ_2 are the Lagrange multipliers. Then, the optimal $r(\omega)$ has to satisfy

$$\frac{\partial}{\partial r}L(r(\omega)) = 2p^2(\omega)r(\omega) + 2\mu_1 r(\omega) + \mu_2 \frac{1}{r(\omega)} = 0$$

$$\text{a.e. } \omega \in [-\pi, \pi]. \tag{57}$$

Thus, for a.e. $\omega \in [-\pi, \pi]$, we need

$$2(p^2(\omega) + \mu_1)r^2(\omega) = -\mu_2. \tag{58}$$

If $p(\omega)$ is almost constant, then $r(\omega)$ has to be almost constant; from (55) $r(\omega) = 1$, implying that $R[\mathrm{z}] = 1$. Hence, the error feedback filter $R[\mathrm{z}] - 1$ is not required and the uniform quantizer is optimal. In the following proof, we only consider $p(\omega)$ that is not almost constant.

As $p(\omega)$ is not almost constant, $p^2(\omega) + \mu_1$ cannot be zero over any interval $[-\pi, \pi]$, having a nonzero measure. As $r(\omega) \neq 0$, μ_2 cannot be zero. Therefore, we obtain

$$r(\omega) = \frac{\theta}{\sqrt{p^2(\omega) + \alpha}} \tag{59}$$

where $\theta = \sqrt{-\mu_2}$ and $\alpha = \mu_1$.

Substituting (59) in (55) results in

$$\int_{-\pi}^{\pi} \left(\ln \theta - \frac{1}{2}\ln(p^2(\omega) + \alpha) d\omega \right) = 0 \tag{60}$$

from which we obtain (24).

APPENDIX B
PROOF OF THEOREM 2

Differentiating $\Phi(\alpha)$ with respect to α, we have

$$\dot{\Phi}(\alpha) = \frac{\dot{N}(\alpha)(\nu - C(\alpha)) + N(\alpha)\dot{C}(\alpha)}{[\nu - C(\alpha)]^2}. \tag{61}$$

With (23), $N(\alpha)$ can be expressed as

$$N(\alpha) = \frac{\theta^2(\alpha)}{2\pi}\int_{-\pi}^{\pi} \frac{p^2(\omega)}{p^2(\omega) + \alpha} d\omega. \tag{62}$$

From

$$\frac{d}{d\alpha}\theta(\alpha) = \frac{\theta(\alpha)}{4\pi}\int_{-\pi}^{\pi} \frac{1}{p^2(\omega) + \alpha} d\omega \tag{63}$$

the derivative of $N(\alpha)$ is found to be

$$\begin{aligned} &\frac{d}{d\alpha}N(\alpha) \\ &= \frac{2\theta(\alpha)}{2\pi}\dot{\theta}(\alpha)\int_{-\pi}^{\pi} \frac{p^2(\omega)}{p^2(\omega) + \alpha} d\omega \\ &\quad - \frac{\theta^2(\alpha)}{2\pi}\int_{-\pi}^{\pi} \frac{p^2(\omega)}{(p^2(\omega) + \alpha)^2} d\omega \qquad (64) \\ &= \frac{\theta^2(\alpha)}{2\pi}\left\{ \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{1}{p^2(\omega) + \alpha} d\omega \int_{-\pi}^{\pi} \frac{p^2(\omega)}{p^2(\omega) + \alpha} d\omega \right. \\ &\quad \left. - \int_{-\pi}^{\pi} \frac{p^2(\omega)}{(p^2(\omega) + \alpha)^2} d\omega \right\} \qquad (65) \end{aligned}$$

It can be seen that $\dot{N}(0) = 0$. To prove

$$\dot{N}(\alpha) < 0 \quad \text{for} \quad \alpha < 0, \quad \dot{N}(\alpha) > 0 \quad \text{for} \quad \alpha > 0. \tag{66}$$

we introduce the next definition and theorem given in [17].

Definition 2: We say that two function ϕ, ψ: $[a, b] \to \mathbb{R}$ are similarly functionally related if and only if there exists a monotonically increasing function $G(\cdot)$ such that $\phi = G(\psi)$ for all $x \in [a, b]$. Similarly, if there exists a monotonically decreasing function such that $\phi = G(\psi)$ for all $x \in [a, b]$, we say that ϕ and ψ are oppositely functionally related.

Theorem 5: If ϕ, ψ: $[a, b] \to \mathbb{R}$ are similarly functionally related, then

$$[b - a]\int_a^b \phi(x)\psi(x)dx \geq \int_a^b \phi(x)dx \int_a^b \psi(x)dx. \tag{67}$$

If ϕ and ψ are oppositely functionally related, then the equality in (67) is reversed. In either case, equality is achieved if and only $\psi(x)$ is almost constant.

We set $\psi(\omega) = \frac{1}{p^2(\omega)+\alpha}$ and $\phi(\omega) = \frac{p^2(\omega)}{p^2(\omega)+\alpha}$ that are related to $\alpha \neq 0$ such that

$$\phi(\omega) = \frac{p^2(\omega)}{p^2(\omega) + \alpha} = 1 - \frac{\alpha}{p^2(\omega) + \alpha} = 1 - \alpha\psi(\omega). \tag{68}$$

Thus, $\phi(\omega)$ and $\psi(\omega)$ are similarly functionally related for $\alpha < 0$, whereas $\phi(\omega)$ and $\psi(\omega)$ are oppositely functionally related for $\alpha > 0$. Then, we can apply theorem 5 to find that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \psi(\omega)d\omega \int_{-\pi}^{\pi} \phi(\omega)d\omega - \int_{-\pi}^{\pi} \phi(\omega)\psi(\omega)d\omega$$

is negative for $\alpha < 0$, whereas it is positive for $\alpha > 0$, proving (66).

On the other hand, differentiating $C(\alpha)$ with respect to α gives

$$\frac{d}{d\alpha}C(\alpha) = \frac{2\theta(\alpha)}{2\pi}\dot{\theta}(\alpha)\int_{-\pi}^{\pi}\frac{1}{p^2(\omega)+\alpha}d\omega - \frac{\theta^2(\alpha)}{2\pi}\int_{-\pi}^{\pi}\frac{1}{(p^2(\omega)+\alpha)^2}d\omega \tag{69}$$

$$= \frac{\theta^2(\alpha)}{2\pi}\left\{\frac{1}{2\pi}\left(\int_{-\pi}^{\pi}\frac{1}{p^2(\omega)+\alpha}d\omega\right)^2 - \int_{-\pi}^{\pi}\frac{1}{(p^2(\omega)+\alpha)^2}d\omega\right\} \tag{70}$$

From the Cauchy-Schwarz inequality, we find that $\dot{C}(\alpha) < 0$.

We note that $\nu - C(\alpha) > 0$ and $N(\alpha) > 0$ in (61). For $\alpha < 0$, from $\dot{N}(\alpha) < 0$ and $\dot{C}(\alpha) < 0$ in (61), $\dot{\Phi}(\alpha) < 0$. At $\alpha = 0$, from $\dot{N}(0) = 0$, we have

$$\dot{\Phi}(0) = \frac{N(0)\dot{C}(0)}{[\nu - C(0)]^2} < 0. \tag{71}$$

As $\Phi(\alpha)$ is continuous in α, the minimum of $\Phi(\alpha)$ is achieved at α greater than zero; i.e., we can conclude that $\alpha_{opt} > 0$.

A necessary condition for α_{opt} is $\dot{\Phi}(\alpha_{opt}) = 0$. As $\dot{N}(\alpha) \neq 0$ for $\alpha > 0$ and $\alpha_{opt} > 0$, we find from (61) that the numerator has to be zero, leading to

$$\nu = \frac{\dot{N}(\alpha_{opt})C(\alpha_{opt}) - N(\alpha_{opt})\dot{C}(\alpha_{opt})}{\dot{N}(\alpha_{opt})}. \tag{72}$$

From (65) and (70), we get

$$\left(\dot{N}(\alpha_{opt})C(\alpha_{opt}) - N(\alpha_{opt})\dot{C}(\alpha_{opt})\right) \Big/ \left(\frac{\theta^2(\alpha_{opt})}{2\pi}\right)^2$$

$$= \left\{\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{p^2(\omega)+\alpha_{opt}}d\omega\int_{-\pi}^{\pi}\frac{p^2(\omega)}{p^2(\omega)+\alpha_{opt}}d\omega - \int_{-\pi}^{\pi}\frac{p^2(\omega)}{(p^2(\omega)+\alpha_{opt})^2}d\omega\right\}\int_{-\pi}^{\pi}\frac{1}{p^2(\omega)+\alpha_{opt}}d\omega$$
$$- \int_{-\pi}^{\pi}\frac{p^2(\omega)}{p^2(\omega)+\alpha_{opt}}d\omega\left\{\frac{1}{2\pi}\left(\int_{-\pi}^{\pi}\frac{1}{p^2(\omega)+\alpha_{opt}}d\omega\right)^2 - \int_{-\pi}^{\pi}\frac{1}{(p^2(\omega)+\alpha_{opt})^2}d\omega\right\}$$
$$= -\int_{-\pi}^{\pi}\frac{p^2(\omega)}{(p^2(\omega)+\alpha_{opt})^2}d\omega\int_{-\pi}^{\pi}\frac{1}{p^2(\omega)+\alpha_{opt}}d\omega + \int_{-\pi}^{\pi}\frac{p^2(\omega)}{p^2(\omega)+\alpha_{opt}}d\omega\int_{-\pi}^{\pi}\frac{1}{(p^2(\omega)+\alpha_{opt})^2}d\omega. \tag{73}$$

Substituting

$$\frac{1}{p^2(\omega)+\alpha_{opt}} = \frac{1}{\alpha_{opt}}\left(1 - \frac{p^2(\omega)}{p^2(\omega)+\alpha_{opt}}\right) \tag{74}$$

in (73) results in

$$-\int_{-\pi}^{\pi}\frac{p^2(\omega)}{(p^2(\omega)+\alpha_{opt})^2}d\omega\int_{-\pi}^{\pi}\frac{1}{\alpha_{opt}}\left(1 - \frac{p^2(\omega)}{p^2(\omega)+\alpha_{opt}}\right)d\omega + \int_{-\pi}^{\pi}\frac{p^2(\omega)}{p^2(\omega)+\alpha_{opt}}d\omega\int_{-\pi}^{\pi}\frac{1}{\alpha_{opt}}\left(1 - \frac{p^2(\omega)}{p^2(\omega)+\alpha_{opt}}\right)\cdot\frac{1}{p^2(\omega)+\alpha_{opt}}d\omega$$
$$= -\frac{2\pi}{\alpha_{opt}}\int_{-\pi}^{\pi}\frac{p^2(\omega)}{(p^2(\omega)+\alpha_{opt})^2}d\omega + \frac{1}{\alpha_{opt}}\int_{-\pi}^{\pi}\frac{p^2(\omega)}{p^2(\omega)+\alpha_{opt}}d\omega\int_{-\pi}^{\pi}\frac{1}{p^2(\omega)+\alpha_{opt}}d\omega$$
$$= \frac{2\pi}{\alpha_{opt}}\dot{N}(\alpha_{opt})\Big/\left(\frac{\theta^2(\alpha_{opt})}{2\pi}\right) \tag{75}$$

which shows that

$$\dot{N}(\alpha_{opt})C(\alpha_{opt}) - N(\alpha_{opt})\dot{C}(\alpha_{opt}) = \frac{\theta^2(\alpha_{opt})}{\alpha_{opt}}\dot{N}(\alpha_{opt}). \tag{76}$$

Substituting this in (72) gives (31).

Appendix C
Proof of Theorem 3

From (31), we obtain

$$\nu = \exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln\left[p_\lambda^2(\omega)+\alpha_{opt}(\nu,\lambda)\right]d\omega\right)\Big/\alpha_{opt}(\nu,\lambda) \tag{77}$$

$$= \exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln\frac{p_\lambda^2(\omega)+\alpha_{opt}(\nu,\lambda)}{\alpha_{opt}(\nu,\lambda)}d\omega\right). \tag{78}$$

Substituting (35) in (78), we have

$$\nu = \exp\left(\frac{1}{2\pi}\int_{-\omega_c}^{\omega_c}\ln\frac{p_1^2(\lambda\omega)+\alpha_{opt}(\nu,\lambda)}{\alpha_{opt}(\nu,\lambda)}d\omega\right). \tag{79}$$

The change of the variable as $\omega' = \lambda\omega$ gives

$$\nu = \exp\left(\frac{1}{2\pi\lambda}\int_{-\pi}^{\pi}\ln\frac{p_1^2(\omega')+\alpha_{opt}(\nu,\lambda)}{\alpha_{opt}(\nu,\lambda)}d\omega'\right). \tag{80}$$

Then we have

$$\nu^\lambda = \exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln\left[p_1^2(\omega')+\alpha_{opt}(\nu,\lambda)\right]d\omega'\right)\Big/\alpha_{opt}(\nu,\lambda) \tag{81}$$

that proves

$$\alpha_{opt}(\nu,\lambda) = \alpha_{opt}(\nu^\lambda, 1) \tag{82}$$

hence (38).

Appendix D
Convex formalization

We show that the optimization problem (50) is cast into a convex. More details could be found in [34].

We denote $\mu_r = ||r_{opt}(\omega)||^2$. The optimization problem is equivalent to minimizing μ_ϵ subject to $R[\infty] = 1$ and

$$||P[\mathrm{z}]R[\mathrm{z}]||^2 \leq \mu_\epsilon \tag{83}$$

$$||R[\mathrm{z}]||^2 \leq \mu_r. \tag{84}$$

Let the order of $P[\mathrm{z}]$ be n and the (A, B, C, D) matrices of a state-space realization of $P[\mathrm{z}]$ be (A_p, B_p, C_p, D_p). The state-space realization of $P[\mathrm{z}]R[\mathrm{z}]$ can be expressed as

$$x_{k+1} = \mathcal{A}x_k + \mathcal{B}w_k \tag{85}$$

$$\epsilon_k = \mathcal{C}x_k + \mathcal{D}w_k \tag{86}$$

where $(A_r, B_r, C_r, 1)$ are (A, B, C, D) matrices of a state-space realization of $R[\mathrm{z}]$ and

$$\mathcal{A} = \begin{bmatrix} A_p & B_pC_r \\ \mathbf{0} & A_r \end{bmatrix}, \ \mathcal{B} = \begin{bmatrix} B_p \\ B_r \end{bmatrix}$$

$$\mathcal{C} = \begin{bmatrix} C_p & D_pC_r \end{bmatrix}, \ \mathcal{D} = D_p.$$

It is known that $||P[\mathrm{z}]R[\mathrm{z}]||_2^2 < \mu_\epsilon$ if and only if there exists a positive definite matrix $\mathcal{P}$ such that [33]

$$\begin{bmatrix} \mathcal{P} & \mathcal{P}\mathcal{A} & \mathcal{P}\mathcal{B} \\ \mathcal{A}^T\mathcal{P} & \mathcal{P} & \mathbf{0} \\ \mathcal{B}^T\mathcal{P} & \mathbf{0} & 1 \end{bmatrix} \succ \mathbf{0} \tag{87}$$

$$\begin{bmatrix} \mu_\epsilon & \mathcal{C} & \mathcal{D} \\ \mathcal{C}^T & \mathcal{P} & \mathbf{0} \\ \mathcal{D}^T & \mathbf{0} & 1 \end{bmatrix} \succ \mathbf{0}. \tag{88}$$

On the other hand, $||R[\mathrm{z}]||_2^2 < \mu_r$ if and only if there exists a positive definite matrix $\mathcal{P}$ that satisfies (87) and

$$\begin{bmatrix} \mu_r - 1 & \tilde{\mathcal{C}} \\ \tilde{\mathcal{C}}^T & \mathcal{P} \end{bmatrix} \succ \mathbf{0} \tag{89}$$

where

$$\tilde{\mathcal{C}} = [\,\mathbf{0} \quad C_r\,]. \tag{90}$$

Equation (87) is a bilinear matrix inequality (BMI), which can be converted into an linear matrix inequality (LMI) using a change of variables [31], [32].

The set of $n \times n$ positive definite matrices is denoted as $PD(n)$. We define the following matrices $\{P_f, P_g, W_f, W_g, W_h, L\}$, where $P_f \in PD(n)$, $P_g \in PD(n)$, $W_f \in \mathbb{R}^{1\times n}$, $W_g \in \mathbb{R}^{n\times 1}$, $W_h \in \mathbb{R}$, $L \in \mathbb{R}^{n\times n}$, with P_f and P_g. Let us also define matrices from $\{P_f, P_g, W_f, W_g, W_h, L\}$ as

$$\mathcal{P}^{-1} = \begin{bmatrix} P_f & S_f \\ S_f & S_f \end{bmatrix} \tag{91}$$

$$U = \begin{bmatrix} P_f & I_n \\ S_f & \mathbf{0} \end{bmatrix} \tag{92}$$

$$P_g = (P_f - S_f)^{-1} \tag{93}$$

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

$$M_\mathcal{A} = \begin{bmatrix} A_pP_f + B_pW_f & A_p \\ L & P_gA_p \end{bmatrix} \tag{94}$$

$$M_\mathcal{B} = \begin{bmatrix} B_p \\ W_g \end{bmatrix} \tag{95}$$

$$M_\mathcal{C} = \begin{bmatrix} C_pP_f + D_pW_f & C_p \end{bmatrix} \tag{96}$$

$$M_\mathcal{P} = \begin{bmatrix} P_f & I_n \\ I_n & P_g \end{bmatrix} \tag{97}$$

$$M_{\tilde{\mathcal{C}}} = \begin{bmatrix} W_f & \mathbf{0} \end{bmatrix}. \tag{98}$$

Then, (87), (88), and (89) can be found to be equivalent to

$$\begin{bmatrix} M_\mathcal{P} & M_\mathcal{A} & M_\mathcal{B} \\ M_\mathcal{A}^T & M_\mathcal{P} & \mathbf{0} \\ M_\mathcal{B}^T & \mathbf{0} & 1 \end{bmatrix} \succ \mathbf{0}, \tag{99}$$

$$\begin{bmatrix} \mu_\epsilon & M_\mathcal{C} & \mathcal{D}^T \\ M_\mathcal{C}^T & M_\mathcal{P} & \mathbf{0} \\ \mathcal{D} & \mathbf{0} & 1 \end{bmatrix} \succ \mathbf{0}, \tag{100}$$

and

$$\begin{bmatrix} \mu_r - 1 & M_{\tilde{\mathcal{C}}} \\ M_{\tilde{\mathcal{C}}}^T & M_\mathcal{P} \end{bmatrix} \succ \mathbf{0}. \tag{101}$$

Since (99), (100), and (101) are convex LMIs, the minimization of μ_ϵ subject to (99), (100), and (101) is a convex optimization.

Finally, once the optimal solution is obtained, we reconstruct (A_r, B_r, C_r) with the inverse transformation.

REFERENCES

[1] K. Sayood, *Introduction to Data Compression*. Oxford, U.K.: Newnes, 2012.

[2] C. Mullis and R. Roberts, "Synthesis of minimum roundoff noise fixed point digital filters," *IEEE Trans. Circuits Syst.*, vol. 23, no. 9, pp. 551–562, Sep. 1976.

[3] T. Laakso and I. 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] 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. 30, no. 6, pp. 963–973, Dec. 1982.

[5] 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.

[6] 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.

[7] 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.

[8] S. Ohno, Y. Wakasa, and M. Nagata, "Optimal error feedback filters for uniform quantizers at remote sensors," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 3866–3870.

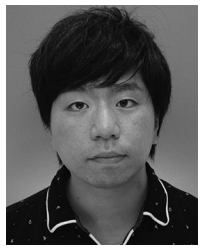
[9] S. Pavan, R. Schreier, and G. C. Temes, *Understanding Delta-sigma Data Converters*, 2nd ed. New York, NY, USA: Wiley, 2017.

[10] E. Janssen and D. Reefman, "Super-audio CD: An introduction," *IEEE Signal Process. Mag.*, vol. 20, no. 4, pp. 83–90, Jul. 2003.

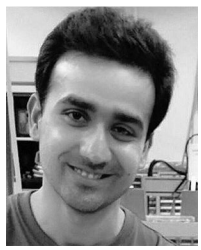
[11] U. Gustavsson, T. Eriksson, and C. Fager, "Quantization noise minimization in $\Sigma\Delta$ modulation based RF transmitter architectures," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 12, pp. 3082–3091, Dec. 2010.
[12] P. T. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in *Proc. 42nd Annu. Conf. Inf. Sci. Syst.*, 2008, pp. 16–21.
[13] A. Fazel, A. Gore, and S. Chakrabartty, "Resolution enhancement in $\Sigma\Delta$ learners for superresolution source separation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1193–1204, Mar. 2010.
[14] N. Thao, "Vector quantization analysis of $\Sigma\Delta$ modulation," *IEEE Trans. Signal Process.*, vol. 44, no. 4, pp. 808–817, Apr. 1996.
[15] I. Daubechies and R. DeVore, "Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order," *Annals Math.*, vol. 158, pp. 679–710, 2003.
[16] P. Deift, F. Krahmer, and C. S. Güntürk, "An optimal family of exponentially accurate one-bit sigma-delta quantization schemes," *Commun. Pure Appl. Math.*, vol. 64, no. 7, pp. 883–919, 2011.
[17] M. Derpich, E. Silva, D. Quevedo, and G. Goodwin, "On optimal perfect reconstruction feedback quantizers," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3871–3890, Aug. 2008.
[18] S. Ohno, T. Shiraki, M. R. Tariq, and M. Nagahara, "Rate-distortion analysis of delta-sigma modulators," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 4581–4585.
[19] B. Friedlander and B. Porat, "The modified Yule-Walker method of ARMA spectral estimation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-20, no. 2, pp. 158–173, Mar. 1984.
[20] B. Widrow and I. Kollár, *Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications*. New York NY, USA: Cambridge Univ. Press, 2008.
[21] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 446–472, Jul. 1948.
[22] R. M. Gray, "Quantization noise spectra," *IEEE Trans. Inf. Theory*, vol. 36, no. 6, pp. 1220–1244, Nov. 1990.
[23] M. Kozak and I. Kale, *Oversampled Delta-Sigma Modulators: Analysis, Applications and Novel Topologies*. New York NY, USA: Springer, 2007.
[24] I. Galton, "Granular quantization noise in a class of delta-sigma modulators," *IEEE Trans. Inf. Theory*, vol. 40, no. 3, pp. 848–859, May 1994.
[25] R. A. Wannamaker, S. P. Lipshitz, J. Vanderkooy, and J. N. Wright, "A theory of nonsubtractive dither," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 499–516, Feb. 2000.
[26] W. Rudin, *Real and Complex Analysis*, 3rd ed. New York, NY, USA: McGraw-Hill, 1987.
[27] P. Noll, "On predictive quantizing schemes," *Bell Syst. Tech. J.*, vol. 57, no. 5, pp. 1499–1532, May 1978.
[28] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, vol. 159. New York, NY, USA: Springer Science & Business Media, 2012.
[29] A. Gersho, "Principles of quantization," *IEEE Trans. Circuits Syst.*, vol. 25, no. 7, pp. 427–436, Jul. 1978.
[30] M. Grant and S. Boyd, "CVX: MATLAB software for disciplined convex programming, version 2.0 beta," Sep. 2012. [Online]. Available: http://cvxr.com/cvx
[31] 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.
[32] 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.
[33] S. Boyd, L. E. Ghaoul, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*. Philadelphia, PA, USA: SIAM, 1997.
[34] 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.

Shuichi Ohno (M'95–SM'11) received the B.E., M.E., and Dr. Eng. degrees in applied mathematics and physics from Kyoto University, Kyoto, Japan, in 1990, 1992, and 1995, respectively.

From 1995 to 1999, he was a Research Associate in 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 at the University of Minnesota as a Visiting Researcher. Since 2010, he has been an Associate Professor with the Department of System Cybernetics, Hiroshima University, Higashihiroshima, Japan. His current research interests include the areas of signal processing in communication, wireless communications, and adaptive signal processing. He is a member of the SICE, IEICE, and ISCIE. He was as an Associated Editor for the IEEE SIGNAL PROCESSING LETTERS from 2001 to 2003.



Teruyuki Shiraki received the Bachelor's degree in electrical engineering in 2014 and the Master's degree in electrical engineering in 2016, both from Hiroshima University, Hiroshima, Japan.



M. Rizwan Tariq (S'14) was born in Islamabad, Pakistan, where he received the B.E. degree in electrical engineering in 2010 and the M.E. degree in 2015 in system cybernetics from Hiroshima University, Higashihiroshima, Japan, where he is currently working toward the Ph.D. degree in signal processing. His current research interests include signal processing, wireless communication, and control systems. He received the Japanese Government Scholarship Program for research students.

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

He is currently a Full Professor in the Institute of Environmental Science and Technology, University of Kitakyushu, Kitakyushu, Japan. He is also a Visiting Professor in Indian Institute of Technology Bombay since 2017. His research interests include control theory, machine learning, and sparse modeling.

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