IEICE TRANSACTIONS

on Fundamentals of Electronics, Communications and Computer Sciences

DOI:10.1587/transfun.2024EAP1085

Publicized:2024/12/19

A PUBLICATION OF THE ENGINEERING SCIENCES SOCIETY
The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

Design of Delta-Sigma modulators for closed loop systems with quantization and saturation*

Shuichi OHNO[†a)], ***Member***, **Shenjian WANG**[†], ***and*** **Kiyotsugu TAKABA**[††], ***Nonmembers***

SUMMARY This paper studies $\Delta\Sigma$ modulators for discrete-time closed loop systems. $\Delta\Sigma$ modulators have been originally developed as efficient analog-to-digital converters (ADCs). Recently, $\Delta\Sigma$ modulators are designed based on the characteristics of the system that uses the $\Delta\Sigma$ modulator. For example in a control system, quantization may degrade control performance due to quantization errors, while the input to any practical system is limited to a range. Then, the saturation of the control input may cause windup phenomena such as overshoots of the system outputs and instability of the control system. In this paper, we propose a design of $\Delta\Sigma$ modulators to mitigate the effects of quantization and saturation in a discrete-time closed loop system. We design the $\Delta\Sigma$ modulator to minimize the norm of the quantization error at the system output to reduce the effects of the quantization error under a stability condition to avoid the saturation of the input on the closed-loop system. Numerical examples are provided to see the effectiveness of our proposed design.

key words: *$\Delta\Sigma$ modulator, quantization, input saturation*

1. Introduction

A $\Delta\Sigma$ modulator consists of a conventional (usually uniform) static quantizer and a filter that feeds back the quantization error of the conventional quantizer to the input to the conventional quantizer. Thanks to this simple structure, $\Delta\Sigma$ modulators have been often utilized as analog-to-digital converters (ADCs) and digital-to-analog converters (DACs). (See, e.g., [1] and references therein).

Recently, the feedback filter has been designed to mitigate the impact of quantization by leveraging the system's inherent characteristics [2], [3]. In this study, we focus on designing a $\Delta\Sigma$ modulator tailored for a discrete-time closed-loop system incorporating quantization and saturation.

To reduce the effect of quantization errors on the system output, *dynamic quantizers* have been employed, where the quantizer parameters are designed using linear programming (LP) [4] or convex optimization [5]. However, most of these methods do not consider stability in the presence of input saturation.

In a control system, with access to ideal actuators capable of achieving any desired controller outputs, designing various types of controllers is feasible. However, practical scenarios entail limitations on actuator outputs, constraining the acceptable input range. When the control input surpasses this range, input saturation occurs, necessitating meticulous input design for systems with such limitations. Overlooking input saturation during controller design can result in significant degradation of control performance, manifesting as output overshoots or instability in closed-loop systems, commonly referred to as windup phenomena.

Input saturation can be effectively represented as a saturator. In many anti-windup controller designs, the input and output of the saturator are looped back to the controller to counteract windup phenomena (for instance, see [6] and related literature). Linear conditions are employed for anti-windup controller design in [7]. For continuous-time systems, static gains are devised for feedback in [8], while the stability region for continuous-time systems is analyzed in [9], with [10] focusing on discrete-time systems. Additionally, [11] diminishes the deviation norm of system response from its ideal counterpart without saturation for continuous-time systems, employing filters for the feedback signal from the saturator, with similar findings extended to discrete-time systems in [12]. In [13], an anti-windup controller has been designed for a discrete-valued input control system with a dynamic quantizer.

In this paper, we reveal that $\Delta\Sigma$ modulators encompass the functionalities of most anti-windup controllers, thus serving not only as quantizers but also as anti-windup controllers. Subsequently, we introduce a methodology for designing the $\Delta\Sigma$ modulator aimed at minimizing the norm of the quantization error at the system output, subject to a stability criterion governing the closed-loop system. The norm is assessed through the transfer function from the quantization error to the system output, while the stability criterion is delineated by the circle criterion as derived in [14]. Both conditions are formulated as bilinear matrix inequalities (BMI) involving the multiplication of design

[†] The author is with the Faculty of Informatics, Osaka Metropolitan University.

[††] The author is with the Faculty Faculty of Science and Engineering, Ritsumeikan University.

[*] A conference version of this paper, entitled as "Delta-Sigma modulators for discrete-time closed loop control systems with quantization and saturation", was presented at 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2023.

a) E-mail: ohno@omu.ac.jp

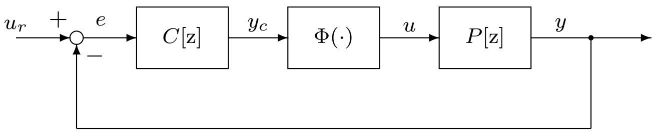


Fig. 1: A feedback control system with a constraint on the range of the input to the plant P[z].

variables. By permitting the sharing of a common Lyapunov matrix, our design objective is transformed into a convex optimization problem, amenable to numerical solution. The efficacy of the proposed methodology is validated through numerical experiments.

2. Closed-loop systems with input quantization and saturation

For a typical application of our quantization, we consider a discretized feedback control system depicted in Fig. 1, where following conventions in control, P[z] and C[z] are respectively transfer functions of the plant and the controller. For simplicity, we assume both the plant and the controller are single-input and single-output linear time-invariant systems. The orders of P[z] and C[z] are n_p and n_c, respectively. We denote the z-transform of a causal discrete-time signal with a lowercase letter represented by its corresponding uppercase letter.

In Fig. 1, $u_r(k) \in \mathbb{R}$ stands for the reference signal at time k, $e(k) \in \mathbb{R}$ represents the control deviation, $y_c(k) \in \mathbb{R}$ signifies the output of the controller, $u(k) \in \mathbb{R}$ denotes the input of the plant, and $y(k) \in \mathbb{R}$ is the control output.

Let us express the state space representations of the plant and the controller as

$$x_p(k+1) = A_p x_p(k) + B_p u(k) \tag{1}$$

$$y(k) = C_p x_p(k) \tag{2}$$

and

$$x_c(k+1) = A_c x_c(k) + B_c e(k) \tag{3}$$

$$y_c(k) = C_c x_c(k) \tag{4}$$

with $A_p \in \mathbb{R}^{n_p \times n_p}$, $B_p \in \mathbb{R}^{n_p \times 1}$, $C_p \in \mathbb{R}^{1 \times n_p}$, $A_c \in \mathbb{R}^{n_c \times n_c}$, $B_c \in \mathbb{R}^{n_c \times 1}$, and $C_c \in \mathbb{R}^{1 \times n_c}$, where $x_p(k) \in \mathbb{R}^{n_p}$ and $x_c(k) \in \mathbb{R}^{n_c}$ are state vectors of the plant and the controller.

We assume that for a real number $L > 0$, the input range of the plant is confined within $[-L, L]$. For the sake of simplicity in presentation, this range is symmetrically centered around 0. Whenever the output of the controller surpasses this input range, denoted as $|y_c(k)| > L$, the input $u(k)$ to the plant becomes saturated. This saturation behavior is modeled using the saturation element $\Phi(\cdot)$ in Fig. 1, defined as

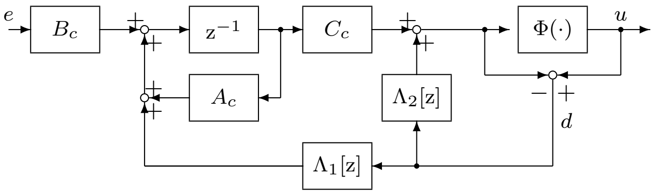


Fig. 2: An anti-windup controller.

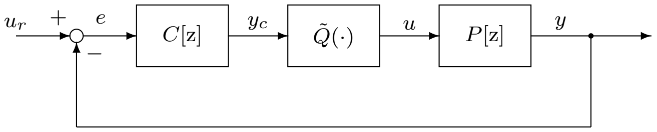


Fig. 3: A feedback control system with a quantizer.

$$\Phi(y_c(k)) = \begin{cases} L\mathrm{sgn}(y_c(k)), & |y_c(k)| > L \\ y_c(k), & |y_c(k)| \leq L \end{cases} \tag{5}$$

where sgn(·) denotes the sign function.

The disparity between the output of the controller and the input to the plant is denoted as

$$w_s(k) := u(k) - y_c(k). \tag{6}$$

The value of $w_s(k)$ is non-zero when $|y_c(k)| > L$. If $w_s(k)$ persists at a considerable magnitude for an extended duration, the control deviation amplifies due to inadequate feedback caused by saturation. Consequently, signals within the controller containing integrators escalate, resulting in phenomena like control output overshoot and system instability.

To mitigate the windup phenomena, anti-windup controllers have been developed. Most of linear time-invariant (LTI) anti-windup controllers can be represented as shown in Fig. 2, where Λ_1[z] and Λ_2[z] are LTI systems to be synthesized. When the saturation occurs, the difference $d(k)$ between the output of the controller and the input to the plant is fed back through Λ_1[z] and Λ_2[z] to reduce the magnitude of the output of the controller. It is important for Λ_2[z] to be strictly proper to avoid an algebraic loop. In the anti-windup controller, Λ_1[z] and Λ_2[z] are design variables. Many methods have been provided to determine Λ_1[z] and Λ_2[z]. (See [6] and references therein.)

We consider the scenario where the controller output is conveyed to the plant via a digital communication channel. When transmitting an analog signal through such a channel, it must undergo quantization to convert it into a digital signal. If the channel's capacity is limited, even the digital signal might be quantized into

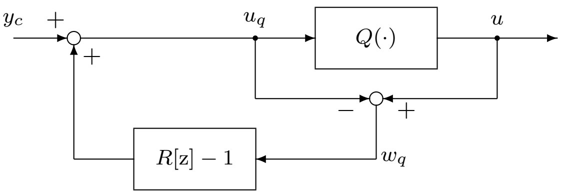


Fig. 4: A $\Delta\Sigma$ modulator.

a low-resolution form.

Let us assume that the saturation levels of the quantizer match those of the plant. Consequently, the feedback control system featuring input saturation and quantization can be depicted as shown in Fig. 3, where $\tilde{Q}(\cdot)$ represents the quantizer satisfying

$$\tilde{Q}(x) = L\mathrm{sgn}(x), \quad |x| > L \tag{7}$$

for a scalar input x.

To mitigate the quantization error and address the windup phenomena, we employ a $\Delta\Sigma$ modulator as our quantizer. The block diagram of a $\Delta\Sigma$ modulator is depicted in Fig. 4. Here, $u_q(k)$ is the input to the static quantizer $Q(\cdot)$, which is often the conventional uniform quantizer, and $w_q(k)$ is the quantization error at time k, defined as

$$w_q(k) = u(k) - u_q(k). \tag{8}$$

The feedback filter $R[\mathrm{z}]$ represents an n_r-order filter, commonly referred to as the noise transfer function (NTF) or noise shaping filter (NSF), satisfying $R[\infty] = 1$. It is important to note that the first term of the impulse response of $R[\mathrm{z}] - 1$ is zero, ensuring the absence of an algebraic loop in the $\Delta\Sigma$ modulator. Furthermore, it is worth noting that when $R[\mathrm{z}] = 1$, implying no feedback, the $\Delta\Sigma$ modulator simplifies to a static quantizer.

It is easy to see that the transfer function from $d(k) = u(k) - y_c(k)$ to the input to $\Phi(\cdot)$ of the anti-windup controller in Fig. 2 is given by

$$\Lambda[\mathrm{z}] := \Lambda_2[\mathrm{z}] + C_c(\mathrm{z}I_{n_c} - A_c)^{-1}\Lambda_1[\mathrm{z}] \tag{9}$$

where I_m is an $m \times m$ identity matrix. Let the order of $\Lambda[\mathrm{z}]$ be n_λ.

If we replace $Q(\cdot)$ and $R[\mathrm{z}] - 1$ in Fig. 4 with $\Phi(\cdot)$ and $\Lambda[\mathrm{z}]$, then we have a system equivalent to the anti-windup controller of Fig. 2. This means that if $n_r \geq n_c + n_\lambda$, then we can find a $\Delta\Sigma$ modulator that would be equivalent to any anti-windup controllers. Conversely, if we set $\Lambda_2[\mathrm{z}] = R[\mathrm{z}] - 1$ and $\Lambda_1[\mathrm{z}] = 0$ and replace $\Phi(\cdot)$ with $\tilde{Q}(\cdot)$, then the system with the anti-windup controller simplifies to the system with the $\Delta\Sigma$ modulator. These relations suggest that we can utilize the $\Delta\Sigma$ modulator not only for the reduction of the effect of the quantization error but also for the mitigation of the windup phenomena.

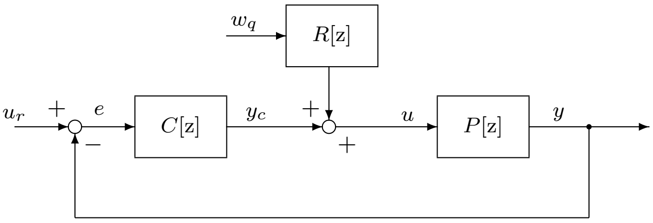


Fig. 5: An additive noise model for the system with a $\Delta\Sigma$ modulator.

3. $\Delta\Sigma$ modulator for quantization and saturation

Let us design a $\Delta\Sigma$ modulator to reduce the effects of quantization errors while adhering to a stability condition to mitigate windup phenomena.

From Fig. 4, one finds that $W_q[\mathrm{z}] = U[\mathrm{z}] - U_q[\mathrm{z}]$ and $U_q[\mathrm{z}] = Y_c[\mathrm{z}] + (R[\mathrm{z}] - 1)W_q[\mathrm{z}]$. By substituting the former into the latter, the z-transform of the output $u(k)$ of the $\Delta\Sigma$ modulator is represented by

$$U[\mathrm{z}] = Y_c[\mathrm{z}] + R[\mathrm{z}]W_q[\mathrm{z}]. \tag{10}$$

Subsequently, we establish the system depicted in Fig. 5 which is equivalent to the system illustrated in Fig. 3.

From Fig. 5, the transfer function from the quantization error of the $\Delta\Sigma$ modulator to the system output can be found as $P[\mathrm{z}]/(1 + P[\mathrm{z}]C[\mathrm{z}])$. Consequently, the z-transform $Y[\mathrm{z}]$ of the system output can be expressed as

$$Y[\mathrm{z}] = \frac{P[\mathrm{z}]C[\mathrm{z}]}{1 + P[\mathrm{z}]C[\mathrm{z}]}U_r[\mathrm{z}] + \frac{P[\mathrm{z}]R[\mathrm{z}]}{1 + P[\mathrm{z}]C[\mathrm{z}]}W_q[\mathrm{z}]. \tag{11}$$

The first term on the right-hand side of (11) corresponds to the system output without quantization, while the second term represents the error signal at the system output, which needs to be minimized.

Our design variable is the NTF $R[\mathrm{z}]$ of the $\Delta\Sigma$ modulator, or equivalently, the feedback filter $R[\mathrm{z}] - 1$. Let us express the state space representation of the filter $R[\mathrm{z}] - 1$ as

$$x_r(k+1) = A_r x_r(k) + B_r w_q(k) \tag{12}$$
$$y_r(k) = C_r x_r(k) \tag{13}$$

where $x_r(k) \in \mathbb{R}^{n_r}$ is the state vector, $A_r \in \mathbb{R}^{n_r \times n_r}$, $B_r \in \mathbb{R}^{n_r \times 1}$, and $C_r \in \mathbb{R}^{1 \times n_r}$.

In (11), the transfer function from the quantization error $w_q(k)$ of the static quantizer to the system output $y(k)$ is denoted as $G_{yw_q}[\mathrm{z}]$, given by

$$G_{yw_q}[\mathrm{z}] = \frac{P[\mathrm{z}]R[\mathrm{z}]}{1 + P[\mathrm{z}]C[\mathrm{z}]}. \tag{14}$$

To mitigate the effect of the quantization error $w_q(k)$ on the system output, we impose a constraint on the norm of $G_{yw_q}[\mathrm{z}]$. Specifically, we consider a performance condition:

$$\|G_{yw_q}[\mathrm{z}]\|_\infty < \epsilon \tag{15}$$

where $\|G_{yw_q}[\mathrm{z}]\|_\infty$ is the H_∞ norm of $G_{yw_q}[\mathrm{z}]$ defined as

$$\|G_{yw_q}[\mathrm{z}]\|_\infty = \max_{|\omega| \le \pi} |G_{yw_q}[e^{j\omega}]| \tag{16}$$

and ϵ is a positive parameter. It should be noted that in place of the H_∞ norm, alternative norms, such as the H_2 norm, can be used to achieve similar results as discussed below.

Let us define an augmented state vector from the state vectors $x_c(k)$, $x_p(k)$, and $x_r(k)$ of the controller, the plant, and the NTF as

$$x(k) = \begin{bmatrix} x_c(k) \\ x_p(k) \\ x_r(k) \end{bmatrix}. \tag{17}$$

Then, the state equation is given by

$$x(k+1) = Ax(k) + Bw_q(k) \tag{18}$$

where

$$A = \begin{bmatrix} \tilde{A} & \tilde{B}C_r \\ \mathbf{0} & A_r \end{bmatrix}, \quad B = \begin{bmatrix} \tilde{B} \\ B_r \end{bmatrix} \tag{19}$$

with $\mathbf{0}$ a zero matrix of an appropriate dimension,

$$\tilde{A} = \begin{bmatrix} A_c & -B_cC_p \\ B_pC_c & A_p \end{bmatrix}, \quad \tilde{B} = \begin{bmatrix} \mathbf{0} \\ B_p \end{bmatrix}. \tag{20}$$

From the definition of the augmented state equation, the system output can be expressed as

$$y(k) = C_{yw_q}x(k) \tag{21}$$

with

$$C_{yw_q} = \begin{bmatrix} C_c & \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{22}$$

It is known that (15) is satisfied if and only if there exists a positive definite matrix of $P_1 \succ 0$ that satisfies

$$\begin{bmatrix} A^TP_1A - P_1 + C_{yw_q}^TC_{yw_q} & A^TP_1B \\ B^TP_1A & B^TP_1B - \epsilon I_n \end{bmatrix} \prec 0 \tag{23}$$

Moreover, using the Schur complement, we can transform (23) into

$$\begin{bmatrix} P_1 & AP_1 & B & \mathbf{0} \\ P_1A^T & P_1 & \mathbf{0} & P_1C_{yw_q}^T \\ B^T & \mathbf{0} & I_n & \mathbf{0} \\ \mathbf{0} & C_{yw_q}P_1 & \mathbf{0} & \epsilon \end{bmatrix} \succ 0 \tag{24}$$

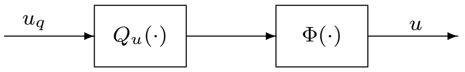


Fig. 6: An equivalent system with the static quantizer $Q(\cdot)$ of the $\Delta\Sigma$ modulator by the saturater $\Phi(\cdot)$.

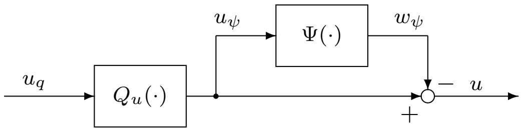


Fig. 7: An equivalent system with the static quantizer $Q(\cdot)$ of the $\Delta\Sigma$ modulator by the dead zone element $\Psi(\cdot)$.

where

$$n = n_c + n_p + n_r. \tag{25}$$

Since there are multiplications of design variables in (24), (24) is a bilinear matrix inequality (BMI).

For simplicity, we assume that the static quantizer $Q(\cdot)$ in Fig. 4 is a mid-tread uniform quantizer whose quantization width is d that satisfies

$$d = \frac{L}{2m+1} \tag{26}$$

for m a positive integer.

Suppose an ideal mid-tread uniform quantizer with a quantization width d that produces

$$Q_u(x) = id, \text{ for } x \in \left[\left(i - \frac{1}{2}\right)d, \left(i + \frac{1}{2}\right)d\right) \tag{27}$$

where x is a scalar input and i is an integer. Then, since $Q(x) = \Phi(Q_u(x))$, the static quantizer $Q(\cdot)$ is equivalent to the cascaded system in Fig. 6.

We introduce a dead zone element $\Psi(\cdot)$ defined as

$$\Psi(x) = x - \Phi(x). \tag{28}$$

Then, the equivalent system can be expressed with the dead zone element $\Psi(\cdot)$ as depicted in Fig. 7. We denote the input and the output of the dead zone element as $u_\psi(k)$ and $w_\psi(k)$, respectievly. Moreover, if we denote the quantization error of $Q_u(\cdot)$ in Fig. 7 as $w_{q_u}(k)$, we have an equivalent system of the original quantizer $Q(\cdot)$ as in Fig. 8, which is known as an additive noise model for the quantization. We replace the quantizer from Fig. 4 with the dead zone module from Fig. 8, as illustrated in Fig. 9.

Fig. 10 shows the input and output relationship of

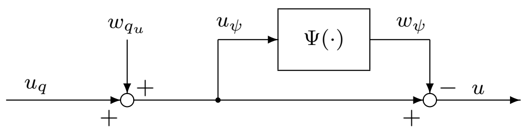


Fig. 8: An additive noise model for the system of Fig. 7.

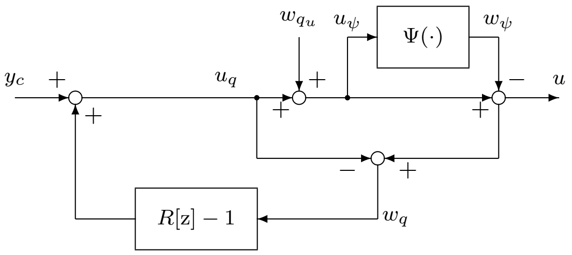


Fig. 9: An additive noise model for the $\Delta\Sigma$ modulator.

the dead zone element. The shaded region corresponds to the sector area enclosed by $y=0$ and $y=\kappa x$ for $0 \leq \kappa \leq 1$. It is evident that $\Psi(x)$ exists in this sector area if and only if the following sector condition is satisfied:

$$\Psi(x)[\Psi(x)-\kappa x] \leq 0. \tag{29}$$

It follows from the small gain theorem that if the saturation error remains the sector area, then the system is l_2-stable [15], [16]. The condition on the saturation error is known as the Tsypkin criterion: Let $G_{u_\psi w_\psi}[\mathrm{z}]$ be the transfer function from $w_\psi(k)$ to $u_\psi(k)$. Then, if

$$(1-\kappa G_{u_\psi w_\psi}(e^{j\theta}))+(1-\kappa G_{u_\psi w_\psi}(e^{j\theta}))^* > 0, \tag{30}$$

for $\theta \in [0, 2\pi)$, then the system is l_2-stable if $w_\psi(k)$ for $\kappa \geq 0$ are in the sector area.

To apply the Tsypkin criterion, we need to obtain a feedback system as shown in Fig. 11, in other words, we need to derive the transfer function from $w_\psi(k)$ to $u_\psi(k)$. To calculate this transfer function, we temporarily set $w_{q_u}(k)=0$, as it does not affect the transfer function. Then, we have

$$u_\psi(k)=y_c(k)+y_r(k)=C_{u_\psi w_\psi}x(k) \tag{31}$$

where

$$C_{u_\psi w_\psi}=\begin{bmatrix} C_c & \mathbf{0} & C_r \end{bmatrix}. \tag{32}$$

One the other hand, $w_q(k)=-w_\psi(k)$. Thus, from (18), the (A,B,C) matrices of $G_{u_\psi w_\psi}[\mathrm{z}]$ are given by $(A,-B)$ where A and B are in (19) and $C_{u_\psi w_\psi}$ is in (32).

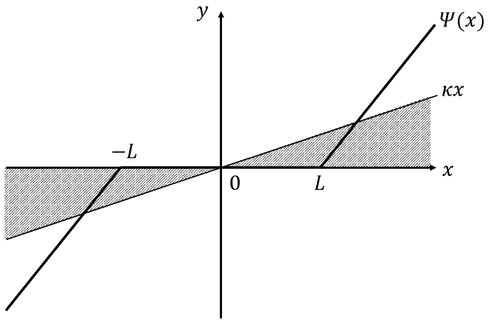


Fig. 10: Dead zone element and sector area.

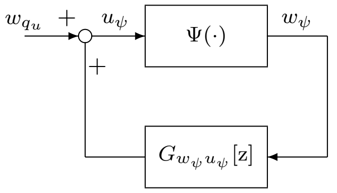


Fig. 11: A feedback connection of a non-linear system $\Psi(\cdot)$ and a linear system $G_{w_\psi u_\psi}$.

Then, using the KYP lemma [17], we can show that the condition (30) is equivalent to the existence of a positive definite matrix $P_2 \succ 0$ that satisfies the following matrix inequality:

$$\begin{bmatrix} -P_2 & B & AP_2 \\ B^T & -2I_n & \kappa C_{u_\psi w_\psi} P_2 \\ P_2 A^T & \kappa P_2 C_{u_\psi w_\psi}^T & -P_2 \end{bmatrix} \prec 0. \tag{33}$$

This matrix inequality is also a BMI.

The value for ϵ should be decreased to reduce the effect of quantization, whereas the value for κ should be increased to enhance stability. In general, there is a trade-off between reducing the effect of quantization and enhancing stability. It should be noted that there is no established criterion for determining the value of κ. A suitablevalue for κ must be found empirically.

Now, let us consider the minimization of ϵ for a given κ. Formally, our problem can be stated as follows:

$$\min_{P_1,P_2,A_r,B_r,C_r} \epsilon \tag{34}$$

subject to (24) and (33). Unfortunately, since the conditions are BMIs, our problem is NP-hard, making it difficult to be solved globally.

One BMI can be converted into a linear matrix inequality (LMI), by using the change of variables [18], [19]. For example, (24) is converted into

$$\begin{bmatrix} M_P & M_A & M_B & \mathbf{0} \\ M_A^T & M_P & \mathbf{0} & M_{C_1}^T \\ M_B^T & \mathbf{0} & I_n & \mathbf{0} \\ \mathbf{0} & M_{C_1} & \mathbf{0} & \epsilon \end{bmatrix} \succ 0. \tag{35}$$

where matrices $\{M_P, M_A, M_B, M_{C_1}\}$ are given by $P_f \succ 0$, $P_g \succ 0$, W_f, W_g and H as follows

$$M_P = \begin{bmatrix} P_f & I_n \\ I_n & P_g \end{bmatrix} \tag{36}$$

$$M_A = \begin{bmatrix} \tilde{A}P_f + \tilde{B}W_f & \tilde{A} \\ H & P_g\tilde{A} \end{bmatrix}, \quad M_B = \begin{bmatrix} \tilde{B} \\ W_g \end{bmatrix} \tag{37}$$

$$M_{C_1} = \begin{bmatrix} [C_c \ \ \mathbf{0}]P_f & [C_c \ \ \mathbf{0}] \end{bmatrix}. \tag{38}$$

However, since P_f and P_g depend on the Lyapunov P_1, the same change of variables can not convert the other BMI (33) into an LMI in general. To address this problem, we adopt the Lyapunov sharing paradigm, as suggested by [18]. By setting the two Lyapunov matrices to be identical, that is, $P_1 = P_2$, we can convert BMI (33) into an LMI given by

$$\begin{bmatrix} -M_P & M_B & M_A \\ M_B^T & -2I_n & \kappa M_{C_2} \\ M_A^T & \kappa M_{C_2}^T & -M_P \end{bmatrix} \prec 0, \tag{39}$$

with M_{C_2} defined as

$$M_{C_2} = \begin{bmatrix} [C_c \ \ \mathbf{0}]P_f & W_f & [C_c \ \ \mathbf{0}] \end{bmatrix}. \tag{40}$$

Finally, our design problem can be stated as

$$\min_{P_f, P_g, W_f, W_g, H} \epsilon \tag{41}$$

subject to (35) and (39), which is a numerically solvable convex optimization. With the optimal $\{P_f, P_g, W_f, W_g, H\}$, the state-space matrices of the optimal feedback filter are given by [18], [19]

$$A_r = [\tilde{B}W_f - P_g^{-1}(H - P_g\tilde{A}P_f)](P_f - P_g^{-1})^{-1} \tag{42}$$

$$B_r = \tilde{B} - P_g^{-1}W_g \tag{43}$$

$$C_r = W_f(P_f - P_g^{-1})^{-1}. \tag{44}$$

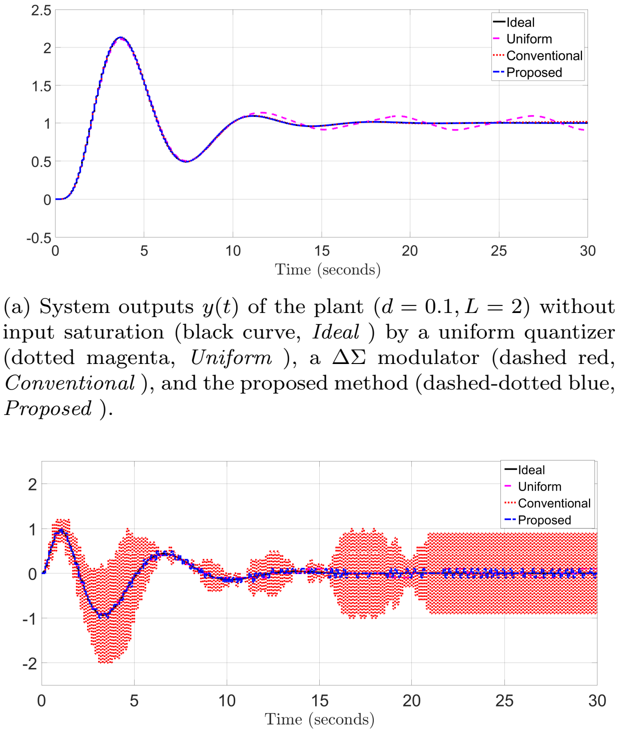
4. Simulation example

To illustrate the performance of our proposed method, we borrow the continuous-time system in [8]. The (A, B, C) matrices of the plant are

$$A_{pc} = \begin{bmatrix} -0.01 & 1 \\ 0 & -0.01 \end{bmatrix}, \quad B_{pc} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C_{pc} = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

and those of the controller are

$$A_{cc} = \begin{bmatrix} 0 & 1.000 & -2.414 & 2.414 \\ -2.414 & -2.414 & -2.000 & 1.000 \\ 1.000 & 0 & -2.414 & 2.414 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$



(a) System outputs $y(t)$ of the plant $(d = 0.1, L = 2)$ without input saturation (black curve, *Ideal*) by a uniform quantizer (dotted magenta, *Uniform*), a $\Delta\Sigma$ modulator (dashed red, *Conventional*), and the proposed method (dashed-dotted blue, *Proposed*).

(b) System inputs $u(t)$ $(d = 0.1, L = 2)$ without input saturation (black curve, *Ideal*) by a uniform quantizer (dotted magenta, *Uniform*), a $\Delta\Sigma$ modulator (dashed red, *Conventional*), and the proposed method (dashed-dotted blue, *Proposed*).

Fig. 12: Comparative analysis of system outputs and inputs using various methods. The outputs are shown in (a), and the inputs are shown in (b).

$$B_{cc} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad C_{cc} = \begin{bmatrix} 2.414 & 2.414 & 1.000 & 0 \end{bmatrix}.$$

The reference signal $u_r(t)$ is set to be a unit step function.

The continuous-time system is discretized using the zero-order hold method and sampled with a period of $T_s = 0.1$ to obtain its equivalent discrete-time model. The coefficient κ, which determines the sector area, is set to be $\kappa = 0.87$.

First, we verify the effectiveness of reducing the influence of quantization errors. By increasing the value of the quantization level L, we relax the conditions of input saturation and perform simulations with a larger quantization step d. We present the results of the proposed method in comparison with results obtained without no constraints on the input saturation as well as those using a uniform quantizer and a Delta-Sigma modulator designed according to the method proposed in [2].

For $L = 2$ and $d = 0.1$, Fig. 12a compares the results without input saturation (labeled as *Ideal*) with those of a uniform quantizer (*Uniform*), a Delta-Sigma modulator (*Conventional*), and the proposed method (*Proposed*). They are plotted using black, dotted magenta, dashed red, and dashed-dotted blue curves, respectively.

In this scenario, with the quantization level set to $L = 2$, both the proposed method and the conventional method yield outputs that closely match the ideal response without causing the windup phenomenon. In contrast, the output from the uniform quantizer exhibits oscillations, indicating its failure to accurately follow the reference signal. This issue arises due to the wide quantization step of $d = 0.1$. Conversely, the outputs using the Delta-Sigma modulator show no significant oscillation for both the proposed and conventional methods.

Fig. 12b shows the input to the system limited to the time range from 0 to 20 corresponding to the outputs of Fig. 12a. The system input with the uniform quantizer and that with our proposed method are almost the same as the ideal system input. On the other hand, the system input with the conventional $\Delta\Sigma$ modulator significantly oscillates around the ideal system input.

Next, we impose a severe condition on the input range with $L = 0.1$. In this case, only two values $-0.1, 0.1$ are available for the system input.

Fig. 13a illustrates the results of the four methods, whereas Fig. 13b depicts the input to the system corresponding to the outputs of Fig. 13a.

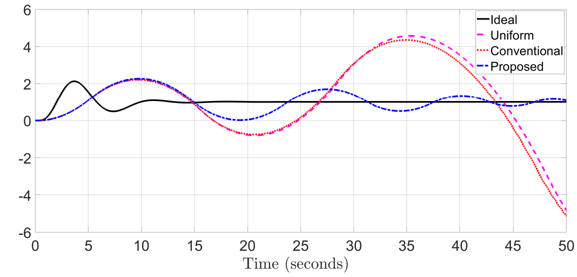
The output of the system with the uniform quantizer diverges, exhibiting typical windup phenomena. The outputs obtained by the two $\Delta\Sigma$ modulators oscillate around the ideal response. The output with the $\Delta\Sigma$ modulator designed by the proposed method converges faster than that with with the conventional $\Delta\Sigma$ modulator, justifying our additional constraint for the input saturation.

5. Conclusions

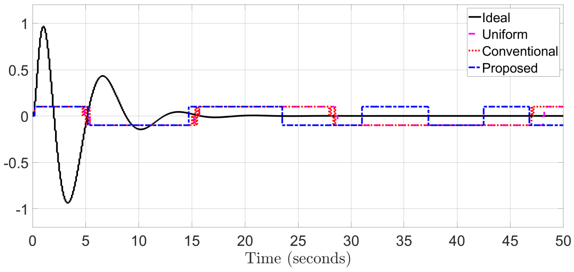
We have introduced a novel design methodology for $\Delta\Sigma$ modulators to mitigate the combined influences of quantization and saturation. Our strategy focuses on minimizing the norm of the quantization error while maintaining stability within the closed-loop system. By leveraging the ability to share a common Lyapunov matrix, we have reformulated our design into a convex optimization problem, enabling numerical solutions. Through numerical examples, we demonstrate the efficacy of our proposed design method.

Acknowledgment

This work was partly supported by JSPS KAKENHI Grant Number 22K04090.



(a) System outputs $y(t)$ of the plant $(d = 0.1, L = 0.1)$ without input saturation (black curve, *Ideal*) by a uniform quantizer (dotted magenta, *Uniform*), a $\Delta\Sigma$ modulator (dashed red, *Conventional*), and the proposed method (dashed-dotted blue, *Proposed*).



(b) System inputs $u(t)$ $(d = 0.1, L = 0.1)$ without input saturation (black curve, *Ideal*) by a uniform quantizer (dotted magenta, *Uniform*), a $\Delta\Sigma$ modulator (dashed red, *Conventional*), and the proposed method (dashed-dotted blue, *Proposed*).

Fig. 13: Comparative analysis of system outputs and inputs using various methods. The outputs are shown in (a), and the inputs are shown in (b).

References

[1] R. Schreier and G. C. Temes, *Understanding Delta-Sigma Data Converters*. Wiley-IEEE Press, 2004.

[2] S. Ohno and M. R. Tariq, “Optimization of noise shaping filter for quantizer with error feedback,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 4, pp. 918–930, April 2017.

[3] S. Ohno, Y. Ishihara, and M. Nagahara, “Min-max design of error feedback quantizers without overloading,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 4, pp. 1395–1405, April 2018.

[4] S. Azuma and T. Sugie, “Synthesis of optimal dynamic quantizers for discrete-valued input control,” *IEEE Transactions on Automatic Control*, vol. 53, no. 9, pp. 2064–2075, Oct. 2008.

[5] K. Sawada and S. Shin, “Dynamic quantizer synthesis based on invariant set analysis for SISO systems with discrete-valued input,” in *the 19th International Symposium on Mathematical Theory of Networks and Systems*, 2010, pp. 1385–1390.

[6] S. Galeani, S. Tarbouriech, M. Turner, and L. Zaccarian, “A tutorial on modern anti-windup design,” *European Journal of Control*, vol. 15, no. 3, pp. 418–440, 2009.

[7] P. Weston and I. Postlethwaite, “Linear conditioning schemes for systems containing saturating actuators,” *IFAC Proceedings Volumes*, vol. 31, no. 17, pp. 675–680, 1998, 4th IFAC Symposium on Nonlinear Control Systems De-

sign 1998 (NOLCOS'98), Enschede, The Netherlands, 1-3 July.

[8] M. Saeki and N. Wada, "Synthesis of a static anti-windup compensator via linear matrix inequalities," *International Journal of Robust and Nonlinear Control*, vol. 12, no. 10, pp. 927–953, 2002.

[9] J. M. G. da Silva and S. Tarbouriech, "Antiwindup design with guaranteed regions of stability: an LMI-based approach," *IEEE Transactions on Automatic Control*, vol. 50, no. 1, pp. 106–111, Jan 2005.

[10] J. G. da Silva and S. Tarbouriech, "Anti-windup design with guaranteed regions of stability for discrete-time linear systems," *Systems & Control Letters*, vol. 55, no. 3, pp. 184–192, 2006.

[11] G. Grimm, J. Hatfield, I. Postlethwaite, A. R. Teel, M. C. Turner, and L. Zaccarian, "Antiwindup for stable linear systems with input saturation: an LMI-based synthesis," *IEEE Transactions on Automatic Control*, vol. 48, no. 9, pp. 1509–1525, Sep. 2003.

[12] G. Grimm, A. R. Teel, and L. Zaccarian, "The l_2 anti-windup problem for discrete-time linear systems: Definition and solutions," *Systems & Control Letters*, vol. 57, no. 4, pp. 356–364, 2008.

[13] A. Matsuya, T. Sato, N. Araki, and Y. Konishi Design of an anti-windup dynamic quantizer for a discrete-valued input control system, Transactions of the Institute of Systems, Control and Information Engineers, vol. 31, no. 2, pp. 58—62, 2018 (in Japanese).

[14] W. M. Haddad and V. Chellaboina, *Nonlinear Dynamical Systems and Control: A Lyapunov-Based Approach.* Princeton University Press, 2008.

[15] Y. Z. Tsypkin, "Fundamentals of the theory of non-linear pluse control systems," *IFAC Proceedings Volumes*, vol. 1, no. 2, pp. 172–180, 1963, 2nd International IFAC Congress on Automatic and Remote Control: Theory, Basle, Switzerland, 1963.

[16] M. Larsen and P. Kokotović, "A brief look at the tsypkin criterion: From analysis to design," *International Journal of Adaptive Control and Signal Processing*, vol. 15, pp. 121–128, 03 2001.

[17] S. Boyd, L. E. Ghaoul, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory.* Society for Industrial and Applied Mathematics, 1997.

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

[19] I. Masubuchi, A. Ohara, and N. Suda, LMI-based controller synthesis: A unified formulation and solution. *International Journal of Robust and Nonlinear Control*, vol. 8, no. 8, 669–686, 1998

Appendix A: A brief derivation of (35) and (39)

Without loss of generality, we can set [19]:

$$P_1 = \begin{bmatrix} P_f & S_f \\ S_f & S_f \end{bmatrix}. \qquad (\mathrm{A}\cdot 1)$$

With P_f and S_f, we define P_g as

$$P_g = (P_f - S_f)^{-1}. \qquad (\mathrm{A}\cdot 2)$$

Using a matrix T given by

$$T = \begin{bmatrix} I_n & P_g \\ \mathbf{0} & -P_g \end{bmatrix}, \qquad (\mathrm{A}\cdot 3)$$

we define a block diagonal matrix $\Theta_1 = \mathrm{diag}(T, T, T, 1)$

After multiplying both side of (24) with the transformation matrix Θ_1 from the right hand side and Θ_1^T from the left hand side, we substitute (42), (43), and (44), and then obtain (35), using (36), (37) and (38).

Similarly, with a block diagonal matrix $\Theta_2 = \mathrm{diag}(T, I_n, T)$, we can show that (33) with $P_2 = P_1$ is transformed into (39).

Shuichi OHNO Shuichi OHNO 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 with Shimane University. From 2002 to 2021, he was an Associate Professor at Hiroshima University. Since 2021, he has been a Professor at Osaka Metropolitan University. His current research interests are in the areas of signal processing. Dr. Ohno is a member of IEICE and IEEE.

Shenjian WANG Shenjian WANG received his B.E. degree in Electronic Information Engineering from Zhejiang University of Technology, Hangzhou, China, in 2018. He earned his M.E. degree in System Cybernetics from Hiroshima University, Hiroshima, Japan, in 2022. Since 2022, he has been a Ph.D. student at Osaka Metropolitan University. His current research interests lie in the area of signal processing. Shenjian Wang is a student member of IEEE.

Kiyotsugu TAKABA Kiyotsugu Takaba received his B.Eng., M.Eng., and Dr.Eng. degrees all from Kyoto University, Japan, in 1989, 1991, and 1995, respectively. From 1991 to 1998, he was an Assistant Professor at the Department of Applied Mathematics and Physics, Kyoto University. From 1998 to 2012, he was an Associate Professor in the same department. In 2012, he joined the Department of Electrical and Electronic Engineering, Ritsumeikan University, where he is currently a Professor. His current research interests include robust/optimal control and optimal filtering for large-scale dynamical systems and their applications. He is a member of SICE, ISCIE, IEEE, and SIAM.