Dynamic Quantization of Nonlinear Control Systems

Shun-ichi Azuma, *Member, IEEE*, and Toshiharu Sugie, *Fellow, IEEE*

***Abstract*— This paper addresses a problem of finding an optimal dynamic quantizer for *nonlinear* control subject to discrete-valued signal constraints, *i.e.*, to the condition that some signals must take a value on a discrete and countable set at each time instant. The quantizers to be studied are in the form of a nonlinear difference equation which maps continuous-valued signals into discrete-valued ones. They are evaluated by a performance index expressing the difference between the resulting quantized system and the unquantized system, in terms of the input-output relation. In this paper, we present a closed-form solution, which globally minimizes the performance index. This result shows the performance limitation of a general class of dynamic quantizers. In addition to this, some results on the structure and the stability are given in order to clarify the mechanism of the best dynamic quantization in nonlinear control systems.**

***Index Terms*— quantized control, dynamic quantizers, nonlinear systems, hybrid systems.**

I. Introduction

QUANTIZED control, *i.e.*, control of systems subject to discrete-valued signal constraints, has become one of the major topics in the systems and control field. The reason lies in its numerous applications, including embedded systems, remote systems, trading systems, and biological systems. In fact, digital devises embedded in them, such as A/D and D/A converters, discrete-level actuators/sensors, and communication channels, are indispensable to make control systems robust, intelligent, and low-cost. Furthermore, it is often the case that the control input is restricted to be one of finite actions, e.g., *sell* or *buy* in trading systems and *activate* or *inhibit* in genetic systems. It is, however, necessary to handle discrete-valued signals as well as continuous-valued ones, which poses challenging control problems.

In this topic, a basic problem is to design the quantizer $Q : \mathbf{U} \to \mathbf{V}$ in such a way that the resulting quantized system (the system including Q) achieves desired performance, where $\mathbf{U}$ and $\mathbf{V}$ are respectively the continuous-valued and discrete-valued signal sets. Various issues arising in quantized control can be reduced into this type of problem with an appropriately selected performance index and quantizer class.

So far, this problem has been studied along two directions: the *networked control* and the *command-driven control*, as summarized in Table I. In the former, the quantizers play a role of the coder-decoder pair in the communication between a plant and a controller. There, the control designer has flexibility in choosing both the map Q and the output signal set $\mathbf{V}$. Several results have been obtained as the minimum data rates for stabilization and estimation [1]–[6] and the (coarsest) quantizers for stabilization and identification [7]–[14]. In the latter, on the other hand, the quantizers are required to adapt continuous-valued signals to the command-driven devices, such as discrete-level actuators, where the quantizer input is assumed to take values on a *fixed* discrete set. So, unlike the former, the map Q is the design parameter and the set $\mathbf{V}$ is a given constraint in the problem. From this standpoint, quantizers have been developed in [15]–[25]. However, the above results have been devoted mainly to linear systems. Namely, except for a few pioneering works, the quantizer design problem has never been studied for nonlinear systems. In fact, in the nonlinear setting, there are some results [26]–[30] for the networked control and *no* result for the command-driven control, as shown in Table I.

This paper thus addresses a quantizer design problem for the command-driven control of a class of nonlinear systems. The quantizers considered here are *dynamic*, *i.e.*, in the form of a nonlinear difference equation which determines its output depending upon the past input sequence. The discrete-valued signal is restricted to take a value on a uniform and countable set at each time instant. The following problem is then considered: when a nonlinear plant and a nonlinear controller are given for the quantized feedback system in Fig. 1 (a), find a quantizer such that the system in (a) *optimally* approximates the usual (unquantized) feedback system in Fig. 1 (b), in terms of the input-output relation. This is a nonlinear version of the authors' quantizer design problem for linear systems [22]–[25], and it is much more challenging.

For the problem, the main contributions of this paper are summarized as follows. First, a *globally optimal solution* is derived as a closed-form expression assuming that the initial state of the system to be quantized is known, even though the problem is nonlinear and nonconvex. The key idea is to analyze the lower and upper bounds of the optimal performance and characterize the dynamic quantizer whose performance is not larger than the lower bound. Second, the structure of the optimal solution is clarified. In particular, it is disclosed that the optimal quantizer is mainly composed of (i) the direct transmission of the input and (ii) an approximated inverse of the error system between the quantized and unquantized systems in Fig. 1. This exhibits the mechanism of the optimal dynamic quantization in nonlinear control systems. Finally, observer-based dynamic quantizers are presented so as to apply our result to the case where the information of the initial state is unavailable. This is provided by fully exploiting the essence of the *optimal* dynamic quantizers.

It is stressed that, although this paper presents a generalization of the result in [23], [25] for linear systems, the solution is

Manuscript submitted April, 2010.

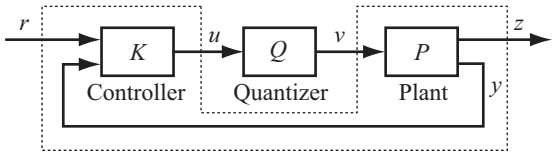
This work was supported by Scientific Research (B) 21360202 and Grant-in-Aid for Young Scientists (B) 21760329 from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

S. Azuma and T. Sugie are with Graduate School of Informatics, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan (e-mail: {sazuma,sugie}@i.kyoto-u.ac.jp).

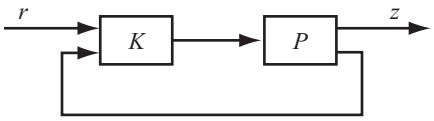
TABLE I

SUMMARY OF DESIGN ISSUES OF QUANTIZER $Q : \mathbf{U} \to \mathbf{V}$

		Networked control	Command-driven control
Purpose of quantization		to transmit signals via digital communication channel	to adapt continuous-valued signals to command-driven devices
Parameters to be designed		Q and $\mathbf{V}$	Q
Existing results	for linear systems	[1]–[14]	[15]–[25]
	for nonlinear systems	[26]–[30]	none (←this paper)



(a) Quantized feedback system.



(b) Unquantized feedback system (Usual system).

Fig. 1. Quantized and unquantized feedback systems.

derived in a different way. In [23], [25], an *exact expression* of the quantizer performance is provided, from which an optimal quantizer is directly derived. In contrast, it is hopeless to obtain such an expression in the nonlinear setting. So, in this paper, we derive an optimal quantizer in an implicit way based on the bound analysis of the optimal performance. Moreover, we have a different result on the optimal structure from the linear case studied in [31].

Also, it should be noted that, to our best knowledge, there is *no* result dealing with both nonlinear systems and (behavioral) performance optimality at the same time, on quantizer design for control. For instance, the main interest of the existing results for nonlinear systems [26]–[30] (see Table I) is the relation to the stability of quantized systems. In this paper, to mathematically clarify an essential mechanism of nonlinear optimal quantization, we mainly consider a somewhat limited case, where the plant and controller are input-affine, the initial states of the systems are available to the quantizer, and the set on which the discrete-valued signal takes a value at each time instant is uniform and countable; but it is remarkable that an exact solution for the unexplored problem is analytically derived. In other words, in the research area of quantized control, this paper provides the first result showing that there exists a nonlinear optimal quantization problem whose solution can be analytically and exactly derived, and the rather restrictive case is regarded as a sufficient condition for the problem to be analytically solved. This will be an important first step to solve the problem for more general situations.

Finally, to avoid misunderstanding, we would like to notify again that the target of this paper is *not* the networked control but the command-driven control, e.g., by discrete-level actuators. This means that typical techniques for networked control may not be applied to our situation. Especially, the zooming/scaling [32], which is a conventional coding technique, cannot be used for the quantizer in Fig. 1 (a), because the zooming/scaling violates the constraint that the quantizer output set $\mathbf{V}$ is *fixed* in advance.

This paper is organized as follows. The quantizer design problem is formulated in Section II. In Section III, a solution, *i.e.*, an optimal quantizer, is presented in an analytical way and is demonstrated by a numerical example. Next, some results on the structure and the stability are given in Section IV. Section V presents observer-based dynamic quantizers and Section VI concludes this paper.

Note that this paper is based on our preliminary version [33], published in a conference proceedings, and contains full explanations and proofs omitted there.

Notation *(i) General mathematical notions:* Let $\mathbf{R}$, $\mathbf{R}_{0+}$, $\mathbf{R}_{+}$, and $\mathbf{N}$ be the real number field, the set of nonnegative real numbers, the set of positive real numbers, and the set of nonnegative integers, respectively. We denote by $0_{n\times m}$ and I_n (or, for simplicity of notation, 0 and I) the $n \times m$ zero matrix and the $n \times n$ identity matrix. Let $\lceil a \rceil$ be the minimum integer greater than or equal to the number $a \in \mathbf{R}$. The vector inequality $x_1 \leq x_2$ represents that each element of $x_1 - x_2$ is nonpositive. For the infinite vector sequences $X := (x_1, x_2, \ldots)$ and $Y := (y_1, y_2, \ldots)$, let $X - Y$ be the vector sequence $(x_1 - y_1, x_2 - y_2, \ldots)$. For the vector x, the matrix M, and the vector sequence X, we use $\|x\|$, $\|M\|$, and $\|X\|$ to express their ∞-norms. Note that $\|M\|$ is the induced norm corresponding to $\max_{x\in\mathbf{R}^n\setminus\{0\}} \|Mx\|/\|x\|$ (where $M \in \mathbf{R}^{m\times n}$), and that $\|X\| := \sup_{i\in\mathbf{N}\setminus\{0\}} \|x_i\|$. When another kind of norm is used or the use of the ∞-norm has to be emphasized, they are denoted with the subscript, *e.g.*, $\|X\|_\rho$ for the ρ-norm. The set of infinite sequences of p-dimensional vectors having finite ∞-norm is denoted by ℓ^p_∞. The function $\psi : \mathbf{R}_{0+} \times \mathbf{R}_{0+} \to \mathbf{R}_{0+}$ is said to be *class-KL* if the following two conditions hold: (a) for each $t \in \mathbf{R}_{0+}$, $\psi(0, t) = 0$ and $\psi(s, t)$ is strictly increasing with respect to s, (b) for each $s \in \mathbf{R}_{0+}$, $\lim_{t\to\infty} \psi(s, t) = 0$ and $\psi(s, t)$ is decreasing with respect to t.

(ii) Notions for dynamical systems: Consider the discrete-time system

$$S : \begin{cases} x(t+1) = f(x(t), u(t)), \\ y(t) \quad = h(x(t), u(t)) \end{cases}$$

where $x(t) \in \mathbf{R}^n$ is the state, $u(t) \in \mathbf{R}^m$ is the input, $y(t) \in \mathbf{R}^p$ is the output, and $f : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n$ and

$h : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^p$ are functions. The system S is said to be *stable* if $(x(1), x(2), \ldots) \in \ell_\infty^n$ holds for every $x(0) \in \mathbf{R}^n$ and $(u(0), u(1), \ldots) \in \ell_\infty^m$. The system S is said to be *output-stable* if $(y(1), y(2), \ldots) \in \ell_\infty^p$ for every $x(0) \in \mathbf{R}^n$ and $(u(0), u(1), \ldots) \in \ell_\infty^m$. These are stability notions based on the boundedness of the state and the output. Note that S is output-stable if h is a continuous function (on its domain) and S is stable. Next, we introduce an equivalence relation between two systems. Consider the systems $S^{(i)}$ $(i = 1, 2, 3)$ given by

$$S^{(i)} : \begin{cases} x^{(i)}(t+1) = f^{(i)}(x^{(i)}(t), u^{(i)}(t)), \\ y^{(i)}(t) \quad = h^{(i)}(x^{(i)}(t), u^{(i)}(t)) \end{cases}$$

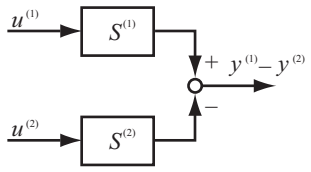
where $x^{(i)}(t) \in \mathbf{R}^{n^{(i)}}$, $u^{(i)}(t) \in \mathbf{R}^{m^{(i)}}$, and $y^{(i)}(t) \in \mathbf{R}^p$. The systems $S^{(1)}$ and $S^{(2)}$ are said to be *equivalent* if the following two conditions hold: (i) $n^{(1)} = n^{(2)} = n$ and $m^{(1)} = m^{(2)} = m$ for some n and m, (ii) $y^{(1)}(t) \equiv y^{(2)}(t)$ for every $x^{(1)}(0) \in \mathbf{R}^n$, $(u^{(1)}(0), u^{(1)}(1), \ldots) \in \ell_\infty^m$, $x^{(2)}(0) \in \mathbf{R}^n$, and $(u^{(2)}(0), u^{(2)}(1), \ldots) \in \ell_\infty^m$ satisfying $x^{(1)}(0) = x^{(2)}(0)$ and $u^{(1)}(t) \equiv u^{(2)}(t)$. The equivalence relation is often denoted by $S^{(1)}(u^{(1)}) = S^{(2)}(u^{(2)})$, which is convenient to express the equivalence between interconnected systems. For instance, when $n^{(1)} + n^{(2)} = n^{(3)}$ and $m^{(1)} + m^{(2)} = m^{(3)}$, the relation $S^{(1)}(u^{(1)}) - S^{(2)}(u^{(2)}) = S^{(3)}(u^{(3)})$ represents the equivalence between the parallel system in Fig. 2 (a) and the system $S^{(3)}$. When $n^{(1)} + n^{(2)} = n^{(3)}$, $u^{(1)}(t)$ is decomposed into $u_1^{(1)}(t) \in \mathbf{R}^{m_1^{(1)}}$ and $u_2^{(1)}(t) \in \mathbf{R}^{m_2^{(1)}}$, *i.e.*, $m_1^{(1)} + m_2^{(1)} = m^{(1)}$ and $u^{(1)}(t) = [(u_1^{(1)}(t))^\top \ (u_2^{(1)}(t))^\top]^\top$, $p = m_2^{(1)}$, and $m_1^{(1)} + m^{(2)} = m^{(3)}$, the relation $S^{(1)}(u_1^{(1)}, S^{(2)}(u^{(2)})) = S^{(3)}(u^{(3)})$ means that the cascade system in Fig. 2 (b) and $S^{(3)}$ are equivalent. Note in the interconnected systems that their state variables are assumed to be $[(x^{(1)}(t))^\top \ (x^{(2)}(t))^\top]^\top$ (not $[(x^{(2)}(t))^\top \ (x^{(1)}(t))^\top]^\top$). In addition, it is worth mentioning that if $S^{(1)}(u^{(1)}) = S^{(2)}(u^{(2)})$ and $S^{(1)}$ is output-stable, then $S^{(2)}$ is output-stable. Finally, an inverse relation is introduced. For the systems $S^{(1)}$ and $S^{(2)}$, assume that $n^{(1)} = n^{(2)} = n$, $m^{(1)} = m^{(2)} = m \geq p$ for some n and m. Let $u_1^{(i)}(t)$ and $u_2^{(i)}(t)$ denote the first $n - p$ elements of $u^{(i)}(t)$ and the others, *i.e.*, $u^{(i)}(t) = [(u_1^{(i)}(t))^\top \ (u_2^{(i)}(t))^\top]^\top \in \mathbf{R}^{m-p} \times \mathbf{R}^p$. The system $S^{(2)}$ is called the *inverse of* $S^{(1)}$ *for the input* $u_2^{(1)}$ if for every $x^{(1)}(0) \in \mathbf{R}^n$ and $x^{(2)}(0) \in \mathbf{R}^n$ satisfying $x^{(1)}(0) = x^{(2)}(0)$ and every $(u^{(2)}(0), u^{(2)}(1), \ldots) \in \ell_\infty^m$, the relation $y^{(1)}(t) \equiv u_2^{(2)}(t)$ holds under $y^{(2)}(t) \equiv u_2^{(1)}(t)$ and $u_1^{(1)}(t) \equiv u_1^{(2)}(t)$ as shown in Fig. 3. The inverse system is denoted by $(S^{(1)})^{-1}_{u_2^{(1)}}$. For example, for the system

$$\begin{cases} x^{(1)}(t+1) = (x^{(1)}(t))^2 + u_1^{(1)}(t) + u_2^{(1)}(t), \\ y^{(1)}(t) \quad = (2 + \sin x^{(1)}(t))^{-1} u_2^{(1)}(t), \end{cases}$$

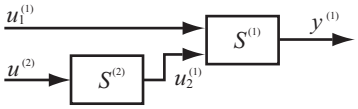
the inverse for the input $u_2^{(1)}$ is given by

$$\begin{cases} x^{(2)}(t+1) = (x^{(2)}(t))^2 + u_1^{(2)}(t) + (2 + \sin x^{(2)}(t)) u_2^{(2)}(t), \\ y^{(2)}(t) \quad = (2 + \sin x^{(2)}(t)) u_2^{(2)}(t). \end{cases}$$

This can be confirmed by the definition and the fact that $x^{(1)}(t) \equiv x^{(2)}(t)$ under $x^{(1)}(0) = x^{(2)}(0)$, $y^{(2)}(t) \equiv u_2^{(1)}(t)$, and $u_1^{(1)}(t) \equiv u_1^{(2)}(t)$.



(a) Parallel system $S^{(1)}(u^{(1)}) - S^{(2)}(u^{(2)})$.



(b) Cascade system $S^{(1)}(u_1^{(1)}, S^{(2)}(u^{(2)}))$.

Fig. 2. Two types of interconnected systems.

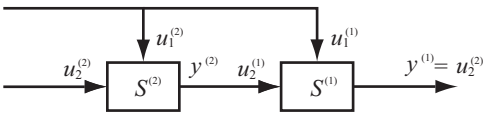


Fig. 3. System $S^{(1)}$ and its inverse for input $u_2^{(1)}$.

II. Problem Formulation

A. System Description

Consider the feedback system Σ_Q shown in Fig. 4 (a), which is composed of the discrete-time nonlinear system G and the dynamic quantizer Q.

The system G is given by

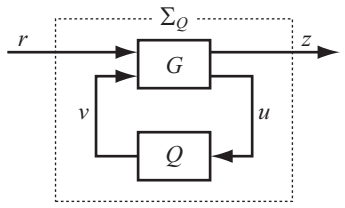
$$G : \begin{cases} x(t+1) = f(x(t)) + g_1(x(t)) r(t) + g_2(x(t)) v(t), \\ z(t) \quad = h_1(x(t)) + k_1(x(t)) r(t), \\ u(t) \quad = h_2(x(t)) + k_2(x(t)) r(t) \end{cases} \tag{1}$$

where $x(t) \in \mathbf{R}^n$ is the state, $r(t) \in \mathbf{R}^p$ and $v(t) \in \mathbf{R}^m$ are the inputs, $z(t) \in \mathbf{R}^l$ and $u(t) \in \mathbf{R}^m$ are the outputs, and $t \in \mathbf{N}$ is the time. Further, $f : \mathbf{R}^n \to \mathbf{R}^n$, $g_1 : \mathbf{R}^n \to \mathbf{R}^{n \times p}$, $g_2 : \mathbf{R}^n \to \mathbf{R}^{n \times m}$, $h_1 : \mathbf{R}^n \to \mathbf{R}^l$, $h_2 : \mathbf{R}^n \to \mathbf{R}^m$, $k_1 : \mathbf{R}^n \to \mathbf{R}^{l \times p}$, and $k_2 : \mathbf{R}^n \to \mathbf{R}^{m \times p}$ are functions. The initial state is given as $x(0) = x_0$ for $x_0 \in \mathbf{R}^n$. In order to show a general formulation of our quantizer design problem first, specific assumptions for the system G will be given at the beginning of the next section, where, for example, $h_1(x) = Cx$ and $k_1(x) = D$ are assumed for constant matrices C and D.

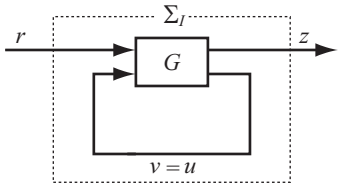
The quantizer Q is of the form

$$Q : \begin{cases} \xi(t+1) = \alpha(\xi(t)) + \beta_1(\xi(t)) u(t) + \beta_2(\xi(t)) v(t), \\ v(t) \quad = q(\gamma(\xi(t)) + \delta(\xi(t)) u(t)) \end{cases} \tag{2}$$

where $\xi(t) \in \mathbf{R}^\nu$, $u(t) \in \mathbf{R}^m$, and $v(t) \in \mathbf{V}^m := \{0, \pm d, \pm 2d, \ldots\}^m$ are the state, the input, and the output, and $\alpha : \mathbf{R}^\nu \to \mathbf{R}^\nu$, $\beta_1, \beta_2 : \mathbf{R}^\nu \to \mathbf{R}^{\nu \times m}$, $\gamma : \mathbf{R}^\nu \to \mathbf{R}^m$, $\delta : \mathbf{R}^\nu \to \mathbf{R}^{m \times m}$ are functions. The set $\mathbf{V}^m$ is a discrete set specified by the quantization interval $d \in \mathbf{R}_+$. The function $q : \mathbf{R}^m \to \mathbf{V}^m$ is the nearest-neighbor static quantizer toward $-\infty$. More precisely, the ith element of the vector $q(\mu)$ is given by $d\lceil(\mu_i/d) - (1/2)\rceil$ where $\mu \in \mathbf{R}^m$ and $\mu_i \in \mathbf{R}$ is the



(a) Quantized feedback system Σ_Q.



(b) Unquantized feedback system Σ_I.

Fig. 4. General expressions of quantized and unquantized feedback systems.

ith element of μ. An example for the case $m := 1$ is shown in Fig. 5. The quantization error of q satisfies

$$\|q(\mu) - \mu\| \leq \frac{d}{2} \qquad (\forall \mu \in \mathbf{R}^m), \tag{3}$$

which will be an important property in this paper. The initial state of Q is given as $\xi(0) = \xi_0 \in \mathbf{R}^\nu$. Note that the quantizer Q determines its output $v(t)$ from the past and present inputs $(u(0), u(1), \ldots, u(t))$ (so Q is dynamic), and Q is equivalent to the typical static quantizer $v(t) = q(u(t))$ if $\gamma(\xi(t)) \equiv 0$ and $\delta(\xi(t)) \equiv I$. Note also that Q includes the self feedback by v as seen in the state equation of (2). In what follows, Q is often regarded as a tuple of the dimension ν, the initial state ξ_0, and the functions α, β_1, β_2, γ, δ, which will be treated as the design parameters.

The system Σ_Q is a generalized version of the quantized feedback system in Fig. 1 (a). It can be seen that Σ_Q is equivalent to the system in Fig. 1 (a) by regarding the part indicated by the dotted line frame (in Fig. 1 (a)) as G. Thus the following discussion holds not only for the feedback system in Fig. 1 (a) but also for various types of quantized systems.

B. Dynamic Quantizer Design Problem

In this paper, the quantizer Q is evaluated by a performance index expressing the difference between the quantized system Σ_Q and the *unquantized* system Σ_I (introduced as an *I*deal system) in Fig. 4. For this, some symbols are prepared. To distinguish the signals of the two systems Σ_Q and Σ_I, we use the symbols x_Q, r_Q, v_Q, z_Q, u_Q and x_I, r_I, v_I, z_I, u_I for x, r, v, z, u. In Σ_Q (with a given Q), when the initial state and the external input are fixed to the specific values $x_0 \in \mathbf{R}^n$ and $R := (r_0, r_1, \ldots) \in \ell^p_\infty$, we denote by $Z_Q(x_0, R)$ the output sequence $(z_Q(1), z_Q(2), \ldots)$ and by $z_Q(t, x_0, R)$ the output at time t. For Σ_I, the symbols $Z_I(x_0, R)$ and $z_I(t, x_0, R)$ are similarly defined. Note here that, though it may look that $Z_Q(x_0, R)$ and $z_Q(t, x_0, R)$ do not depend on the initial state ξ_0 of Q, their subscripts Q correspond to the all design parameters $(\nu, \xi_0, \alpha, \beta_1, \beta_2, \gamma, \delta)$ (including ξ_0) and so the dependency on ξ_0 is expressed at $Z_Q(x_0, R)$ and $z_Q(t, x_0, R)$ in a proper fashion.

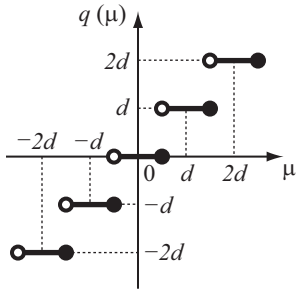


Fig. 5. Static quantizer $q(\mu)$ for $\mu \in \mathbf{R}^1$.

Then the following problem is considered.

Problem 1: For the system Σ_Q, suppose that the system G and the quantization interval $d \in \mathbf{R}_+$ are given. Then, find a quantizer Q (that is, a dimension ν, an initial state ξ_0, and functions α, β_1, β_2, γ, δ) minimizing the performance index

$$E(Q) := \sup_{(x_0, R) \in \mathbf{R}^n \times \ell^p_\infty} \|Z_Q(x_0, R) - Z_I(x_0, R)\|. \tag{4}$$

■

In this problem, the performance index $E(Q)$ corresponds to the difference between the quantized and unquantized systems Σ_Q and Σ_I in terms of the input-output relation. If $E(Q)$ is small, we can conclude that the system Σ_Q behaves similarly to the ideal system Σ_I.

Solving the problem provides us a practical design method of nonlinear control systems with discrete-valued signal constraints. For example, consider the feedback system in Fig. 1 (a), and suppose that the input of P is restricted to be a discrete-valued signal on $\mathbf{V}^m$. Then, in spite of the severe restriction, Σ_Q would have good performance with

- a controller K achieving desirable performance in the unquantized system in Fig. 1 (b) (where it is supposed that the input of P is continuous-valued),
- a dynamic quantizer Q such that $E(Q)$ is small.

Therefore, the combination of the conventional (nonlinear) control theory and the solution to Problem 1 enables us to construct high-performance quantized systems.

Finally, four remarks on Problem 1 are given. First, the system model G can represent a combination of an input-affine plant model and input-affine controller in discrete-time. Plants described in G include the cart-spring-damper system in [34] and the stirred tank reactor system in [35] for example. The discrete-time models in the literatures are provided by the Euler approximation of the continuous-time models. Second, the quantizer output set $\mathbf{V}^m$ is a uniform lattice in $\mathbf{R}^m$, which fits quantized control problems with D/A converters or discrete-level actuators. Even if the quantization intervals are different for each input channels as $d_i \in \mathbf{R}_+$ $(i = 1, 2, \ldots, m)$, the following discussion holds by replacing the input matrix $g_2(x)$ with the scaled matrix $g_2(x)\,\mathrm{diag}\{1, d_2/d_1, d_3/d_1, \ldots, d_m/d_1\}$ [24]. Third, Σ_I is an ideal system and so Σ_I should be a stable system in common situations. Meanwhile, the stability of Σ_I is not assumed in Problem 1 because, with or without the assumption, there

exists a solution Q to Problem 1 such that $E(Q) < \infty$ under a condition not implying the stability of Σ_I. This fact will be shown in Theorem 1. Fourth, in our setting, the performance is evaluated based on not the 1- or 2-norm but the ∞-norm. This is because the signal v is restricted to be a value on the *uniform lattice* $\mathbf{V}^m$ and thus the asymptotic stability of Σ_Q is not always possible, e.g., in the case where P is unstable in Fig. 1 (a). More concretely, when Σ_I is (globally) asymptotically stable and Σ_Q cannot be asymptotically stable with any Q, we have $\lim_{t\to\infty} z_Q(t, x_0, R) \neq z_I(t, x_0, R)$, *i.e.*, $\|Z_Q(x_0, R) - Z_I(x_0, R)\|_\rho = \infty$ for $\rho = 1, 2$, under an observability condition (note that $Z_Q(x_0, R)$ and $Z_I(x_0, R)$ are infinite sequences). In contrast, we may have $\|Z_Q(x_0, R) - Z_I(x_0, R)\|_\infty < \infty$ in the same situation, which means that the index based on the ∞-norm can capture the performance of Q more precisely.

III. Optimal Dynamic Quantizers

A. Assumptions and Outline of Derivation

In this paper, we aim at obtaining an analytical solution to Problem 1 in order to clarify an essential mechanism of optimal quantization. To this end, the problem is considered under the following assumptions:

- (A1) $h_1(x) = Cx$ and $k_1(x) = D$ for constant matrices $C \in \mathbf{R}^{l\times n}$ and $D \in \mathbf{R}^{l\times p}$.
- (A2) The matrix $k_2(x)$ is square and nonsingular for every $x \in \mathbf{R}^n$.
- (A3) The matrix $Cg_2(x)$ is square and nonsingular for every $x \in \mathbf{R}^n$, where C is given in (A1).

The first assumption means that the controlled output z is given as a linear combination of x and r, *i.e.*, $z(t) = Cx(t) + Dr(t)$. The others are technical assumptions for the existence of the inverses of $k_2(x)$ and $Cg_2(x)$, which also imply that r, v, z, and u have all the same dimensions. Roughly speaking, in the feedback system in Fig. 1 (a), these two are usually satisfied in the case where v and z have all the same dimensions and r is directly transmitted to u in K (though it depends on how to get the discrete-time model G).

The idea to find the solution is outlined as follows:

$$\overbrace{\phi(d) \le \underbrace{\min_Q E(Q)}_{\text{(Step 2)}} \le \phi(d).}^{\text{(Step 1)}} \tag{5}$$

In Step 1, we derive a lower bound of $\min_Q E(Q)$, which is a function of d and is denoted by $\phi(d)$. In Step 2, it is shown that, if a condition called (Ω) is satisfied, the lower bound $\phi(d)$ becomes an upper bound of $\min_Q E(Q)$. These steps prove that $\min_Q E(Q) = \phi(d)$ holds under (Ω), from which a solution to Problem 1 is provided.

B. Lower and Upper Bound Analysis of Optimal Performance

1) Step 1: Lower bound: For the system Σ_Q, suppose that Q is given. By the definition of $E(Q)$, we have

$$\sup_{(x_0,R)\in\mathbf{R}^n\times\ell^p_\infty} \|z_Q(1, x_0, R) - z_I(1, x_0, R)\| \le E(Q). \tag{6}$$

From (1), (2), and (A1), the term $z_Q(1, x_0, R) - z_I(1, x_0, R)$ is expressed as

$$\begin{aligned} &z_Q(1, x_0, R) - z_I(1, x_0, R) \\ &= Cg_2(x_0)(v_Q(0) - v_I(0)) \\ &= Cg_2(x_0)(q(a(x_0, \xi_0, r_0)) - a(x_0, \xi_0, r_0)) + Cg_2(x_0)\gamma(\xi_0) \\ &\quad + Cg_2(x_0)(\delta(\xi_0) - I)(h_2(x_0) + k_2(x_0)r_0) \end{aligned} \tag{7}$$

for

$$a(x_0, \xi_0, r_0) := \gamma(\xi_0) + \delta(\xi_0)(h_2(x_0) + k_2(x_0)r_0). \tag{8}$$

In fact, (1) and (2) give $v_Q(0) = q(\,\gamma(\xi_0) + \delta(\xi_0)\,u_Q(0)\,) = q(\gamma(\xi_0) + \delta(\xi_0)(h_2(x_0) + k_2(x_0)r_0))$ and $v_I(0) = u_I(0) = h_2(x_0) + k_2(x_0)r_0$, from which (7) is confirmed. Note that, for the first term of (7), we have

$$\|Cg_2(x_0)(q(a(x_0, \xi_0, r_0)) - a(x_0, \xi_0, r_0))\| \le \|Cg_2(x_0)\|\frac{d}{2}$$

from (3). It can be shown by (7) that

$$\sup_{(x_0,R)\in\mathbf{R}^n\times\ell^p_\infty} \|z_Q(1, x_0, R) - z_I(1, x_0, R)\| \begin{cases} \ge \sup\limits_{x_0\in\mathbf{R}^n} \|Cg_2(x_0)\|\dfrac{d}{2} & \text{if } \delta(\xi_0) = I, \\ = \infty & \text{otherwise} \end{cases} \tag{9}$$

holds under (A1)–(A3) (see Appendix I for the exact proof of (9)). Equations (6) and (9), which hold for any Q, establish a lower bound of $\min_Q E(Q)$ as

$$\sup_{x\in\mathbf{R}^n} \|Cg_2(x)\|\frac{d}{2} \le \min_Q E(Q). \tag{10}$$

This completes Step 1 for $\phi(d) := \sup_{x\in\mathbf{R}^n} \|Cg_2(x)\|(d/2)$.

2) Step 2: Upper bound: Next, we show that the lower bound is an upper bound of $\min_Q E(Q)$ under a suitable condition.

Let

$$f_{cl}(x) := f(x) + g_2(x)h_2(x), \tag{11}$$

$$g_{cl}(x) := g_1(x) + g_2(x)k_2(x). \tag{12}$$

Under (A1) and $r_Q(t) \equiv r_I(t) \equiv r(t)$ for some $r(t)$, the output difference between Σ_Q and Σ_I is described by

$$\begin{aligned} &z_Q(t+1) - z_I(t+1) \\ &= Cx_Q(t+1) + D\,r(t+1) - Cx_I(t+1) - D\,r(t+1) \\ &= C\big(f(x_Q(t)) + g_1(x_Q(t))r(t) + g_2(x_Q(t))v(t)\big) \\ &\quad - C\big(f_{cl}(x_I(t)) + g_{cl}(x_I(t))r(t)\big). \end{aligned} \tag{13}$$

If

$$v(t) = q\big(\sigma(x_Q(t), x_I(t), r(t))\big) \tag{14}$$

for

$$\sigma(x_Q, x_I, r) := -(Cg_2(x_Q))^{-1}[C \;\; -C]\begin{bmatrix} f(x_Q) + g_1(x_Q)r \\ f_{cl}(x_I) + g_{cl}(x_I)r \end{bmatrix} \tag{15}$$

(where $(Cg_2(x_Q))^{-1}$ is given under (A3)), then

$$z_Q(t+1) - z_I(t+1) = Cg_2(x_Q(t))(q(\sigma(t)) - \sigma(t))$$

and (3) yield

$$\|z_Q(t+1) - z_I(t+1)\| \le \|Cg_2(x_Q(t))\|\ \|q(\sigma(t)) - \sigma(t)\| \le \sup_{x\in\mathbf{R}^n} \|Cg_2(x)\|\ \frac{d}{2} \quad (\forall t \in \mathbf{N}), \tag{16}$$

where $\sigma(t)$ stands for $\sigma(x_Q(t), x_I(t), r(t))$. That is,

$$\|z_Q(t+1, x_0, R) - z_I(t+1, x_0, R)\| \le \sup_{x\in\mathbf{R}^n} \|Cg_2(x)\|\ \frac{d}{2}$$

holds for every $t \in \mathbf{N}$ and $(x_0, R) \in \mathbf{R}^n \times \ell_\infty^p$. Therefore, if the condition

(Ω) there exists a Q satisfying (14)

holds, we have

$$\min_Q E(Q) \le \sup_{x\in\mathbf{R}^n} \|Cg_2(x)\|\frac{d}{2}. \tag{17}$$

This achieves Step 2 in (5).

C. Optimal Dynamic Quantizers

Equations (10) and (17) establish the relation

$$\min_Q E(Q) = \sup_{x\in\mathbf{R}^n} \|Cg_2(x)\|\frac{d}{2} \tag{18}$$

subject to the condition (Ω), which presents the following result.

Theorem 1: For the system Σ_Q, suppose that G and d are given and assume (A1)–(A3). Then the following statements hold.
(i) A solution to Problem 1 is given by

$$Q^* := (\nu^*, \xi_0^*, \alpha^*, \beta_1^*, \beta_2^*, \gamma^*, \delta^*) \tag{19}$$

where

$$\nu^* := 2n, \tag{20}$$

$$\xi_0^* := \begin{bmatrix} x_0 \\ x_0 \end{bmatrix}, \tag{21}$$

$$\alpha^*(\xi(t)) := \begin{bmatrix} f(\xi_1(t)) - g_1(\xi_1(t))\, k_2^{-1}(\xi_1(t))\, h_2(\xi_1(t)) \\ f_{cl}(\xi_2(t)) - g_{cl}(\xi_2(t))\, k_2^{-1}(\xi_1(t))\, h_2(\xi_1(t)) \end{bmatrix}, \tag{22}$$

$$\beta_1^*(\xi(t)) := \begin{bmatrix} g_1(\xi_1(t)) k_2^{-1}(\xi_1(t)) \\ g_{cl}(\xi_2(t)) k_2^{-1}(\xi_1(t)) \end{bmatrix}, \tag{23}$$

$$\beta_2^*(\xi(t)) := \begin{bmatrix} g_2(\xi_1(t)) \\ 0_{n\times m} \end{bmatrix}, \tag{24}$$

$$\gamma^*(\xi(t)) := -(Cg_2(\xi_1(t)))^{-1}[C \quad -C]\alpha^*(\xi(t)), \tag{25}$$

$$\delta^*(\xi(t)) := -(Cg_2(\xi_1(t)))^{-1}[C \quad -C]\beta_1^*(\xi(t)), \tag{26}$$

$\xi_1, \xi_2 \in \mathbf{R}^n$ are the first half and the second half of the vector ξ, *i.e.*, $\xi = [\xi_1^\top \ \ \xi_2^\top]^\top$, and f_{cl} and g_{cl} are the functions given in (11) and (12).
(ii) The minimum value of $E(Q)$ is given by

$$E(Q^*) = \sup_{x\in\mathbf{R}^n} \|Cg_2(x)\|\frac{d}{2}. \tag{27}$$

■

Proof: We prove that (14) holds for $Q := Q^*$, since this fact implies that Q^* is a solution to Problem 1 and (27) holds. Under (A2), the third equation of (1) is rewritten as

$$r(t) = k_2^{-1}(x(t))(u(t) - h_2(x(t))). \tag{28}$$

From this and the first equation of (1), it follows that x_Q evolves according to

$$\begin{cases} x_Q(t+1) = f(x_Q(t)) - g_1(x_Q(t))\, k_2^{-1}(x_Q(t))\, h_2(x_Q(t)) \\ \qquad\qquad + g_1(x_Q(t))\, k_2^{-1}(x_Q(t))u(t) + g_2(x_Q(t))v(t), \\ x_Q(0) = x_0. \end{cases}$$

By comparing this equation with the dynamics of ξ_1 in Q^*, it turns out that

$$\xi_1(t) = x_Q(t) \quad (\forall t \in \mathbf{N}) \tag{29}$$

holds for Q^*. In a similar way to the above, we also get

$$\xi_2(t) = x_I(t) \quad (\forall t \in \mathbf{N}). \tag{30}$$

Thus, applying (15), (19), (22), (23), (25), (26), (28), (29), and (30) to the output equation in (2), we have $v(t) = q(\gamma^*(\xi(t)) + \delta^*(\xi(t))u(t)) = q(\sigma(\xi_1(t), \xi_2(t), r(t))) = q(\sigma(x_Q(t), x_I(t), r(t)))$, which implies that (14) holds for Q^*. ■

Theorem 1 provides an analytical solution to Problem 1 (which is globally optimal) and an expression of the minimum value of $E(Q)$. The latter corresponds to the performance limitation of the dynamic quantizers in the form of (2), which shows the relation between the achievable performance and the problem parameters G and d.

An intuitive interpretation of the optimal quantizer Q^* is as follows. As shown in (29) and (30), the states of Σ_Q and Σ_I are estimated in the state equation of Q^*. They are in general different due to the quantization by Q. Considering that $[C \quad -C]\xi(t)$ is equal to the output difference between the two systems, we see that the term $[C \quad -C]\alpha^*(\xi(t)) + [C \quad -C]\beta_1^*(\xi(t))u(t)$ expresses the difference expected at the next time (time $t+1$). Then if the multiplication by $-(Cg_2(\xi_1(t)))^{-1}$, *i.e.*, $\gamma^*(\xi(t)) + \delta^*(\xi(t))u(t)$, is applied to Σ_Q, the signal completely cancels out the expected difference in Σ_Q. Namely, the quantizer output $v(t) = q(\gamma^*(\xi(t)) + \delta^*(\xi(t))u(t))$ is the optimal discrete-valued signal to reduce the difference between Σ_Q and Σ_I.

It should be noted that the exact information of the initial state x_0 of G is required to construct the optimal quantizer Q^*, as seen in (21). So it can be directly applied only to systems whose state is measurable and available to Q or to systems which operates from a fixed initial state such as robot manipulators for a repetitive work. An extended version of Q^*, which do not use the information of x_0, will be provided in Section V.

It is also notified that the optimally quantized system Σ_{Q^*}, Σ_Q with $Q := Q^*$, is not always stable in the stability concept defined in Section I. However, it can be shown that Σ_{Q^*} is stable under a suitable condition, and even when the condition does not hold, there is a practical method to avoid instability. This will be detailed in Section IV.

Remark 1: Theorem 1 is a generalized version of the authors' previous result [23], [25] for linear G, which has

$f(x) := Ax$, $g_1(x) := B_1$, $g_2(x) := B_2$, $h_1(x) := C_1x$, $h_2(x) := C_2x$, $k_1(x) := D_1$, $k_2(x) := D_2$. In fact, by substituting the linear functions and constant matrices and eliminating redundant states in Q^*, we have a solution to Problem 1 as $(\nu, \xi_0, \alpha, \beta_1, \beta_2, \gamma, \delta) = (n, 0, (A + B_2C_2)\xi, -B_2, B_2, -(C_1B_2)^{-1}C_1(A+B_2C_2)\xi, I)$, which is the same as given in [23], [25]. ■

Remark 2: Since the optimal quantizer Q^* depends on x_0, one may consider that the optimal performance $E(Q^*)$ must depend on x_0. However, as seen in (4), the argument of the function $E(Q)$, *i.e.*, $(\nu, \xi_0, \alpha, \beta_1, \beta_2, \gamma, \delta)$, specifies the *function* to be maximized with respect to $(x_0, R) \in \mathbf{R}^n \times \ell^p_\infty$, that is, $\|Z_Q(x_0, R) - Z_I(x_0, R)\|$. Thus, even if Q^* depends on x_0, $E(Q^*)$ does not depend on x_0. ■

Remark 3: The solution in Theorem 1 is derived by fully utilizing the fact that $\mathbf{V}^m$ is a uniform discrete set. However, even when $\mathbf{V}^m$ is not uniform, a similar result can be obtained as long as $\|q(\mu) - \mu\| \le \Delta$ $(\forall \mu \in \mathbf{R}^m)$ for some $\Delta \in \mathbf{R}_{0+}$. In fact, it is trivial in this case that, instead of (16), $\|z_Q(t + 1) - z_I(t + 1)\| \le \sup_{x\in\mathbf{R}^n} \|Cg_2(x)\|\Delta$ $(\forall t \in \mathbf{N})$ and so

$$E(Q^*) \le \sup_{x\in\mathbf{R}^n} \|Cg_2(x)\|\Delta \tag{31}$$

for the proposed quantizer Q^* in (19). Although Q^* may not be optimal in this case, it will be a practical quantizer in the sense of (31). ■

D. Example

Consider the quantized system Σ_Q for the feedback system in Fig. 1 (a). The plant P and the controller K are given by

$$P: \begin{cases} \begin{bmatrix} x_1(t+1) \\ x_2(t+1) \end{bmatrix} = \begin{bmatrix} 1.0x_1(t)+0.1x_2(t)+0.4e^{-|x_2(t)|}\cos^3 x_1(t) \\ 0.2x_1(t)+1.1x_2(t)+0.4e^{-|x_1(t)|}\sqrt{|\cos x_1(t)|} \end{bmatrix} \\ \qquad\qquad + \begin{bmatrix} \sigma(x_1(t)) \\ \sigma(x_1(t)) \end{bmatrix} v(t), \\ z(t) = 1.45x_1(t) + x_2(t), \\ y(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}, \end{cases}$$

$$K: \quad u(t) = -\frac{1}{\sigma(x_1(t))}[0.2 \quad 0.5]y(t) + r(t)$$

for $x_1, x_2 \in \mathbf{R}$ and

$$\sigma(x_1(t)) := 0.01\left(1 + \frac{1}{(x_1(t))^4 + 0.1}\right).$$

The quantizer Q is of (19) with the quantization interval $d := 2$.

Fig. 6 shows the simulation result on the time responses of Σ_Q for $x_0 := [0.1 \quad -0.2]^\top$ and $r(t) \equiv 0$. In the third figure, the output response of the unquantized system Σ_I in Fig. 4 (b) (Fig. 1 (b)) is also depicted by the thin line, where the same condition is imposed. Though the coarse discrete-valued signal is applied to Σ_Q, the output behavior of Σ_Q is quite similar to that of Σ_I. This result is quantified as $E(Q^*) = 0.2695$ by (27) (for the worst (x_0, R)) and $\max_{t\in\{0,1,\ldots,75\}} |z_Q(t) - z_I(t)| = 0.2565$ by the simulation (for the given (x_0, R)).

For comparison, we also consider the static quantizer case $Q = q$, *i.e.*, the case of $\gamma(\xi(t)) \equiv 0$ and $\delta(\xi(t)) \equiv I$. Fig. 7 illustrates the responses in the same fashion. We see that the result for the static quantizer case is different from that for the dynamic quantizer case.

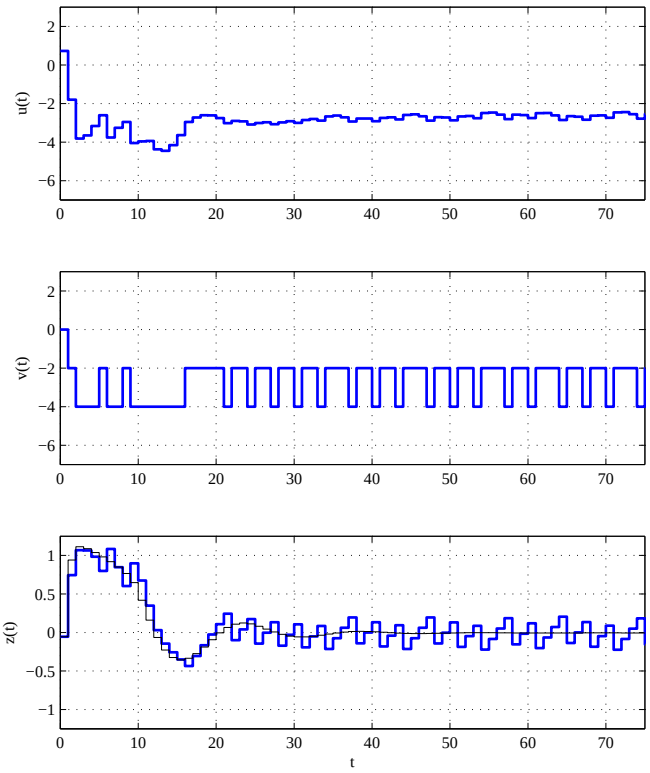


Fig. 6. Responses of Σ_{Q^*} (thick lines) and output response of Σ_I (thin line).

This example shows that, even if the control input is restricted to be a coarse signal, high performance, which cannot be achieved by the static quantizer, is obtained by the optimal dynamic quantizer.

It should be remarked that the above system is an academic example selected to show our result more clearly. As stated in Theorem 1, a similar result can be obtained for any systems satisfying (A1)–(A3).

IV. Structural Analysis of Optimal Dynamic Quantizers

In this section, we analyze the structure of the optimal quantizer Q^* in order to understand the mechanism. Based on this, a stability condition for the optimally quantized system Σ_{Q^*} is provided.

A. Structure of Optimal Dynamic Quantizers

Consider the quantizer Q in (2). To express the static quantization error, *i.e.*, produced by the static quantizer q, we introduce the new variable

$$w(t) := q(\,\gamma(\xi(t))+\delta(\xi(t))u(t)\,) - (\,\gamma(\xi(t))+\delta(\xi(t))u(t)\,), \tag{32}$$

which satisfies

$$\|w(t)\| \le \frac{d}{2} \quad (\forall t \in \mathbf{N}) \tag{33}$$

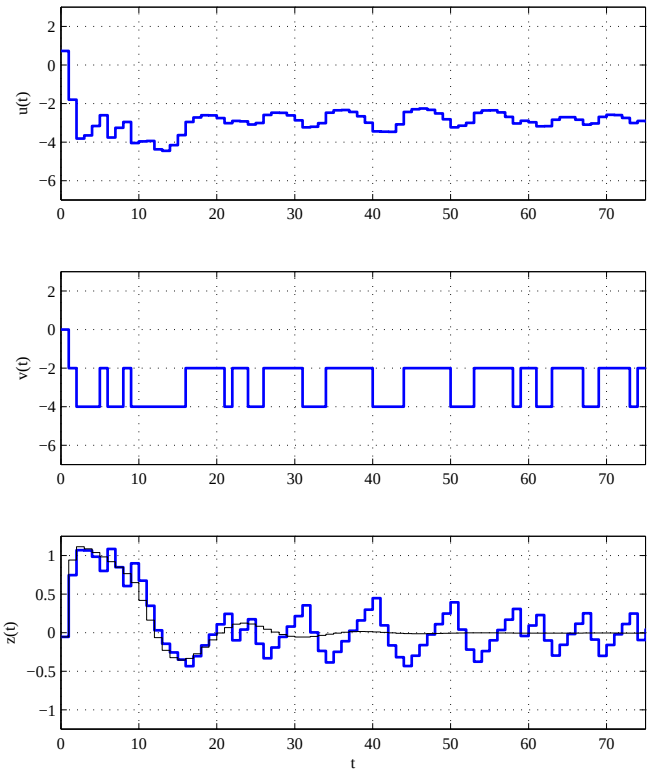


Fig. 7. Responses of Σ_Q with $Q := q$ (thick lines) and output response of Σ_I (thin line).

from (3). With this variable, Q is equivalently represented as

$$Q : \begin{cases} \xi(t+1) = \alpha(\xi(t)) + \beta_2(\xi(t))\gamma(\xi(t)) \\ \qquad\qquad + (\beta_1(\xi(t)) + \beta_2(\xi(t))\delta(\xi(t)))u(t) \\ \qquad\qquad + \beta_2(\xi(t))w(t), \\ v(t) \quad = \gamma(\xi(t)) + \delta(\xi(t)))u(t) + w(t). \end{cases} \tag{34}$$

If (29) is assumed and the third equation of (1) is applied to (34), we have

$$Q : \begin{cases} \xi(t+1) = \tilde{\alpha}(\xi(t)) + \beta_2(\xi(t))w(t) + \beta_3(\xi(t))r(t), \\ v(t) \quad = \tilde{\gamma}(\xi(t)) + u(t) + w(t) + \delta_3(\xi(t))r(t) \end{cases}$$

where

$$\begin{aligned} &\tilde{\alpha}(\xi) := \alpha(\xi) + \beta_1(\xi)h_2(\xi_1) + \beta_2(\xi)(\gamma(\xi) + \delta(\xi)h_2(\xi_1)), \\ &\beta_3(\xi) := (\beta_1(\xi) + \beta_2(\xi)\delta(\xi))k_2(\xi_1), \\ &\tilde{\gamma}(\xi) := \gamma(\xi) + (\delta(\xi) - I)h_2(\xi_1), \\ &\delta_3(\xi) := (\delta(\xi) - I)k_2(\xi_1). \end{aligned}$$

So, under the condition (29), Q can be formally regarded as a nonlinear system driven by its original input u, the static quantization error w, and the external input r for G, though these are not independent each other. Then Q is expressed as Fig. 8[1] by defining the subsystem

$$H : \begin{cases} \xi(t+1) = \tilde{\alpha}(\xi(t)) + \beta_2(\xi(t))w(t) + \beta_3(\xi(t))r(t), \\ s(t) \quad = \tilde{\gamma}(\xi(t)) + w(t) + \delta_3(\xi(t))r(t). \end{cases} \tag{35}$$

[1]In Figs. 8 and 9, w is not purely exogenous as shown in (32) but the dependency on the other signals is omitted in the figure, because we will consider the signal transfer from w to e and it would be helpful for us to regard w as a virtual exogenous signal in order to understand the following discussion.

Furthermore, the error system between Σ_Q and Σ_I is illustrated as Fig. 9. Based on these expressions, the following result is obtained for the structure of Q^*.

Theorem 2: For the system Σ_Q, suppose that G and d are given and assume (A1)–(A3). Let Σ denote the system in Fig. 10, which is a subsystem of the error system in Fig. 9. Then the optimal quantizer Q^* is composed of
(a) the direct transmission of u,
(b) a system H such that

$$H(r, w) = (\mathbf{z}\Sigma)_s^{-1}(r, Cg_2(x_Q)w) \tag{36}$$

where $\mathbf{z}\Sigma$ is the time-shifted system of Σ,
(c) the initial state $\xi(0) = [x_0^\top \;\; x_0^\top]^\top$.
Note in (36) that $(\mathbf{z}\Sigma)_s^{-1}$ is the inverse of $\mathbf{z}\Sigma$ for the input s, $Cg_2(x_Q)$ is the time-varying gain (because of $x_Q(t)$), and the right-hand side is the system $(\mathbf{z}\Sigma)_s^{-1}$ whose inputs are r and $Cg_2(x_Q)w$. Note also that the equivalence and the inverse are introduced in Section I, and the notions are not restricted to the structured initial state in (c).

Proof: Fig. 8 shows that Q is equivalent to the sum of the direct transmission of u and the system H, which implies (a). Furthermore, (c) corresponds to (21). So, in what follows, we prove (b) by showing that the system H for $(\alpha, \beta_1, \beta_2, \gamma, \delta) := (\alpha^*, \beta_1^*, \beta_2^*, \gamma^*, \delta^*)$ satisfies (36). Consider the system Σ in Fig. 10 and suppose that $x_Q(0), x_I(0) \in \mathbf{R}^n$ are given. Note that we do *not* restrict the case $x_Q(0) = x_I(0)$. From Fig. 10, (A1), (1), (11), and (12), the output $e(t+1)$ is represented as

$$\begin{aligned} e(t+1) &= z_Q(t+1) - z_I(t+1) \\ &= Cx_Q(t+1) + Dr(t+1) - Cx_I(t+1) - Dr(t+1) \\ &= [C \;\; -C]\begin{bmatrix} f_{cl}(x_Q(t)) + g_{cl}(x_Q(t))r(t) \\ f_{cl}(x_I(t)) + g_{cl}(x_I(t))r(t) \end{bmatrix} \\ &\quad + Cg_2(x_Q(t))s(t). \end{aligned} \tag{37}$$

On the other hand, we consider the system H for $(\alpha^*, \beta_1^*, \beta_2^*, \gamma^*, \delta^*)$ and $\xi(0) := [x_Q^\top(0) \;\; x_I^\top(0)]^\top$. For the state $\xi = [\xi_1^\top \;\; \xi_2^\top]^\top$, we can obtain the relations

$$\xi_1(t) = x_Q(t), \quad \xi_2(t) = x_I(t) \quad (\forall t \in \mathbf{N}) \tag{38}$$

in a similar way to (29) and (30), where $x_Q(t)$ and $x_I(t)$ are the states of Σ (for the following s). Moreover, the output $s(t)$ is expressed as

$$\begin{aligned} s(t) &= \gamma^*(\xi(t)) + w(t) \\ &\quad + (\delta^*(\xi(t)) - I)(h_2(\xi_1(t)) + k_2(\xi_1(t))r(t)) \\ &= -(Cg_2(\xi_1(t)))^{-1}[C \;\; -C]\begin{bmatrix} f(\xi_1(t)) + g_1(\xi_1(t))r(t) \\ f_{cl}(\xi_2(t)) + g_{cl}(\xi_2(t))r(t) \end{bmatrix} \\ &\quad - h_2(\xi_1(t)) - k_2(\xi_1(t))r(t) + w(t), \end{aligned} \tag{39}$$

where the first equality is given by (35) and the second one is done by (1), (22), (23), (25), and (26). Applying (38) and (39) to (37) provides

$$e(t+1) = Cg_2(x_Q(t))w(t). \tag{40}$$

This implies (36) (note the definition of the inverse). ■

Theorem 2 shows the components of the optimal quantizer Q^*. Part (a) is the same as the unity feedback in the unquantized system Σ_I and plays a role to imitate Σ_I. Part (b) is for

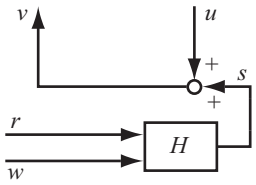


Fig. 8. Equivalent representation of Q under assuming (29) (where actually, w is not an exogenous signal but the static quantization error depending on u and ξ).

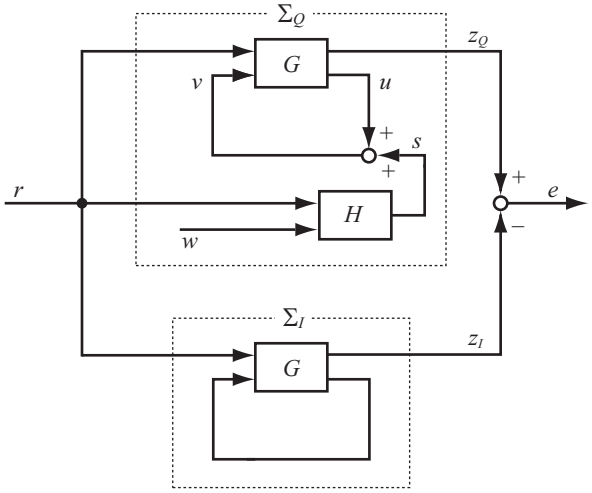


Fig. 9. Error system between Σ_Q and Σ_I under assuming (29) (where note again that w is not exogenous).

minimizing the influence of the static quantization error on the performance index $E(Q)$. In fact, Fig. 9 shows that the error system is a cascade system of Σ and H (in the form of Fig. 3), and (36) means that H for $Q := Q^*$ is a cascade system of the inverse of $\mathbf{z}\Sigma$ and the gain $Cg_2(x_Q)$. So, in the error system, the signal transfer from w to e is reduced by Q^* as shown in (40). Note that (40) corresponds to (27) by considering that the output error e specifies $E(Q)$ and the quantization error w satisfies (33). Finally, (c) comes from the definition of $E(Q)$ in (4). In this way, we have structural interpretation of Q^*.

The above result is somewhat unexpected due to the following reasons. First, the optimal structure of Q can be *rigorously* explained by an approximate inverse of Σ, even though Q is a continuous-to-discrete map but Σ is a continuous-to-continuous map. Second, Theorem 2 is not the same as the result in [31] given for linear G; it has been shown in [31] that the optimal quantizers for linear G include an approximate inverse not of Σ but of Σ'_I given in Fig. 11.

B. Stability of Optimally Quantized Feedback Systems

Now, a stability condition of the optimally quantized system Σ_{Q^*} is given. We employ the stability notion defined in Section I, because the system Σ_Q for some G cannot be asymptotically stable with any Q, as stated at the end of Section II.

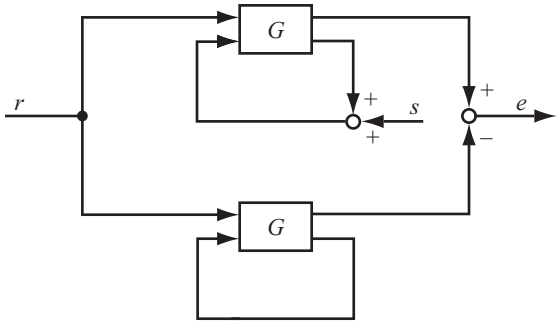


Fig. 10. System Σ.

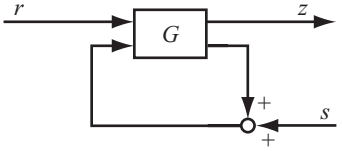


Fig. 11. System Σ'_I.

Theorem 3: For the system Σ_Q, suppose that G and d are given and assume (A1)–(A3). Let Σ'_I denote the system in Fig. 11, which is the unquantized system Σ_I with the new input s. Then the optimally quantized system Σ_{Q^*} is stable if f, g_1, g_2, h_2, and k_2 are continuous functions (on their domains), $E(Q^*) < \infty$ (i.e., $\sup_{x\in\mathbf{R}^n} \|Cg_2(x)\| < \infty$), and the systems Σ'_I and $(\mathbf{z}\Sigma'_I)_s^{-1}$ are stable.

Proof: The proof directly follows from the five facts: (i) under the condition (29), Σ_Q is stable if H (in (35)) and Σ'_I are stable and H is output-stable, (ii) (29) holds for $Q := Q^*$, (iii) H^* (*i.e.*, H for $Q := Q^*$) is stable if Σ'_I is stable and H^* is output-stable, (iv) H^* is output-stable if $(\mathbf{z}\Sigma)_s^{-1}$ is output-stable and $E(Q^*) < \infty$, (v) $(\mathbf{z}\Sigma)_s^{-1}$ is output-stable if Σ'_I and $(\mathbf{z}\Sigma'_I)_s^{-1}$ are stable. These facts are shown as follows. (i) Note that the cascade system in the form of Fig. 3, where $S^{(2)}$ is not necessarily an inverse of $S^{(1)}$, is stable if $S^{(1)}$ and $S^{(2)}$ are stable and $S^{(2)}$ is output-stable. Using this, it is proven by the fact that, under (29), Σ_Q is a cascade system of H and Σ'_I as shown in Fig. 9. (ii) It is shown in the proof of Theorem 1. (iii) Consider the state $[x_Q^\top(t) \;\; x_I^\top(t)]^\top$ of the error system in Fig. 9 where $H := H^*$. As can imagine from the figure, $x_I(t)$ is finite if Σ'_I is stable, while $x_Q(t)$ is finite if Σ'_I is stable and the output $s(t)$ of H^* is finite (H^* is output-stable). On the other hand, (38) holds between the state of H^* and the state of the error system. These complete the proof. (iv) It turns out from (36) that H^* is output-stable if the system in the right-hand side of (36) is output-stable. The system in the right-hand side is the cascade system of $(\mathbf{z}\Sigma)_s^{-1}$ and the time-varying gain $Cg_2(x_Q)$, and thus it is output-stable if $(\mathbf{z}\Sigma)_s^{-1}$ is output-stable and $Cg_2(x_Q(t))$ is finite. On the other hand, from (27), $E(Q^*) < \infty$ implies that $Cg_2(x_Q(t))$ is finite. So (iv) holds. (v) From Figs. 10 and 11, $\Sigma(r,s) = \Sigma'_I(r,s) - \Sigma'_I(r,0)$. This and the definition of the inverse system (Section I) give $(\mathbf{z}\Sigma)_s^{-1}(r,w) = (\mathbf{z}\Sigma'_I)_s^{-1}(r, w + \mathbf{z}\Sigma'_I(r,0))$. Namely, $(\mathbf{z}\Sigma)_s^{-1}(r,w)$ is equivalent to the cascade system of $(\mathbf{z}\Sigma'_I)_s^{-1}$

and $w + \mathbf{z}\Sigma_I'(r,0)$. Then (33) holds, $\mathbf{z}\Sigma_I'(r,0)$ is stable if and only if $\Sigma_I'(r,0)$ is stable, and $\mathbf{z}\Sigma_I'(r,0)$ is output-stable if $\mathbf{z}\Sigma_I'(r,0)$ is stable. Moreover, the output map of $(\mathbf{z}\Sigma_I')_s^{-1}$ is continuous for continuous f, g_1, g_2, h_2, and k_2. Thus we have (v). ∎

From this result, it follows that the stability of the optimally quantized system Σ_{Q^*} can be verified by the stability of the two usual systems (which do not involve discrete-valued signals). In particular, the stability of $(\mathbf{z}\Sigma_I')_s^{-1}$ is essential because Σ_I' (*i.e.,* Σ_I) is a reference system for Σ_Q and is usually provided as a stable system.

An example is given. Consider the optimally quantized system Σ_{Q^*} for

$$G:\begin{cases} x(t+1) = \begin{bmatrix} x_1(t)\sin(x_2(t)) \\ 0.5x_2(t) + 1.4x_1(t)\sin(x_2(t)) \end{bmatrix} \\ \qquad\qquad + \begin{bmatrix} -1 \\ 1 \end{bmatrix} r(t) + \begin{bmatrix} 1 \\ 1 \end{bmatrix} v(t), \\ z(t) \;\; = [1 \;\; 0]x(t), \\ u(t) \;\; = -1.4x_1(t)\sin(x_2(t)) + r(t) \end{cases}$$

and $d := 10$, where $x(t) := [x_1(t)\ x_2(t)]^\top \in \mathbf{R}^2$. The corresponding Σ_I' is given by

$$\Sigma_I':\begin{cases} x_I(t+1) = \begin{bmatrix} -0.4x_{I1}(t)\sin(x_{I2}(t)) \\ 0.5x_{I2}(t) \end{bmatrix} \\ \qquad\qquad + \begin{bmatrix} 0 \\ 2 \end{bmatrix} r(t) + \begin{bmatrix} 1 \\ 1 \end{bmatrix} s(t), \\ z_I(t) \;\; = [1 \;\; 0]x_I(t), \end{cases}$$

for $x_I(t) := [x_{I1}(t)\ x_{I2}(t)]^\top \in \mathbf{R}^2$. Furthermore, $(\mathbf{z}\Sigma_I')_s^{-1}$ is

$$(\mathbf{z}\Sigma_I')_s^{-1}:\begin{cases} \chi(t+1) = \begin{bmatrix} 0 \\ 0.5\chi_2(t) + 0.4\chi_1(t)\sin(\chi_2(t)) \end{bmatrix} \\ \qquad\qquad + \begin{bmatrix} 0 \\ 2 \end{bmatrix} r(t) + \begin{bmatrix} 1 \\ 1 \end{bmatrix} \omega(t), \\ \eta(t) \;\; = 0.4\chi_1(t)\sin(\chi_2(t)) + \omega(t) \end{cases}$$

where $\chi(t) := [\chi_1(t)\ \chi_2(t)]^\top \in \mathbf{R}^2$, $\omega(t) \in \mathbf{R}$, and $\eta(t) \in \mathbf{R}$ are the state, input, and output. Then the system Σ_I' is stable because

$$\begin{aligned} \|x_I(t+1)\| &\le \left\| \begin{bmatrix} -0.4\sin(x_{I2}(t)) & 0 \\ 0 & 0.5 \end{bmatrix} \right\| \|x_I(t)\| \\ &\qquad + \left\| \begin{bmatrix} 0 \\ 2 \end{bmatrix} r(t) + \begin{bmatrix} 1 \\ 1 \end{bmatrix} s(t) \right\| \\ &\le 0.5\|x_I(t)\| + 2\|r(t)\| + \|s(t)\| \end{aligned}$$

and thus $\|x_I(t)\| \le 0.5^t\|x_I(0)\| + \sum_{i=0}^{t-1} 0.5^{t-1-i}(2\|r(i)\| + \|s(i)\|)$. By the same way, it can be shown that $(\mathbf{z}\Sigma_I')_s^{-1}$ is stable. Moreover, $E(Q^*) = \sup_{x\in\mathbf{R}^n} \|Cg_2(x)\|(d/2) = 5 < \infty$. Therefore, we conclude from Theorem 3 that Σ_{Q^*} is stable.

Remark 4: In Theorem 3, the stability of $(\mathbf{z}\Sigma_I')_s^{-1}$ is concerned with the minimum phase property of Σ_I'. In particular, in Fig. 1 (a), $(\mathbf{z}\Sigma_I')_s^{-1}$ will be unstable if P is non-minimum phase. Meanwhile, even when P is non-minimum phase, the following technique is useful to avoid the instability. It is well-known that the parallel connection of a given non-minimum phase system and some compensator could be minimum phase. So by constructing a parallel connection for P so as to be minimum phase and regarding it as a new plant, a stable quantized system can be obtained by Theorem 1. ∎

V. Observer-Based Dynamic Quantizers

As seen in Theorem 1, the optimal dynamic quantizer Q^* contains the initial state x_0 of G in its inside. This means that the exact information of x_0 is required to construct Q^* and so it may be a limitation in practice. In this section, we extend the result of Theorem 1 to the case where the information of x_0 is not available.

The idea for the extention is as follows. As shown in (29), the first half of the state equation of Q^* corresponds to a perfect state estimator for Σ_Q with the information of x_0. Thus, by replacing the perfect estimator with an asymptotic observer, it can be expected to obtain a sub-optimal quantizer without the information of x_0.

Now, this idea is formalized. Consider the quantized system Σ_Q in Fig. 12, which is a modified version of that in Fig. 4 (a). The system G is the same as in (1). The quantizer Q is an extended version of (2) so that r is available and the state equation is of the more general form

$$Q:\begin{cases} \xi(t+1) = \hat{\alpha}(\xi(t), u(t), v(t), r(t)), \\ v(t) \;\; = q(\,\gamma(\xi(t)) + \delta_1(\xi(t))u(t) + \delta_2(\xi(t))r(t)\,) \end{cases} \tag{41}$$

where $\hat{\alpha} : \mathbf{R}^\nu \times \mathbf{R}^m \times \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}^\nu$, $\gamma : \mathbf{R}^\nu \to \mathbf{R}^m$, $\delta_1 : \mathbf{R}^\nu \to \mathbf{R}^{m\times m}$, and $\delta_2 : \mathbf{R}^\nu \to \mathbf{R}^{m\times p}$ are functions. The quantizers in this form allow us to construct an observer for the system G in its inside. The initial state of Σ_Q is given as $x(0) = x_0$ and $\xi(0) = \xi_0$, and ν, ξ_0, $\hat{\alpha}$, γ, δ_1, and δ_2 are the parameters to be designed.

For the system Σ_Q, we employ the performance index

$$\hat{E}_T(Q) := \sup_{\substack{(x_0,R)\in\mathbf{X}_0\times\ell_\star^p \\ \hat{x}_0\in\mathbf{R}^n \text{s.t.} \|x_0-\hat{x}_0\|\le\kappa}} \sup_{t\in\{T+1,T+2,\ldots\}} \|z_Q(t,x_0,R) - z_I(t,\hat{x}_0,R)\| \tag{42}$$

where $T \in \mathbf{N}$ is the time specifying the time interval on which Q is evaluated, $\mathbf{X}_0 \subseteq \mathbf{R}^n$ and $\ell_\star^p \subseteq \ell_\infty^p$ are the sets of the initial states and the input sequences of interest, and $\kappa \in \mathbf{R}_+$ is the upper bound of the initial estimation error $\|x_0 - \hat{x}_0\|$, that is, the difference between the true value and the initial guess of the initial state of G. The symbols $z_Q(t,x_0,R)$ and $z_I(t,\hat{x}_0,R)$ are similarly defined as before, but note that the former is for the system Σ_Q in Fig. 12 and the latter is for the system Σ_I in Fig. 4 (b) and the initial state $\hat{x}_0$. The index $\hat{E}_T(Q)$ represents the maximum output difference on the time interval $\{T+1, T+2, \ldots\}$ between the system Σ_Q with $x_Q(0) = x_0$ and the system Σ_I with $x_I(0) = \hat{x}_0$. The time T will be related to the settling time of an observer, in order to purely capture the quantization performance without the transient performance of the observer. Note from (42) and (4) that $\hat{E}_T(Q) = E(Q)$ if $T = 0$, $\mathbf{X}_0 = \mathbf{R}^n$, $\ell_\star^p = \ell_\infty^p$, and $\kappa = 0$. Note also that $\hat{E}_T(Q)$ depends on $\mathbf{X}_0$, $\ell_\star^p$, and κ but which are assumed to be fixed in advance; so the dependence is not explicitly denoted in the symbol "$\hat{E}_T(Q)$".

In considering the quantizer design problem with $\hat{E}_T(Q)$, we assume (A1), (A3), and

(A4) there exists a $\rho \in \mathbf{R}_+$ such that $\|R\| < \rho$ for every $R \in \ell_\star^p$, and the reachable set of the system Σ_I, i.e., $\mathbf{X}_I := \{x \in \mathbf{R}^n | \exists (t, x_0, R) \in \mathbf{N} \times \mathbf{X}_0 \times \ell_\star^p \text{ s.t. } x = x_I(t, x_0, R)\}$, is bounded,

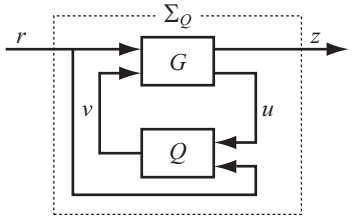


Fig. 12. Quantized feedback system Σ_Q for observer-based dynamic quantizer design.

(A5) the function $\sigma(x_Q, x_I, r)$ in (15) is differentiable with respect to $x_Q \in \mathbf{R}^n$, and there exists an $M \in \mathbf{R}_+$ such that

$$\sup_{(x_Q,x_I,r)\in\mathbf{R}^n\times\mathbf{X}_I\times\bigcup_{t\in\mathbf{N}}\pi_t(\ell^p_\star)} \left\| \frac{\partial\sigma}{\partial x_Q}(x_Q, x_I, r) \right\| \le M \tag{43}$$

where $\pi_t(\ell^p_\star)$ is the projection of $\ell^p_\star$ onto the r_t-space,

(A6) there exists a function $\hat{\alpha}_1 : \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}^n$ such that $\xi_1(t+1) = \hat{\alpha}_1(\xi_1(t), u(t), v(t), r(t))$ is an asymptotic observer for G such that $\|x_Q(t) - \xi_1(t)\|$ is bounded by a class-KL function $\psi(\max\{\|x_Q(0) - \xi_1(0)\|, \|(r(0), r(1), \ldots)\|\}, t)$.

Assumption (A4) is concerned with the boundedness of the input sequence set $\ell^p_\star$ and the reachable set $\mathbf{X}_I$. The condition on $\mathbf{X}_I$ may not be easily checked but it holds if $\mathbf{X}_0$ and $\ell^p_\star$ are bounded and Σ_I is globally exponentially stable. (A5) enables us to estimate the influence of the estimation error of an observer, and (A6) guarantees the existence of an asymptotic observer for G (see, e.g., [36] for nonlinear observers).

Then, we obtain the following result.

Theorem 4: For the system Σ_Q with (1) and (41), suppose that G, d, $\mathbf{X}_0$, $\ell^p_\star$, and κ are given. Assume (A1) and (A3)–(A6), and let ρ, M, $\hat{\alpha}_1$, and ψ be given by (A4)–(A6). Consider the dynamic quantizer

$$Q^\circ := (\nu^\circ, \xi^\circ_0, \hat{\alpha}^\circ, \gamma^\circ, \delta^\circ_1, \delta^\circ_2) \tag{44}$$

with

$$\nu^\circ := 2n, \tag{45}$$

$$\xi^\circ_0 := \begin{bmatrix} \hat{x}_0 \\ \hat{x}_0 \end{bmatrix}, \tag{46}$$

$$\hat{\alpha}^\circ(\xi(t), u(t), v(t), r(t)) := \begin{bmatrix} \hat{\alpha}_1(\xi_1(t), u(t), v(t), r(t)) \\ f_{cl}(\xi_2(t)) + g_{cl}(\xi_2(t))r(t) \end{bmatrix}, \tag{47}$$

$$\gamma^\circ(\xi(t)) := -(Cg_2(\xi_1))^{-1}[C \ \ -C] \begin{bmatrix} f(\xi_1) \\ f_{cl}(\xi_2) \end{bmatrix}, \tag{48}$$

$$\delta^\circ_1(\xi(t)) := 0, \tag{49}$$

$$\delta^\circ_2(\xi(t)) := -(Cg_2(\xi_1))^{-1}[C \ \ -C] \begin{bmatrix} g_1(\xi_1) \\ g_{cl}(\xi_2) \end{bmatrix} \tag{50}$$

where $\xi_1, \xi_2 \in \mathbf{R}^n$ are the first half and the second half of the vector ξ and $\hat{x}_0 \in \mathbf{R}^n$ is arbitrarily given so that $\|\hat{x}_0 - x_0\| \le \kappa$. Then

$$\hat{E}_T(Q^\circ) \le 3 \sup_{x\in\mathbf{R}^n} \|Cg_2(x)\| \frac{d}{2} \tag{51}$$

holds for every T satisfying

$$\psi(\max\{\kappa, \rho\}, T) \le \frac{d}{2M}. \tag{52}$$

Proof: For the system Σ_Q with $Q := Q^\circ$, suppose that $(x_0, \hat{x}_0) \in \mathbf{X}_0 \times \mathbf{R}^n$ and $R := (r_0, r_1, \ldots) \in \ell^p_\star$ are given and assume that $\|\hat{x}_0 - x_0\| \le \kappa$, $x_Q(0) = x_0$, and $(r(0), r(1), \ldots) = R$. From (41), (44), (46), (47), and (A6), the following relations hold for the state ξ of Q°:

$$\xi_1(t) = x_Q(t, x_0, R) + \varepsilon(t) \quad (\forall t \in \mathbf{N}), \tag{53}$$

$$\xi_2(t) = x_I(t, \hat{x}_0, R) \quad (\forall t \in \mathbf{N}) \tag{54}$$

where $\varepsilon(t) \in \mathbf{R}^n$ is the estimation error such that

$$\begin{aligned} \|\varepsilon(t)\| &\le \psi(\max\{\|x_0 - \hat{x}_0\|, \|R\|\}, t) \\ &\le \psi(\max\{\kappa, \rho\}, t) \quad (\forall t \in \mathbf{N}) \end{aligned} \tag{55}$$

and $x_Q(t, x_0, R)$ and $x_I(t, \hat{x}_0, R)$ are the states of Σ_Q in Fig. 12 and of Σ_I in Fig. 4 (b), defined in a similar way to $z_Q(t, x_0, R)$ and $z_I(t, \hat{x}_0, R)$. So ξ_1 and ξ_2 correspond to the estimation of x_Q and the copy of x_I. Then (15), (41), (44), (48), (49), (50), (53), (54), (A5), and the mean value theorem enable us to express $v(t)$ as

$$\begin{aligned} v(t) &= q\big(\sigma(\xi_1(t), \xi_2(t), r(t))\big) \\ &= q\big(\sigma(x_Q(t, x_0, R) + \varepsilon(t), x_I(t, \hat{x}_0, R), r_t)\big) \\ &= q\Big(\sigma(x_Q(t, x_0, R), x_I(t, \hat{x}_0, R), r_t) \\ &\qquad + \frac{\partial\sigma}{\partial x_Q}(\check{x}_Q(t, x_0, R), x_I(t, \hat{x}_0, R), r_t)\, \varepsilon(t)\Big) \end{aligned} \tag{56}$$

where $\check{x}_Q(t, x_0, R)$ is some vector on the line segment between $x_Q(t, x_0, R)$ and $x_Q(t, x_0, R) + \varepsilon(t)$. Meanwhile,

$$\frac{\partial\sigma}{\partial x_Q}(\check{x}_Q(t, x_0, R), x_I(t, \hat{x}_0, R), r_t)\, \varepsilon(t) \in \left[-\frac{d}{2}, \frac{d}{2}\right]^m \tag{57}$$

holds for $t \in \{T, T+1, \ldots\}$, since (43) and (55) provide

$$\begin{aligned} \left\| \frac{\partial\sigma}{\partial x_Q}(\check{x}_Q(t), x_I(t), r_t)\varepsilon(t) \right\| &\le \left\| \frac{\partial\sigma}{\partial x_Q}(\check{x}_Q(t), x_I(t), r_t) \right\| \|\varepsilon(t)\| \\ &\le M\psi(\max\{\kappa, \rho\}, t) \end{aligned}$$

and (52) means that $M\psi(\max\{\kappa, \rho\}, t) \le Md/(2M) = d/2$ for $t \in \{T, T+1, \ldots\}$. Therefore, we have the following expression of $v(t)$ from (56), (57), and the definition of q (in Section II-A):

$$v(t) = q\big(\sigma(x_Q(t, x_0, R), x_I(t, \hat{x}_0, R), r_t)\big) + \theta(t) \quad (\forall t \ge T) \tag{58}$$

where $\theta(t)$ is some vector such that

$$\theta(t) \in \{-d, 0, d\}^m. \tag{59}$$

By applying (15) and (58) to (13), the output difference $z_Q(t+1, x_0, R) - z_I(t+1, \hat{x}_0, R)$ for $t \ge T$ is calculated as

$$\begin{aligned} &z_Q(t+1, x_0, R) - z_I(t+1, \hat{x}_0, R) \\ &= C\big(f(x_Q(t, x_0, R)) + g_1(x_Q(t, x_0, R))r_t \\ &\qquad + g_2(x_Q(t, x_0, R))v(t)\big) \\ &\quad - C\big(f_{cl}(x_I(t, \hat{x}_0, R)) + g_{cl}(x_I(t, \hat{x}_0, R))r_t\big) \\ &= Cg_2(x_Q(t, x_0, R))\big(q(\sigma(t)) - \sigma(t) + \theta(t)\big) \end{aligned}$$

where $\sigma(t)$ stands for $\sigma(x_Q(t,x_0,R),x_I(t,\hat{x}_0,R),r_t)$ and we note that (13) holds under (A1). Since $q(\sigma(t))-\sigma(t)+\theta(t) \in [-3d/2,3d/2]^m$ from (3) and (59), it follows that

$$\|z_Q(t,x_0,R)-z_I(t,\hat{x}_0,R)\| \le 3\|Cg_2(x_Q(t-1,x_0,R))\|\frac{d}{2} \quad (\forall t \ge T+1),$$

which, together with (42), proves (51). ∎

This result presents an observer-based quantizer Q° together with the performance evaluation of (51). In (51), T expresses the time when the estimation error $\|\xi_1(t)-x_Q(t)\|$ becomes sufficiently small, and the right-hand side is an upper bound of the output difference after T. An interesting point in this result is that the right-hand side of (51) is the triple of the right-hand side of (27). This means that introducing an observer gives a dynamic quantizer without the information of the initial state x_0 but degrades the performance in terms of the time T and the three-times larger bound. The former degradation is caused by the transient of the observer and the latter comes from the nonzero estimation error in the steady state, *i.e.*, the fact that $\|\xi_1(t)-x_Q(t)\|$ will be nearly zero but not be just zero for any $t \in \mathbf{N}$.

Remark 5: Though the dynamic quantizer in the form of (41) is a generalized version of that in (2), it is a fact under (A1)–(A3) that the right-hand side of (27) is a lower bound of the minimum value of $E(Q)$ with respect to Q in the form of (41). This can be proven in the same way as in Section III-B.1. ∎

Remark 6: A lower bound of the minimum value of $\hat{E}_T(Q)$, which holds for any $T \in \mathbf{N}$, is given as

$$\inf_{x\in\mathbf{R}^n}\|Cg_2(x)\|\frac{d}{2} \le \min_Q \hat{E}_T(Q)$$

subject to (A1), (A2), and

(A3') $p=l=m=1$ and $C(g_1(x_1)-g_{cl}(x_2)) \neq 0$ for every $(x_1,x_2)\in\mathbf{R}^n\times\mathbf{R}^n$.

This can be derived by the fact that the right-hand side of (13) is equal to $Cg_2(x_Q(t))((d/2)+v(t))$ for $r(t) := (C(g_1(x_Q(t))-g_{cl}(x_I(t))))^{-1}(Cg_2(x_Q(t))(d/2)-C(f(x_Q(t))-f_{cl}(x_I(t))))$ and $v(t)\in\{0,\pm d,\pm 2d,\ldots\}$ for any Q. Thus, in addition to the above observation, it turns out that, if (A1), (A2), (A3'), and (A4)–(A6) hold and $Cg_2(x)$ is a constant (i.e., $\inf_{x\in\mathbf{R}^n}\|Cg_2(x)\| = \sup_{x\in\mathbf{R}^n}\|Cg_2(x)\|$), the right-hand side of (51) is the triple of a lower bound of $\min_Q \hat{E}_T(Q)$. ∎

VI. Conclusion

This paper has discussed a dynamic quantizer design problem for command-driven nonlinear control. Based on the bound analysis of the optimal performance, we have obtained an optimal dynamic quantizer in a closed form. This has also shown the performance limitation of a general class of nonlinear dynamic quantizers. Moreover, the structure of the optimal quantizer and the stability of the optimally quantized system have been disclosed. Finally, observer-based dynamic quantizers have been presented so as to be utilized in many practical situations. We expect that the result will be a foundation for the dynamic quantization of nonlinear control systems.

Since this paper has aimed at obtaining analytical results, the problem has been solved in somewhat limited cases. In the future, a method to solve the problem in its full generality should be developed. For such an issue, the idea of numerical optimization, which has been proposed in [24] for linear systems, will be useful.

Appendix I
Proof of relation (9)

A. Notation

For the vector x and the matrix M, let $\langle x\rangle_i$ and $\langle M\rangle_i$ denote the ith element of x and the ith row vector of M, respectively. The symbol $\mathrm{sign}(x)$ expresses the vector obtained by elementwisely applying the signum function to the vector x.

B. Main Part

The first case of (9) is the direct consequence of the following lemma.

Lemma 1: For the system Σ_Q, suppose that G, Q, and d are given. Then the following statements hold.
(i) Let $R^*(\epsilon)\in\ell_\infty^p$ be an external input sequence parameterized by a number $\epsilon\in(0,d/2)$. Then

$$\sup_{(x_0,R)\in\mathbf{R}^n\times\ell_\infty^p}\|z_Q(1,x_0,R)-z_I(1,x_0,R)\| \ge \sup_{x_0\in\mathbf{R}^n}\sup_{\epsilon\in(0,d/2)}\|z_Q(1,x_0,R^*(\epsilon))-z_I(1,x_0,R^*(\epsilon))\|.$$

(ii) Let $R^*(\epsilon) := (r_0^*(\epsilon),r_1,r_2,\ldots)$ for

$$r_0^*(\epsilon) := k_2^{-1}(x_0)\bigg(-\mathrm{sign}(\langle Cg_2(x_0)\gamma(\xi_0)\rangle_{i^*}) \times \mathrm{sign}(\langle Cg_2(x_0)\rangle_{i^*})^\top\left(\frac{d}{2}-\epsilon\right)-h_2(x_0)-\gamma(\xi_0)\bigg)$$

and arbitrarily given $(r_1,r_2,\ldots)\in\ell_\infty^p$ where $i^*\in\{1,2,\ldots,m\}$ is defined by (61) and $\Lambda := Cg_2(x_0)$ in the next subsection. If $\delta(\xi_0)=I$, then

$$\sup_{x_0\in\mathbf{R}^n}\sup_{\epsilon\in(0,d/2)}\|z_Q(1,x_0,R^*(\epsilon))-z_I(1,x_0,R^*(\epsilon))\| \ge \sup_{x_0\in\mathbf{R}^n}\|Cg_2(x_0)\|\frac{d}{2}. \quad (60)$$

Proof: Statement (i) is trivial, while (ii) is proven in Appendix I-C. ∎

The second case of (9) is given by the fact that the right-hand side of (7) is *not* a bounded function with respect to $r_0\in\mathbf{R}^p$. Note here that $Cg_2(x_0)(\delta(\xi_0)-I)k_2(x_0)\neq 0_{l\times p}$ under (A2), (A3), and $\delta(\xi_0)\neq I$.

C. Proof of Lemma 1 (ii)

1) Preliminary: First, we provide a preliminary result.

Lemma 2: (i) Suppose that a matrix $\Lambda\in\mathbf{R}^{l\times m}$ and a positive number $\zeta\in\mathbf{R}_+$ are given. Let

$$i^* := \ \mathrm{argmax}_{i\in\{1,2,\ldots,m\}}\sum_{j=1}^m|\Lambda_{ij}| \quad (61)$$

where Λ_{ij} is the (i,j)th element of Λ. Then

$$\left\|\Lambda \operatorname{sign}(\langle\Lambda\rangle_{i^*})^\top \zeta\right\| = \|\Lambda\|\,\zeta.$$

(ii) Suppose that vectors $\lambda_1, \lambda_2 \in \mathbf{R}^l$ are given. If there exists an $i \in \{1, 2, \ldots, m\}$ such that $\|\lambda_1\| = |\langle\lambda_1\rangle_i|$ and $\langle\lambda_1\rangle_i\langle\lambda_2\rangle_i \geq 0$, then

$$\|\lambda_1 + \lambda_2\| \geq \|\lambda_1\|.$$

Proof: The statements are straightforwardly proven by the definition of the ∞-norm. ∎

2) Main Part: From (8), we have

$$\begin{aligned}
&a(x_0, \xi_0, r_0^*(\epsilon)) \\
&= (I - \delta(\xi_0))\gamma(\xi_0) \\
&\quad - \delta(\xi_0)\operatorname{sign}(\langle Cg_2(x_0)\gamma(\xi_0)\rangle_{i^*})\operatorname{sign}(\langle Cg_2(x_0)\rangle_{i^*})^\top\left(\frac{d}{2} - \epsilon\right)
\end{aligned}$$

for $R := R^*(\epsilon)$. Since $q(a(x_0, \xi_0, r_0^*(\epsilon))) = 0$ under $\delta(\xi_0) = I$, (7) provides

$$\begin{aligned}
&z_Q(1, x_0, R^*(\epsilon)) - z_I(1, x_0, R^*(\epsilon)) \\
&\quad = Cg_2(x_0)\Bigg(\operatorname{sign}(\langle Cg_2(x_0)\gamma(\xi_0)\rangle_{i^*}) \\
&\qquad\qquad \times \operatorname{sign}(\langle Cg_2(x_0)\rangle_{i^*})^\top\left(\frac{d}{2} - \epsilon\right)\Bigg) \\
&\qquad + Cg_2(x_0)\gamma(\xi_0) \qquad (62)
\end{aligned}$$

subject to $\delta(\xi_0) = I$. Note here that $\operatorname{sign}(\langle Cg_2(x_0)\gamma(\xi_0)\rangle_{i^*})$ is a scalar, and $\langle Cg_2(x_0)\operatorname{sign}(\langle Cg_2(x_0)\rangle_{i^*})^\top\rangle_{i^*} \geq 0$. By applying Lemma 2 to (62) with $\Lambda := Cg_2(x_0)$, $\zeta := (d/2) - \epsilon$, $\lambda_1 := Cg_2(x_0)\operatorname{sign}(\langle Cg_2(x_0)\gamma(\xi_0)\rangle_{i^*})\operatorname{sign}(\langle Cg_2(x_0)\rangle_{i^*})^\top((d/2) - \epsilon)$, and $\lambda_2 := Cg_2(x_0)\gamma(\xi_0)$, it follows that

$$\|z_Q(1, x_0, R^*(\epsilon)) - z_I(1, x_0, R^*(\epsilon))\| \geq \|Cg_2(x_0)\|\left(\frac{d}{2} - \epsilon\right).$$

This proves (60).

REFERENCES

[1] H. Ishii and B. A. Francis, "Stabilizing a linear system by switching control with dwell time," *IEEE Transactions on Automatic Control*, vol. 47, no. 12, pp. 1962–1973, 2002.

[2] H. Ishii and T. Basar, "Remote control of LTI systems over networks with state quantization," in *Proceedings of the 41st IEEE Conference on Decision and Control*, pp. 830–835, 2002.

[3] G.N. Nair and R.J. Evans, "Exponential stabilisability of finite-dimensional linear systems with limited data rates," *Automatica*, vol. 39, no. 4, pp.585–593, 2003.

[4] G.N. Nair and R.J. Evans, "Stabilizability of stochastic linear systems with finite feedback data rates," *SIAM Journal on Control and Optimization*, vol. 43, no.2, pp. 413–436, 2004.

[5] S. Tatikonda and S. Mitter, "Control under communication constraints," *IEEE Transactions on Automatic Control*, vol. 49, no.7, pp. 1056–1068, 2004.

[6] S. Tatikonda and S. Mitter, "Control over noisy channels," *IEEE Transactions on Automatic Control*, vol. 49, no.7, pp. 1196–1201, 2004.

[7] R.W. Brockett and D. Liberzon, "Quantizer feedback stabilization of linear systems," *IEEE Transactions on Automatic Control*, vol. 45, no. 7, pp. 1279–1289, 2000.

[8] D. Liberzon and D. Nesic, "Input-to-state stabilization of linear systems with quantized state measurements," *IEEE Transactions on Automatic Control*, vol. 52, no. 5, pp. 767–781, 2007.

[9] N. Elia and S.K. Mitter, "Stabilization of linear systems with limited information," *IEEE Transactions on Automatic Control*, vol. 46, no. 9, pp. 1384–1400, 2001.

[10] M. Fu and L. Xie, "The sector bound approach to quantized feedback control," *IEEE Transactions on Automatic Control*, vol. 50, no. 11, pp. 1698–1711, 2005.

[11] K. Tsumura and J. Maciejowski, "Optimal quantization of signals for system identification," in *Proceedings of the European Control Conference 2003*, 2003.

[12] K. Tsumura, "Asymptotic property of optimal quantization for system identification," *Mathematical Engineering Technical Reports of the University of Tokyo*, METR 2004-10, 2004.

[13] H. Ishii and T. Basar, "An analysis on quantization effects in H^∞ parameter identification," in *Proceedings of IEEE International Conference on Control Applications*, pp. 468–473, 2004.

[14] K. Tsumura, "Criteria for system identification with quantized data and the optimal quantization schemes," in *Proceedings of the 16th IFAC World Congress*, Mo-M02-TP/6, 2005.

[15] D.E. Quevedo and G.C. Goodwin, "Audio quantization from a receding horizon control perspective," in *Proceedings of the 2003 American Control Conference*, pp. 4131–4136, 2003.

[16] D.E. Quevedo, G.C. Goodwin, and J.A. De Dona, "Finite constraint set receding horizon quadratic control," *International Journal of Robust and Nonlinear Control*, vol. 14, no. 4, pp. 355–377, 2004.

[17] G.I. Bourdopoulos, *Delta-Sigma Modulators: Modeling, Design and Applications*, Imperial College Press, 2003.

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

[19] C. Canudas-de-Wit, F.R. Rubio, J. Fornés, and F. Gómez-Estern, "Differential coding in networked controlled linear systems," in *Proceedings of the 2006 American Control Conference*, pp. 4177–4182, 2006.

[20] F. Bullo and D. Liberzon, "Quantized control via locational optimization," *IEEE Transactions on Automatic Control*, vol. 51, no. 1, pp. 2–13, 2006.

[21] B. Picasso and A. Bicchi, "On the stabilization of linear systems under assigned I/O quantization," *IEEE Transactions on Automatic Control*, vol. 52, no. 10, pp. 1994–2000, 2007.

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

[23] Y. Minami, S. Azuma, and T. Sugie, "Optimal dynamic quantizers for discrete-valued input feedback control," in *Proceedings of the 46th IEEE Conference on Decision and Control*, pp. 2259–2264, 2007.

[24] S. Azuma and T. Sugie, "Synthesis of optimal dynamic quantizers for discrete-valued input control," *IEEE Transactions on Automatic Control*, vol. 53, no. 12, pp. 2064–2075, 2008.

[25] S. Azuma, Y. Minami, and T. Sugie, "Optimal dynamic quantizers for feedback control with discrete-level actuators: unified solution and experimental evaluation," *Transactions of the ASME, Journal of Dynamic Systems, Measurement and Control*, vol. 133, no. 2, art. 021005, 2010.

[26] G.N. Nair, R.J. Evans, I.M.Y. Mareels, and W Moran, "Topological feedback entropy and nonlinear stabilization," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1585–1597, 2004.

[27] D. Liberzon and J.P. Hespanha, "Stabilization of nonlinear systems with limited information feedback," *IEEE Transactions on Automatic Control*, vol. 50, no. 6, pp. 910–915, 2005.

[28] C. De Persis, "n-bit stabilization of n-dimensional nonlinear systems in feedforward form," *IEEE Transactions on Automatic Control*, vol. 50, no. 3, pp. 299–311, 2005.

[29] D. Liberzon, "Quantization, time delays, and nonlinear stabilization," *IEEE Transactions on Automatic Control*, vol. 51, no. 7, pp. 1190–1195, 2006.

[30] C. De Persis, "Minimal data rate control of nonlinear systems over networks with large delay," *International Journal of Robust and Nonlinear Control*, vol. 20, no. 10, pp. 1097–1111, 2010.

[31] S. Azuma and T. Sugie, "Stability analysis of optimally quantized LFT-feedback systems," *International Journal of Control*, vol. 83, no. 6, pp. 1125–1135, 2010.

[32] D. Liberzon, *Switching in systems and control*, Birkhauser, Boston, 2003.

[33] S. Azuma and T. Sugie, "An analytical solution to dynamic quantization problem of nonlinear control systems," in *Proceedings of the combined 48th IEEE Conference on Decision Control and 28th Chinese Control Conference*, pp. 3914–3919, 2009.

[34] L. Magni, G. De Nicolao, R. Scattolini, and F. Allgower, "Robust model predictive control of nonlinear discrete-time systems," *International Journal of Robust and Nonlinear Control*, vol. 13, no. 3–4, pp. 229–246, 2003.

[35] Y. Pan and J. Wang, "Model predictive control for nonlinear affine systems based on the simplified dual neural network," in *Proceedings of 2009 IEEE International Symposium on Intelligent Control*, pp. 683–688, 2009.

[36] P.E. Moraal and J.W. Grizzle, "Observer design for nonlinear systems with discrete-time measurements," *IEEE Transactions on Automatic Control*, vol. 40, no. 3, pp. 395–404, 1995.

Shun-ichi Azuma (S'03-M'05) was born in Tokyo, Japan, in 1976. He received the B.Eng. degree in electrical engineering from Hiroshima University, Higashi Hiroshima, Japan, in 1999, and the M.Eng. and Ph.D. degrees in control engineering from Tokyo Institute of Technology, Tokyo, Japan, in 2001 and 2004, respectively. He was a research fellow of the Japan Society for the Promotion of Science at Tokyo Institute of Technology from 2004 to 2005 and an Assistant Professor in the Department of Systems Science, Graduate School of Informatics, Kyoto University, Uji, Japan, from 2005 to 2011. He is currently an Associate Professor at Kyoto University. He held visiting positions at Georgia Institute of Technology, Atlanta GA, USA, from 2004 to 2005 and at University of Pennsylvania, Philadelphia PA, USA, from 2009 to 2010. He serves as an Associate Editor of IEEE CSS Conference Editorial Board from 2011. His research interests include analysis and control of hybrid systems.

Toshiharu Sugie (F'07) received the B.E., M.E., and Ph.D. degrees in engineering form Kyoto University, Japan, in 1976, 1978 and 1985, respectively. From 1978 to 1980, he was a research member of Musashino Electric Communication Laboratory in NTT, Musashino, Japan. From 1984 to 1988, he was a research associate of Department of Mechanical Engineering, University of Osaka Prefecture, Osaka. In 1988, he joined Kyoto University, where he is currently a Professor of Department of Systems Science. He serves as an Editor of Automatica, and was also an Associate Editor of Asian Journal of Control and International Journal of Systems Science. His research interests are in robust control, identification for control, and control application to mechanical systems. He is a Fellow of the Society of Instrument and Control Engineers, Japan.