Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Optimal Dynamic Quantizers for 2D Systems with Discrete-Valued Input and Its Application to Generation of Binary Halftone Images

Yuki Minami * **Shun-ichi Azuma** * **Toshiharu Sugie** *

* *Kyoto University, Uji, Kyoto 611-0011, Japan,*
(Tel: +81-774-38-3954; e-mail: minami@robot.kuass.kyoto-u.ac.jp,
{sazuma, sugie}@i.kyoto-u.ac.jp)

Abstract: This paper addresses an optimal design problem of dynamic quantizers for a class of 2D systems with discrete-valued control input. First, we derive a closed form expression of the performance of a class of dynamic quantizers. Next, based on that, an optimal dynamic quantizer is provided. Finally, we apply the optimal dynamic quantizer to generate binary halftone images.

1. INTRODUCTION

In recent years, analysis and design problems of quantizers have been actively discussed from a control point of view. So various results on this topic have been derived (e.g., Nair and Evans (2003); Tatikonda and Mitter (2004); Bullo and Liberzon (2006); Quevedo and Goodwin (2003); Azuma and Sugie (2008); Minami, Azuma, and Sugie (2007)).

These results, however, have been derived only for 1D systems with discrete-valued signal constrains. So no quantizer for 2D systems has been proposed so far, although 2D systems have been received extensive attention in several modern engineering fields such as image processing, modeling of partial differential equations, and control of repetitive systems (e.g., Roesser (1975); Galkowski et al. (1999)). Therefore, the design problem of quantizers for 2D system is one of novel and interesting research topics.

Motivated by the above background, this paper addresses an optimization problem of a class of quantizers for 2D systems with discrete-valued input, i.e., for 2D systems whose input takes only values on a fixed discrete set. In particular, we focus on an optimal design problem of a class of dynamic quantizers, since dynamic quantizers can be much better than static ones to achieve high performance. In this paper, we consider the following problem as an extension of our previous work (Minami, Azuma, and Sugie (2007)): when a 2D plant and a 2D controller are given in the feedback system in **Fig. 1** (a), find a dynamic quantizer such that the system in **Fig. 1** (a) optimally approximates the ideal feedback system in (b) in terms of the controlled output.

To this problem, the main contributions of this paper are as follows. First, we analytically derive an optimal dynamic quantizer for a class of 2D systems with discrete-valued input. The optimal dynamic quantizer proposed in this paper allows us to use the conventional controller design theory for 2D systems, even though 2D systems have discrete-valued input constraints. Second, we apply the optimal dynamic quantizer to generate binary halftone images. The halftoning is a process of transforming grayscale images to binary images, and halftone images resemble original gray images in appearance. In this paper, a halftone image is generated by the optimal quantizer.

Notation: Let $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{N}$ denote the real number field, the set of positive real numbers, and the set of positive integers, respectively. For the matrix $H := \{H_{ij}\}$, let $\mathrm{abs}(H)$ be the matrix composed of the absolute values of the elements, i.e., $\mathrm{abs}(H) := \{|H_{ij}|\}$, and we use I, 0, and $\mathbf{1}$ to express the identity matrix, the zero matrix, and the vector whose all elements are one. For $i, j, h, k \in \mathbb{N}$, let $(i, j) = (h, k)$ express $i = h$ and $j = k$. For the vector x, $\mathrm{sign}(x)$ expresses the vector obtained by elementwisely applying the signum function to x. Finally, for the vector x, the matrix H, and the sequence of the vectors $X := \{x_1, x_2, \ldots, x_f\}$, the symbols $\|x\|$, $\|H\|$, and $\|X\|$ express their ∞-norms (note that $\|X\| := \sup_{i\in\{1,2,\ldots,f\}} \|x_i\|$).

2. PROBLEM FORMULATION

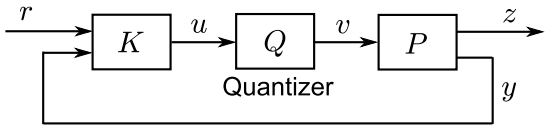
Let us consider the feedback system Σ_Q in **Fig. 1** (a) composed of the linear 2D plant P, the controller K, and the quantizer Q. The plant P is the Fornasini-Marchesini first model (Fornasini and Marchesini (1978)) given by

$$P : \begin{cases} x(i+1,j+1) = A_0x(i,j) + A_1x(i,j+1) \\ \qquad\qquad\qquad +A_2x(i+1,j) + Bv(i,j), \\ z(i,j) = C_1x(i,j), \\ y(i,j) = C_2x(i,j) \end{cases} \tag{1}$$

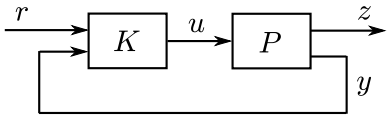
where $x \in \mathbb{R}^n$ is the state, $v \in \mathbb{R}^m$ is the input, $z \in \mathbb{R}^{l_1}$ is the controlled output, $y \in \mathbb{R}^{l_2}$ is the measured output, A_0, A_1, $A_2 \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C_1 \in \mathbb{R}^{l_1\times n}$, $C_2 \in \mathbb{R}^{l_2\times n}$ are constant matrices, and $(i, j) \in \{0, 1, \ldots, M-1\} \times \{0, 1, \ldots, N-1\}$ for M, $N \in \mathbb{N}$. For given M, $N \in \mathbb{N}$, let $X(0)$ be represented by

$$X(0) := [\ x(0,0)\ x(1,0)\ \cdots\ x(M,0) \\ x(0,1)\ x(0,2)\ \cdots\ x(0,N)\ ]. \tag{2}$$

Then the boundary condition is given as $X(0) = X_0$ for $X_0 \in \mathbb{R}^{n\times(M+N+1)}$. The controller K is given by



(a) Feedback system Σ_Q with an input quantizer



(b) Ideal feedback system Σ

Fig. 1. Two feedback control systems

$$K : u(i,j) = F_0 y(i,j) + F_1 y(i,j+1) + F_2 y(i+1,j) + G r(i,j) \tag{3}$$

where $u \in \mathbb{R}^m$ is the output, $r \in \mathbb{R}^p$ is the reference, and $F_0,\ F_1,\ F_2 \in \mathbb{R}^{m\times l_2}$, $G \in \mathbb{R}^{m\times p}$ are constant matrices. We suppose that the controller K internally stabilizes the plant P in **Fig. 1** (b). Finally, Q is the dynamic quantizer in the form of

$$Q : \begin{cases} \xi(i+1,j+1) = \mathcal{A}_0\xi(i,j) + \mathcal{A}_1\xi(i,j+1) + \mathcal{A}_2\xi(i+1,j) \\ \qquad\qquad +\mathcal{B}_1 u(i,j) + \mathcal{B}_2 v(i,j), \\ v(i,j) = \mathrm{q}\Big[\mathcal{C}_0\xi(i,j) + \mathcal{C}_1\xi(i,j+1) \\ \qquad\qquad +\mathcal{C}_2\xi(i+1,j) + u(i,j)\Big] \end{cases} \tag{4}$$

where $\xi \in \mathbb{R}^n$ is the state, $u \in \mathbb{R}^m$ is the input, $v \in \mathbb{V}^m$ is the output, $\mathbb{V}^m \subset \mathbb{R}^m$ is the discrete set on which the output takes values, $\mathcal{A}_0,\ \mathcal{A}_1,\ \mathcal{A}_2 \in \mathbb{R}^{n\times n}$, $\mathcal{B}_1,\ \mathcal{B}_2 \in \mathbb{R}^{n\times m}$, $\mathcal{C}_0,\ \mathcal{C}_1,\ \mathcal{C}_2 \in \mathbb{R}^{m\times n}$ are constant matrices, and $\mathrm{q} : \mathbb{R}^m \to \mathbb{V}^m$ is the static quantizer. Using different fonts, we distinguish the symbols $(\mathcal{A},\mathcal{B},\mathcal{C})$ used in Q from (A,B,C) in P. For given $M,\ N \in \mathbb{N}$, the bounday condition is given as

$$\begin{cases} \xi(0,0) = 0 \\ \xi(i,0) = 0 \quad (i = 1,2,\ldots,M) \\ \xi(0,j) = 0 \quad (j = 1,2,\ldots,N) \end{cases} \tag{5}$$

for guaranteeing that Q is drift-free, i.e., $v(i,j) = 0$ for $u(i,j) = 0\,(i = 0,1,2,\ldots,M{-}1,\ j = 0,1,2,\ldots,N{-}1)$. This quantizer is a 2D version of the dynamic quantizer for 1D plant (Azuma and Sugie (2008)). Also it is an extension of the usual static quantizer; in fact, if $\mathcal{C}_0 = \mathcal{C}_1 = \mathcal{C}_2 := 0$, Q is the same as the static quantizer, i.e., $v = \mathrm{q}[\,u\,]$.

In this paper, we consider a problem of finding an optimal dynamic quantizer with the following assumptions:

- **(A1)** The matrix C_1B is square and nonsingular.
- **(A2)** The matrix G is full row rank.
- **(A3)** For given $d \in \mathbb{R}_+$, the set $\mathbb{V}^m$ is defined as $\mathbb{V}^m := \{0, \pm d, \pm 2d, \ldots\}^m$. In addition, q is the nearest neighbor quantizer[1], i.e., $\mathrm{q}[\mu] := \arg\min_{v\in\mathbb{V}^m} (v-\mu)^\top (v-\mu)$ for $\mu \in \mathbb{R}^m$.

The assumptions **(A1)** and **(A2)** are given for P and K, under which the inverse of C_1B can be defined and the pseudo-inverse of G is given by $G^\dagger := G^\top (GG^\top)^{-1}$. These are the essential assumptions in this paper.

[1] Note that if the value of $\mathrm{q}[\mu]$ is not uniquely determined, $\mathrm{q}[\mu]$ is given as the smallest vector (in the sense of the sum of the all elements) of the solutions to $\min_{v\in\mathbb{V}^m} (v-\mu)^\top (v-\mu)$.

On the other hand, **(A3)** is imposed for q in Q from a practical point of view. The first half implies that the set $\mathbb{V}^m$ is a lattice whose interval is d in $\mathbb{R}^m$. The last half is given since the nearest quantizer is the simplest in the sense that the output can be obtained in short computation time.

It should be noticed that under **(A3)**, we have the relation

$$\mathrm{abs}(\mathrm{q}[\mu] - \mu) = \mathrm{abs}(\mu - \mathrm{q}[\mu]) \le \frac{d}{2}\mathbf{1} \quad (\forall \mu \in \mathbb{R}^m), \tag{6}$$

which will be used in this paper.

Before describing the problem discussed here, some symbols are defined. Given $M,\ N \in \mathbb{N}$, the reference sequence

$$R := \{r_{0,0},\ r_{0,1},\ldots,r_{0,(N-1)},\ r_{1,0},\ r_{1,1},\ \ldots,r_{1,(N-1)}, \ldots, r_{(M-1),0},\ldots,r_{(M-1),(N-1)}\} \in \mathbb{R}^{pMN} \tag{7}$$

is applied to the systems in **Fig. 1**, i.e., $r(i,j) = r_{i,j} \in \mathbb{R}^p$ $(i = 0,1,\ldots,M-1,\ j = 0,1,\ldots,N-1)$. For the system Σ_Q in **Fig. 1** (a) with the boundary condition $X(0) = X_0 \in \mathbb{R}^{n\times(M+N+1)}$ and the reference sequence $R \in \mathbb{R}^{pMN}$, let $Z_Q(X_0,R)$ be the output sequence from $(i,j) = (1,1)$ to $(i,j) = (M,N)$, and let $z_Q(i,j,X_0,R)$ be the output at (i,j). In addition, we consider the system Σ in **Fig. 1** (b), and for which, the output sequence is expressed by $Z(X_0,R)$ and the output at (i,j) is expressed by $z(i,j,X_0,R)$. Then, we denote by $Z(X_0,R) - Z_Q(X_0,R)$ the vector sequence of $z(i,j,X_0,R) - z_Q(i,j,X_0,R)$ from $(i,j) = (1,1)$ to $(i,j) = (M,N)$, and denote by $\|Z(X_0,R) - Z_Q(X_0,R)\|$ its ∞-norm, i.e.,

$$\|Z(X_0,R) - Z_Q(X_0,R)\| := \sup_{\substack{i\in\{1,2,\ldots,M\},\\ j\in\{1,2,\ldots,N\}}} \|z(i,j,X_0,R) - z_Q(i,j,X_0,R)\|. \tag{8}$$

Then the following problem is considered.

Problem 1. For the system Σ_Q, suppose that $M,\ N \in \mathbb{N}$ and $d \in \mathbb{R}_+$ are given, and we consider the maximum controlled output difference

$$E(Q) := \sup_{\substack{(X_0,R)\in\mathbb{R}^{n\times(M+N+1)}\\ \times\mathbb{R}^{pMN}}} \|Z(X_0,R) - Z_Q(X_0,R)\|. \tag{9}$$

Then,

(i) determine the value of $E(Q)$ for a given Q,
(ii) find a dynamic quantizer Q (i.e., $\mathcal{A}_0,\ \mathcal{A}_1,\ \mathcal{A}_2,\ \mathcal{B}_1,\ \mathcal{B}_2,\ \mathcal{C}_0,\ \mathcal{C}_1,\ \mathcal{C}_2$) minimizing $E(Q)$, and determine the minimum value of $E(Q)$.

In Problem 1, the two items (i) and (ii) correspond to an analysis problem and a design problem, respectively. The performance index $E(Q)$ represents the difference between Σ and Σ_Q in terms of the controlled output. Thus, if the minimum value of $E(Q)$ is sufficiently small, the output response of Σ_Q is close to that of Σ. This, for example, implies that if a good controller K for the ideal feedback system in Fig. 1 (b) is designed via conventional 2D system control theory, then the performance of (a) with the same K and the optimal Q is still good enough.

3. AN OPTIMAL DYNAMIC QUANTIZER

In this section, we provide a solution to Problem 1. Before solving the problem, the solution of the 2D system in (1)

is prepared.

$$\begin{aligned} x(\mu,\nu)=&\sum_{i=1}^{\mu-1} A^{\mu-i-1,\nu-1}\Big(A_0x(i,0)+A_2x(i+1,0)\Big)\\ &+\sum_{j=1}^{\nu-1} A^{\mu-1,\nu-j-1}\Big(A_0x(0,j)+A_1x(0,j+1)\Big)\\ &+A^{\mu-1,\nu-1}\Big(A_0x(0,0)+A_1x(0,1)+A_2x(1,0)\Big)\\ &+\sum_{i=0}^{\mu-1}\sum_{j=0}^{\nu-1} A^{\mu-i-1,\nu-j-1}Bu(i,j) \end{aligned} \tag{10}$$

where $A^{i,j}\in\mathbb{R}^{n\times n}$ is the transition matrix defined as

$$A^{i,j}:=\begin{cases} I & \text{if} \quad i=0 \ \text{ and } \ j=0,\\ 0 & \text{if} \quad i<0 \ \text{ or } \ j<0,\\ A_0A^{i-1,j-1}+A_1A^{i-1,j}+A_2A^{i,j-1} & \\ \qquad\qquad \text{otherwise.} & \end{cases} \tag{11}$$

Then, we derive a closed form expression of the performance $E(Q)$ as a solution to Problem 1 (i).

As a preliminary, let us derive the state space models of the systems Σ_Q and Σ in **Fig. 1**. For obtaining a state space representation of Σ_Q, we first rewrite the quantizer (4) to an equivalent linear system. The equation (4) can be rewritten by

$$Q:\begin{cases} \xi(i+1,j+1)=\mathcal{A}_0\xi(i,j)+\mathcal{A}_1\xi(i,j+1)+\mathcal{A}_2\xi(i+1,j)\\ \qquad\qquad +\mathcal{B}_1u(i,j)+\mathcal{B}_2v(i,j),\\ v(i,j)=\mathcal{C}_0\xi(i,j)+\mathcal{C}_1\xi(i,j+1)\\ \qquad\qquad +\mathcal{C}_2\xi(i+1,j)+u(i,j)+w(i,j) \end{cases} \tag{12}$$

where the variable $w\in\mathbb{R}^m$ is defined as

$$w(i,j):=\mathrm{q}\,[\,\tilde{u}(i,j)\,]-\tilde{u}(i,j) \tag{13}$$

for

$$\tilde{u}(i,j):=\mathcal{C}_0\xi(i,j)+\mathcal{C}_1\xi(i,j+1)+\mathcal{C}_2\xi(i+1,j)+u(i,j). \tag{14}$$

Notice that $w(i,j)$ expresses the quantization error generated by the static quantizer q in Q, and that $w(i,j)\in[-d/2,\ d/2]^m$.

By using (1), (3), and (12), the system Σ_Q is expressed as

$$\begin{cases} \begin{bmatrix}x(i+1,j+1)\\ \xi(i+1,j+1)\end{bmatrix}=\big(\bar{A}_0+\bar{B}_1\bar{F}_0C_2\big)\begin{bmatrix}x(i,j)\\ \xi(i,j)\end{bmatrix}\\ \qquad\qquad +\big(\bar{A}_1+\bar{B}_1\bar{F}_1C_2\big)\begin{bmatrix}x(i,j+1)\\ \xi(i,j+1)\end{bmatrix}\\ \qquad\qquad +\big(\bar{A}_2+\bar{B}_1\bar{F}_2C_2\big)\begin{bmatrix}x(i+1,j)\\ \xi(i+1,j)\end{bmatrix}\\ \qquad\qquad +\left(\begin{bmatrix}BG\\ 0\end{bmatrix}+\bar{B}_1G\right)r(i,j)+\bar{B}_2w(i,j),\\ z_Q(i,j)=\bar{C}\begin{bmatrix}x(i,j)\\ \xi(i,j)\end{bmatrix} \end{cases} \tag{15}$$

for

$$\begin{aligned} &\bar{A}_0:=\begin{bmatrix}A_0^* & B\mathcal{C}_0\\ 0 & \mathcal{A}_0+\mathcal{B}_2\mathcal{C}_0\end{bmatrix},\quad \bar{A}_1:=\begin{bmatrix}A_1^* & B\mathcal{C}_1\\ 0 & \mathcal{A}_1+\mathcal{B}_2\mathcal{C}_1\end{bmatrix},\\ &\bar{A}_2:=\begin{bmatrix}A_2^* & B\mathcal{C}_2\\ 0 & \mathcal{A}_2+\mathcal{B}_2\mathcal{C}_2\end{bmatrix},\\ &\bar{B}_1:=\begin{bmatrix}0\\ \mathcal{B}_1+\mathcal{B}_2\end{bmatrix},\quad \bar{B}_2:=\begin{bmatrix}B\\ \mathcal{B}_2\end{bmatrix},\quad \bar{C}:=[C_1\ 0],\\ &\bar{F}_0:=[F_0\ 0],\quad \bar{F}_1:=[F_1\ 0],\quad \bar{F}_2:=[F_2\ 0] \end{aligned} \tag{16}$$

where A_0^*, A_1^*, and A_2^* are given by

$$\begin{aligned} &A_0^*:=A_0+BF_0C_2,\qquad A_1^*:=A_1+BF_1C_2,\\ &A_2^*:=A_2+BF_2C_2. \end{aligned} \tag{17}$$

Here, let $\bar{x}(i,j):=[\,x(i,j)^\top\ \xi(i,j)^\top\,]^\top\in R^{2n}$, and we define $\bar{A}^{i,j}\in\mathbb{R}^{2n\times 2n}$ as $A^{i,j}$ in (11) for $A_0:=\bar{A}_0+\bar{B}_1\bar{F}_0C_2$, $A_1:=\bar{A}_1+\bar{B}_1\bar{F}_1C_2$, and $A_2:=\bar{A}_2+\bar{B}_1\bar{F}_2C_2$. Then, for given $\mu,\nu\in\mathbb{N}$, $X_0\in\mathbb{R}^{n\times(\mu+\nu+1)}$, and $R\in\mathbb{R}^{p\mu\nu}$, the output $z_Q(\mu,\nu,X_0,R)$ of Σ_Q at $(i,j):=(\mu,\nu)$ is described as

$$\begin{aligned} &z_Q(\mu,\nu,X_0,R)\\ &=\sum_{i=1}^{\mu-1}\bar{C}\bar{A}^{\mu-i-1,\nu-1}\Big((\bar{A}_0+\bar{B}_1\bar{F}_0C_2)\bar{x}(i,0)\\ &\qquad\qquad +(\bar{A}_2+\bar{B}_1\bar{F}_2C_2)\bar{x}(i+1,0)\Big)\\ &\quad+\sum_{j=1}^{\nu-1}\bar{C}\bar{A}^{\mu-1,\nu-j-1}\Big((\bar{A}_0+\bar{B}_1\bar{F}_0C_2)\bar{x}(0,j)\\ &\qquad\qquad +(\bar{A}_1+\bar{B}_1\bar{F}_1C_2)\bar{x}(0,j+1)\Big)\\ &\quad+\bar{C}\bar{A}^{\mu-1,\nu-1}(\bar{A}_0+\bar{B}_1\bar{F}_0C_2)\bar{x}(0,0)\\ &\quad+\bar{C}\bar{A}^{\mu-1,\nu-1}(\bar{A}_1+\bar{B}_1\bar{F}_1C_2)\bar{x}(0,1)\\ &\quad+\bar{C}\bar{A}^{\mu-1,\nu-1}(\bar{A}_2+\bar{B}_1\bar{F}_2C_2)\bar{x}(1,0)\\ &\quad+\sum_{i=0}^{\mu-1}\sum_{j=0}^{\nu-1}\bar{C}\bar{A}^{\mu-i-1,\nu-j-1}\\ &\qquad\qquad\cdot\left\{\left(\begin{bmatrix}BG\\ 0\end{bmatrix}+\bar{B}_1G\right)r(i,j)+\bar{B}_2w(i,j)\right\}. \end{aligned} \tag{18}$$

On the other hand, by using (17), Σ is represented as

$$\Sigma:\begin{cases} x(i+1,j+1)=A_0^*\,x(i,j)+A_1^*\,x(i,j+1)\\ \qquad\qquad +A_2^*\,x(i+1,j)+BGr(i,j),\\ z_Q(i,j)=C_1x(i,j). \end{cases} \tag{19}$$

Furthermore, we define $(A^*)^{i,j}\in\mathbb{R}^{n\times n}$ as $A^{i,j}$ in (11) for $A_0:=A_0^*$, $A_1:=A_1^*$, and $A_2:=A_2^*$. Then, the output $z(\mu,\nu,X_0,R)$ of Σ is given by

$$\begin{aligned} &z(\mu,\nu,X_0,R)\\ &=\sum_{i=1}^{\mu-1}C_1(A^*)^{\mu-i-1,\nu-1}\Big(A_0^*x(i,0)+A_2^*x(i+1,0)\Big)\\ &\quad+\sum_{j=1}^{\nu-1}C_1(A^*)^{\mu-1,\nu-j-1}\Big(A_0^*x(0,j)+A_1^*x(0,j+1)\Big)\\ &\quad+C_1(A^*)^{\mu-1,\nu-1}\Big(A_0^*x(0,0)+A_1^*x(0,1)+A_2^*x(1,0)\Big)\\ &\quad+\sum_{i=0}^{\mu-1}\sum_{j=0}^{\nu-1}C_1(A^*)^{\mu-i-1,\nu-j-1}BGr(i,j). \end{aligned} \tag{20}$$

The difference between $z(\mu,\nu,X_0,R)$ and $z_Q(\mu,\nu,X_0,R)$ is given by

$$z(\mu,\nu,X_0,R)-z_Q(\mu,\nu,X_0,R)$$

$$= W(X_0, R) - \sum_{i=0}^{\mu-1}\sum_{j=0}^{\nu-1} \bar{C}\bar{A}^{\mu-i-1,\nu-j-1}\bar{B}_2 w(i,j) \quad (21)$$

where $W(X_0, R)$ describes all the terms depending on X_0 and R.

Then the following result is obtained as a solution to Problem 1 (i).

Theorem 1. For the system Σ_Q, suppose that M, $N \in \mathbb{N}$ and $d \in \mathbb{R}_+$ are given, and assume **(A2)** and **(A3)**. If

$$\bar{C}\bar{A}^{i,j}\bar{B}_1 = 0 \quad (\forall (i,j) \in \{0,1,\ldots,M-1\}\times\{0,1,\ldots,N-1\}), \quad (22)$$

then

$$E(Q) = \left\| \sum_{i=0}^{M-1}\sum_{j=0}^{N-1} \mathrm{abs}\left(\bar{C}\bar{A}^{i,j}\bar{B}_2\right) \right\| \frac{d}{2}; \quad (23)$$

otherwise

$$E(Q) = \infty. \quad (24)$$

Proof. See Appendix A.

Theorem 1 provides a closed form expression of the maximum output difference $E(Q)$. In the right hand side of (23), each element of the matrix $\mathrm{abs}(\bar{C}\bar{A}^{i,j}\bar{B}_2)$ is nonnegative for every i, $j \in \{0\}\cup\mathbb{N}$, and the term $\bar{C}\bar{A}^{0,0}\bar{B}_2 = C_1 B$ does not include the parameters of Q. This means that if there exists a dynamic quantizer Q satisfying the relations (22) and

$$\bar{C}\bar{A}^{i,j}\bar{B}_2 = 0 \quad (\forall (i,j) \in \{0,1,\ldots,M-1\}\times\{0,1,\ldots,N-1\} \text{ except for } (i,j) = (0,0)), \quad (25)$$

such a Q is a solution to Problem 1 (ii).

Theorem 2. For the system Σ_Q, suppose that M, $N \in \mathbb{N}$ and $d \in \mathbb{R}_+$ are given, and assume **(A1)**–**(A3)**. Then, an optimal dynamic quantizer and the minimum value of $E(Q)$ are given by

$$Q^{\mathrm{OPT}}: \begin{cases} \mathcal{A}_0 := A_0^*, \quad \mathcal{A}_1 := A_1^*, \quad \mathcal{A}_2 := A_2^*, \\ \mathcal{B}_1 := -B, \quad \mathcal{B}_2 := B, \\ \mathcal{C}_0 := -(C_1 B)^{-1} C_1 A_0^*, \\ \mathcal{C}_1 := -(C_1 B)^{-1} C_1 A_1^*, \\ \mathcal{C}_2 := -(C_1 B)^{-1} C_1 A_2^*, \end{cases} \quad (26)$$

and

$$E(Q^{\mathrm{OPT}}) = \left\| \mathrm{abs}\left(C_1 B\right) \right\| \frac{d}{2} \quad (27)$$

where A_0^*, A_1^*, and A_2^* are given by (17).

4. GENERATION OF BINARY HALFTONE IMAGES

Digital halftoning is a process of transforming grayscale images (multi-level images) to monochrome images (binary images), and which is necessary for display of multi-level images in media in which the direct rendition of the tone is impossible. Also, this technique is used for compressing images since the storage capacity of binary images is smaller than that of multi-level images.

In this section, we apply an optimal dynamic quantizer for 2D systems to generate a halftone image. More concretely, for the system shown in **Fig. 2** (a), we regard the input r of Q, the output v, and the plant P as an original grayscale image, a halftone image, and a human visual system, respectively. So the output z of P corresponds to an image obtained by viewing a halftone image v, i.e., an image is perceived by human eyes. On the other hand, **Fig. 2** (b) is the ideal system with continuous-valued input, and on which the output z corresponds to an image obtained by a sight of an original image r. For Q, we use the optimal dynamic quantizer Q^{OPT} such that the system in **Fig. 2** (a) optimally approximates the system in (b). Hence, a halftone image generated by Q^{OPT} closely resembles an original image in appearance.

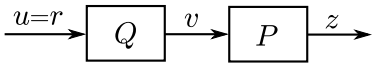
In this paper, the system P is given by

$$P: \begin{cases} A_0 := \begin{bmatrix} 0.2 & 0 \\ 0 & 0 \end{bmatrix}, \ A_1 := \begin{bmatrix} 0.3 & 0 \\ 0 & 1 \end{bmatrix}, \\ A_2 := \begin{bmatrix} 0.3 & 0.1 \\ 0 & 0 \end{bmatrix}, \ B := \begin{bmatrix} 0.1 \\ 0 \end{bmatrix}, \ C_1 := \begin{bmatrix} 1 & 0 \end{bmatrix}. \end{cases} \quad (28)$$

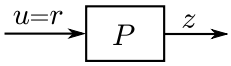
This model has the spatial low pass property, and which is given by identifying the visual system of the first author in this paper [2]. Note here that the state $x(i,j)$ of P is defined as $x(i,j) := [\tilde{x}(i,j) \ \tilde{x}(i+1,j)]^\top \in \mathbb{R}^2$ where $\tilde{x}(i,j) \in \mathbb{R}$ corresponds to the output $z_Q(i,j) \in \mathbb{R}$ of P, i.e., $\tilde{x}(i,j) = z_Q(i,j)$. Then, the optimal dynamic quantizer is given by (26) where A_0^*, A_1^*, and A_2^* are given by (17) with $F_0 = F_1 = F_2 := 0$. Note that if $F_0 = F_1 = F_2 := 0$ and $G := I$, the system in **Fig. 1** implies the system in **Fig. 2**. In addition, we use the image shown in **Fig. 3** as the original 8-bit image whose the minimum and maximum luminance values equal to 0 and 255, respectively, and we set $d := 255$. Then, using the optimal dynamic quantizer, we have the halftone image shown in **Fig. 4**. It is remarked that the picture in **Fig. 4** looks like the grayscale image although **Fig. 4** is the binary image composed of 0 and 255, i.e., black and white pixels; see **Fig. 5** (left and center figures).

Furthermore, we compare with another halftone image obtained by a conventional halftoning process called error diffusion algorithm (Eschbach et al. (2003)). **Fig. 6** shows the halftone image generated by Floyd & Steinberg filter which is one of error diffusion algorithms. Compared **Fig. 4** with **Fig. 6**, we can verify that our result in **Fig. 4** is similar to the result obtained by the error diffusion algorithm. Here, to evaluate the two images quantitatively, we use WSNR (Kite, et al. (2000)) which is one of the useful metrics of halftone images. The WSNR of **Fig. 4** and that of **Fig. 6** are 27.820[dB] and 27.816[dB], respectively, which means that there is not so much of a difference between **Fig. 4** and **Fig. 6**. On the other hand, the value of $E(Q)$ for **Fig. 4** is 12.75 and that for **Fig. 6** is 27.58. From this, the proposed method has a superiority over the error diffusion algorithm in the sense of $E(Q)$. This reason is that the proposed quantizer is optimized for the visual model in (28). Finally, the optimal dynamic quantizer has the parameters of P. So we can construct another optimal quantizer, which has a different property from the quantizer used in this paper, by using another model P such as printer models. Namely, our dynamic quantizer has a flexibility of designing model based halftone filters, while the error diffusion filter does not have it. Therefore, we conclude that the optimal dynamic quantizer is very useful in halftone image processing.

[2] By using a circular zone plate image, we have estimated a spatial cutoff frequency.



(a) Feedforward system with an input quantizer



(b) Ideal feedforward system

Fig. 2. Two feedforward systems

Fig. 3. Original image (lena)

Fig. 4. Halftone image by the proposed method

Fig. 5. Closeup images of the original image (**Fig. 3**) and two halftone images (**Figs. 4, 6**)

Fig. 6. Halftone image by the error diffusion method

5. CONCLUSION

In this paper, we have considered analysis and optimization problems of dynamic quantizers for a class of 2D systems with discrete-valued input. First, as a solution to the analysis problem, we derive a closed form expression of the performance of dynamic quantizers, which is represented by the parameters of P, K, and Q. Next, based on the result, we have proposed an optimal dynamic quantizer. Finally, we have applied the optimal dynamic quantizer to the generation of binary halftone images, and we have verified that the optimal dynamic quantizer for 2D systems is applicable to image processing.

REFERENCES

G. N. Nair and R. J. Evans: Exponential stabilizability of finite-dimensional linear systems with limited data rates, *Automatica*, Vol. 39, No. 4, pp. 583–593, 2003

S. Tatikonda and S. Mitter: Control under communication constraints, *IEEE Transactions on Automatic Control*, Vol. 49, No. 7, pp. 1056–1068, 2004

F. Bullo and D. Liberzon: Quantized control via locational optimization, *IEEE Transactions on Automatic Control*, Vol. 51, No. 1, pp. 2–13, 2006

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

S. Azuma and T. Sugie: Optimal dynamic quantizers for discrete-valued input control, *Automatica*, Vol. 44, No. 2, pp. 396–406, 2008

Y. Minami, S. Azuma, and T. Sugie: An optimal dynamic quantizer for feedback control with discrete-valued signal constraints, *Proceedings of the 46th IEEE Conference on Decision and Control*, pp. 2259–2264, 2007

Robert P. Roesser: A discrete state-space model for linear image processing, *IEEE Transactions on Automatic Control*, Vol. 20, No. 1, pp. 1–10, 1975

K. Galkowski, E. Rogers, and D. H. Owens: New 2D models and a transition matrix for discrete linear repetitive processes, *International Journal of Control*, Vol. 72, No. 15, pp. 1365–1380, 1999

E. Fornasini and G. Marchesini: Doubly-indexed dynamical systems: state-space models and structural properties, *Math. systems theory*, Vol. 12, pp. 59–72, 1978

R. Eschbach, Z. Fan, K. T. Knox, and G. Marcu: Threshold modulation and stability in error diffusion, *IEEE Signal processing magazine*, pp. 39–50, 2003

T. D. Kite, B. L. Evans, and A. C. Bovik: Modeling and quality assessment of halftoning by error diffusion, *IEEE Transactions on Image Processing*, Vol. 9, No. 5, pp. 909–922, 2000

Appendix A. PROOF OF THEOREM 1

For proving (23) and (24), the following lemma, which is an extension of the result in (Azuma and Sugie (2008)), is prepared.

Lemma 1. Suppose that matrices $H^{i,j} \in \mathbb{R}^{m\times m}$, vectors $z_{i,j} \in \mathbb{R}^m$ ($i \in \{0,1,\ldots,\mu-1\}$, $j \in \{0,1,\ldots,\nu-1\}$), and a positive number $\zeta \in \mathbb{R}_+$ are given. Then, the following statements hold.

(i) If $\mathrm{abs}(z_{i,j}) \le \zeta\mathbf{1}$, then

$$\left\|\sum_{i=0}^{\mu-1}\sum_{j=0}^{\nu-1} H^{i,j} z_{i,j}\right\| \le \left\|\sum_{i=0}^{\mu-1}\sum_{j=0}^{\nu-1} \mathrm{abs}(H^{i,j})\right\| \zeta. \tag{A.1}$$

(ii) Let $H^{i,j}_{\alpha\beta}$ and $\langle H^{i,j}\rangle_\alpha$ be the (α,β)-th element and the α-th row vector of $H^{i,j}$, respectively, and let

$$\alpha' := \arg\max_{\alpha\in\{1,2,\ldots,m\}} \sum_{\beta=1}^{m}\sum_{i=0}^{\mu-1}\sum_{j=0}^{\nu-1} \left|H^{i,j}_{\alpha\beta}\right|. \tag{A.2}$$

If $z_{i,j} := \mathrm{sign}(\langle H^{i,j}\rangle_{\alpha'})^\top \zeta$, the equality holds in (A.1).

By Lemma 1, (23) and (24) are proven as follows.

Proof of (23): Since $W(X_0,R) = 0$ holds in (21) if the relation (22) holds, we have

$$\begin{aligned}&\|z(\mu,\nu,X_0,R) - z_Q(\mu,\nu,X_0,R)\| \\ &\quad= \left\|\sum_{i=0}^{\mu-1}\sum_{j=0}^{\nu-1} \bar{C}\bar{A}^{\mu-i-1,\nu-j-1}\bar{B}_2 w(i,j)\right\|\end{aligned} \tag{A.3}$$

for given μ, $\nu \in \mathbb{N}$, a boundary condition $X_0 \in \mathbb{R}^{n\times(M+N+1)}$, and a reference sequence $R \in \mathbb{R}^{pMN}$. From (6), (13), and Lemma 1 (i), the following inequality holds under **(A3)**.

$$\begin{aligned}&\|z(\mu,\nu,X_0,R) - z_Q(\mu,\nu,X_0,R)\| \\ &\le \left\|\sum_{i=0}^{\mu-1}\sum_{j=0}^{\nu-1} \mathrm{abs}(\bar{C}\bar{A}^{\mu-i-1,\nu-j-1}\bar{B}_2)\right\| \frac{d}{2}.\end{aligned} \tag{A.4}$$

Now, let $H^{i,j} := \bar{C}\bar{A}^{\mu-i-1,\nu-j-1}\bar{B}_2$, and we define α' by (A.2). For $r(i,j) := r_{i,j}$, we define $\tilde{u}_{i,j}$ in a similar way to $\tilde{u}(i,j)$ in (14) with (1) and (3), that is,

$$\tilde{u}_{i,j} := \mathcal{C}\xi_{i,j} + FC_2x_{i,j} + Gr_{i,j} \tag{A.5}$$

where

$$\mathcal{C} := [\mathcal{C}_0\ \mathcal{C}_1\ \mathcal{C}_2] \in \mathbb{R}^{m\times 3n},\ F := [F_0\ F_1\ F_2] \in \mathbb{R}^{m\times 3l_2},$$

$$\xi_{i,j} := \begin{bmatrix}\xi(i,j)\\ \xi(i,j+1)\\ \xi(i+1,j)\end{bmatrix} \in \mathbb{R}^{3n},\ x_{i,j} := \begin{bmatrix}x(i,j)\\ x(i,j+1)\\ x(i+1,j)\end{bmatrix} \in \mathbb{R}^{3n}.$$

Let $w_{i,j}$ be described as

$$w_{i,j} = \mathrm{q}[\tilde{u}_{i,j}] - \tilde{u}_{i,j}, \tag{A.6}$$

and we consider the reference defined as

$$\begin{aligned}r_{i,j} := G^\dagger\Big\{ &-\mathrm{sign}(\langle H^{i,j}\rangle_{\alpha'})^\top \left(\frac{d}{2}-\delta\right) \\ &- \mathcal{C}\xi_{i,j} - FC_2x_{i,j} + \mathrm{q}\big[\mathcal{C}\xi_{i,j} + FC_2x_{i,j}\big]\Big\}\end{aligned} \tag{A.7}$$

for an arbitrarily given small number $\delta \in (0, d/2)$. Then under **(A2)**, we have

$$w_{i,j} = \mathrm{sign}\big(\langle H^{i,j}\rangle_{\alpha'}\big)^\top \left(\frac{d}{2}-\delta\right) \tag{A.8}$$

because

$$\begin{aligned}\mathrm{q}[\tilde{u}_{i,j}] &= \mathrm{q}\Big[-\mathrm{sign}(\langle H^{i,j}\rangle_{\alpha'})^\top \left(\frac{d}{2}-\delta\right) \\ &\qquad + \mathrm{q}[\mathcal{C}\xi_{i,j} + FC_2x_{i,j}]\Big] \\ &= \mathrm{q}[\mathcal{C}\xi_{i,j} + FC_2x_{i,j}].\end{aligned} \tag{A.9}$$

Thus it follows from Lemma 1 (ii) that the relation

$$\begin{aligned}&\|z(\mu,\nu,X_0,R) - z_Q(\mu,\nu,X_0,R)\| \\ &= \left\|\sum_{i=0}^{\mu-1}\sum_{j=0}^{\nu-1} \mathrm{abs}(\bar{C}\bar{A}^{\mu-i-1,\nu-j-1}\bar{B}_2)\right\| \left(\frac{d}{2}-\delta\right)\end{aligned} \tag{A.10}$$

holds for $R \in \mathbb{R}^{pMN}$ defined by $r_{i,j}$ in (A.7). Moreover, for every (sufficiently small) $\epsilon \in \mathbb{R}_+$, there exists $\delta \in (0, d/2)$ satisfying

$$\begin{aligned}&\left\|\sum_{i=0}^{\mu-1}\sum_{j=0}^{\nu-1} \mathrm{abs}(\bar{C}\bar{A}^{\mu-i-1,\nu-j-1}\bar{B}_2)\right\| \frac{d}{2} - \epsilon \\ &\le \left\|\sum_{i=0}^{\mu-1}\sum_{j=0}^{\nu-1} \mathrm{abs}(\bar{C}\bar{A}^{\mu-i-1,\nu-j-1}\bar{B}_2)\right\| \left(\frac{d}{2}-\delta\right).\end{aligned} \tag{A.11}$$

Hence from this and (A.4), we obtain

$$\begin{aligned}&\sup_{(X_0,R)} \|z(\mu,\nu,X_0,R) - z_Q(\mu,\nu,X_0,R)\| \\ &\qquad= \left\|\sum_{i=0}^{\mu-1}\sum_{j=0}^{\nu-1} \mathrm{abs}(\bar{C}\bar{A}^{\mu-i-1,\nu-j-1}\bar{B}_2)\right\| \frac{d}{2}.\end{aligned} \tag{A.12}$$

Since each element of the matrix $\mathrm{abs}(\bar{C}\bar{A}^{\mu-i-1,\nu-j-1}\bar{B}_2)$ is nonnegative and the value of the right hand side of (A.12) is a monotone nondecreasing function with respect to μ, $\nu \in \mathbb{N}$, the following relation holds.

$$\begin{aligned}&\sup_{\substack{\mu\in\{1,2,\ldots,M\},\\ \nu\in\{1,2,\ldots,N\}}}\ \sup_{(X_0,R)} \|z(\mu,\nu,X_0,R) - z_Q(\mu,\nu,X_0,R)\| \\ &\qquad= \left\|\sum_{i=0}^{M-1}\sum_{j=0}^{N-1} \mathrm{abs}(\bar{C}\bar{A}^{i,j}\bar{B}_2)\right\| \frac{d}{2}.\end{aligned} \tag{A.13}$$

This proves (23).

Proof of (24): Applying Lemma 1 (i) to (21) under **(A3)**, we have the following relation.

$$\begin{aligned}\|W(X_0,R)\| - &\left\|\sum_{i=0}^{\mu-1}\sum_{j=0}^{\nu-1} \mathrm{abs}(\bar{C}\bar{A}^{\mu-i-1,\nu-j-1}\bar{B}_2)\right\| \frac{d}{2} \\ &\le \|z(\mu,\nu,X_0,R) - z_Q(\mu,\nu,X_0,R)\|.\end{aligned} \tag{A.14}$$

Here, let μ', ν' be defined by smallest integers μ, $\nu \in \mathbb{N}$ satisfying $\bar{C}\bar{A}^{\mu-1,\nu-1}\bar{B}_1 \ne 0$ (note that there exists a pair (μ',ν') since (22) does not hold), and recall that $W(X_0,R)$ in (21) describes the terms depending on X_0 and R; see (18) and (20). According to the term depending on $r(0,0)$ in $W(X_0,R)$, we have

$$\begin{aligned}&C_1(A^*)^{\mu'-1,\nu'-1}BGr(0,0) \\ &\qquad - \bar{C}\bar{A}^{\mu'-1,\nu'-1}\left(\begin{bmatrix}BG\\0\end{bmatrix} + \bar{B}_1G\right) r(0,0) \\ &= -\bar{C}\bar{A}^{\mu'-1,\nu'-1}\bar{B}_1Gr(0,0).\end{aligned} \tag{A.15}$$

Since G is the full row rank under **(A2)**, the relation $\bar{C}\bar{A}^{\mu'-1,\nu'-1}\bar{B}_1G \ne 0$ holds. Hence the supremum of $\|W(X_0,R)\|$ with respect to $R \in \mathbb{R}^{pMN}$ is infinity, i.e., the left hand side of (A.14) with respect to R is infinity. This implies $\sup_{(X_0,R)} \|z(\mu',\nu',X_0,R) - z_Q(\mu',\nu',X_0,R)\| = \infty$, which leads to (24).