stability and robustness analysis problems to nondifferentiable convex programs. They have also provided efficient numerical methods to solve the resulting convex programs. In particular, the results in [2] can unambiguously determine whether an uncertain system with scalar uncertain diagonal blocks is $QS(\Delta_R)$. Some of these ideas can also be extended to synthesis problems. See, for example, Bernussou *et al.* [1] and Packard and Becker [9] for the solution of a state-feedback quadratic stabilizability problem, in the presence of real uncertain parameters, via convex programming.

References

[1] J. Bernussou, P. L. D. Peres, and J. C. Geromel, "An LP oriented procedure for quadratic stabilization of uncertain systems," *Syst. Contr. Lett.*, vol. 13, no. 3, pp. 65–72, 1989.

[2] S. P. Boyd and Q. Yang, "Structured and simultaneous Lyapunov functions for system stability problems," *Int. J. Contr.*, vol. 49, no. 6, pp. 2215–2240, 1989.

[3] P. Dorato, Ed., *Robust Control.* New York: IEEE Press, 1987.

[4] P. Dorato and R. K. Yedavalli, Ed., *Recent Advances in Robust Control. A Volume of Selected Conference Papers.* New York: IEEE Press, 1990.

[5] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, "State-space solutions to standard $\mathscr{H}_2$ and $\mathscr{H}_\infty$ control problems," *IEEE Trans. Automat. Contr.*, vol. 34, no. 8, pp. 831–847, 1989.

[6] D. Hinrichsen and A. J. Pritchard, "Stability radii of linear systems," *Syst. Contr. Lett.*, vol. 7, pp. 1–10, 1986.

[7] P. P. Khargonekar, I. R. Petersen, and K. Zhou, "Robust stabilization of uncertain linear systems: quadratic stabilizability and $\mathscr{H}_\infty$ control theory," *IEEE Trans. Automat. Contr.*, vol. 35, no. 3, pp. 356–361, 1990.

[8] A. Packard, Ph.D. dissertation, Univ. of California, Berkeley, 1988.

[9] A. Packard and G. Becker, "State-feedback solutions to quadratic stabilization," in *Proc. 28th Annual Allerton Conf. Communication, Contr., Comput.*, 1990, pp. 768–769.

[10] A. Packard and J. C. Doyle, "Quadratic stability with real and complex perturbations," *IEEE Trans. Automat. Contr.*, vol. 35, no. 2, pp. 198–201, 1990.

[11] M. A. Rotea, "Quadratic stability and $\mathscr{H}_\infty$ norm tests," in preparation.

[12] K. Zhou and P. P. Khargonekar, "An algebraic Riccati equation approach to $\mathscr{H}_\infty$ optimization," *Syst. Contr. Lett.*, vol. 11, pp. 85–91, 1988.

Comparative Study of Finite Wordlength Effects in Shift and Delta Operator Parameterizations

Gang Li and Michel Gevers

Abstract—**This note analyzes the sensitivity of transfer functions w.r.t. finite wordlength effect errors in the implementation of the coefficients of both shift operator and delta operator parameterizations. The relationships between optimal realization sets in shift and delta operator are established. It is shown that the optimal realizations in delta operator have better performance than those in shift operator when the poles of the systems are clustered around $z = +1$. A numerical example is given.**

I. Introduction

In the last few years, both Peterka [1] and Middleton and Goodwin ([2], [3]) have promoted the use of the delta operator as opposed to the shift operator in estimation and control applications. Two major advantages are claimed for the use of delta operator parameterizations: a theoretically interesting unified formulation of continuous-time and discrete-time control theory which entails a better understanding of discrete-time control under fast sampling, and a range of practically interesting numerical advantages connected with finite wordlength (FWL) effects. One problem not studied in [3] is that of comparing the sensitivity of the transfer function of a state-variable model w.r.t. coefficient errors in the (A, B, C) state-space matrices when the state-variable model is implemented in either a shift-operator parameterization or a delta-operator parameterization. This problem, which of course is of interest in FWL implementations, is the object of the present note.

The effect of FWL errors in the state space matrices (A, B, C) on the transfer function has been studied by various authors ([3]–[6]), and has been extended to the effect of FWL errors in the regulator coefficients on the closed-loop transfer function [7], [10]. This has led to a commonly accepted measure for the sensitivity of a transfer function w.r.t. to the coefficients of (A, B, C) (see [4]–[6]), and to the search for optimal realizations $(A^{\text{opt}}, B^{\text{opt}}, C^{\text{opt}})$, among the equivalence class $(T^{-1}AT, T^{-1}B, CT)$ of similarity transforms, that minimize this sensitivity. This problem has been solved by Thiele ([6], [12]). In [13], a frequency weighted sensitivity minimization problem has been investigated. A range of other sensitivity minimization problems have been solved in [8]. All of these results relate solely to shift operator state space representations.

Here we first define in Section II the delta operator representations of a system. The relationships between shift and delta operator parameterizations, both in transfer function and in state-space form, are established. In Section III we show that the set of optimal delta realizations can be connected in a simple way and therefore derived from the set of optimal shift realizations. It is then shown that, by a proper choice of the degree of freedom available in the definition of the delta operator, the sensitivity with the delta operator state space models can be made smaller than that with shift operator state space models. In Section IV, the comparison between shift and delta operators is illustrated by a numerical example. Some concluding remarks are given in Section V.

II. Delta Operator Parameterizations

Throughout this note we consider scalar strictly proper time invariant discrete-time transfer functions. In the old days (i.e., before Middleton and Goodwin [3]) it was customary to represent such transfer functions as follows:

$$H(z) = \frac{\sum_1^n b_i z^{n-i}}{z^n + \sum_1^n a_i z^{n-i}} = \frac{\sum_1^n b_i z^{-i}}{1 + \sum_1^n a_i z^{-i}}. \tag{2.1}$$

Such discrete-time transfer functions are often obtained from a continuous-time transfer function $H_s(s)$ as the result of a discretization procedure with a sampling period T_s. It has been shown [3] that when fast sampling is used better numerical properties can be achieved by reparameterizing (2.1) with z replaced by $(z-1)/T_s$. Here, we will consider that the starting point is a discrete-time (rather than continuous-time) transfer function. Motivated by [3], we shall introduce the following definition for the δ-operator:

$$\delta \triangleq \frac{z-1}{\Delta}, \tag{2.2}$$

Manuscript received February 28, 1990; revised October 7, 1991. Paper recommended by Associate Editor, D. F. Delchamps.

M. Gevers is with Laboratoire d'Automatique et d'Analyse des Systèmes, Louvain University, Bâtiment Maxwell, B-1348 Louvain-la-Neuve, Belgium.

G. Li is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 2263.

IEEE Log Number 9205256.

where Δ is any positive number, not necessarily a sampling period. Thus (2.2) should be seen purely as a linear operator. Discussions about the implementation of the δ-operator can be found in [3], [8].

With the definition (2.2) for δ, the transfer function $H(z)$ of (2.1) can be reexpressed in δ-form as follows:

$$H(z) = \frac{\sum_1^n b_i z^{n-i}}{z^n + \sum_1^n a_i z^{n-i}} = \frac{\sum_1^n \beta_i \delta^{n-i}}{\delta^n + \sum_1^n \alpha_i \delta^{n-i}} \triangleq H_\delta(\delta). \quad (2.3)$$

Our aim in this note will then be to compare the sensitivity of the transfer functions $H(z)$ and $H_\delta(\delta)$, respectively, w.r.t. numerical errors in the coefficients of their respective state-space implementations. The choice of a value for Δ and its role in improving these sensitivities will be a central feature of our note.

We note that the coefficients $\{\alpha_i, \beta_i\}$ are obtained from the $\{a_i, b_i\}$ by substituting $z = 1 + \Delta\delta$ in $H(z)$. This yields the following relationships:

$$\bar{\beta} = \begin{pmatrix} 0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} = T \begin{pmatrix} 0 \\ b_1 \\ \vdots \\ b_n \end{pmatrix}, \quad \bar{\alpha} = \begin{pmatrix} 1 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = T \begin{pmatrix} 1 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} \quad (2.4)$$

where

$$T = \begin{pmatrix} 1 & 0 & & \cdots & 0 \\ t_{21} & t_{22} & 0 & \cdots & 0 \\ t_{31} & t_{32} & t_{33} & & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ t_{n+1,1} & t_{n+1,2} & & & t_{n+1,n+1} \end{pmatrix} \quad (2.5a)$$

with

$$t_{ij} = C_{n+1-j}^{i-j} \Delta^{-(i-1)}, \quad i \geq j; \qquad C_m^i = \frac{m!}{(m-i)!i!}. \quad (2.5b)$$

Going back to (2.3), we observe that $H(z)$ and $H_\delta(\delta)$ are two different but equivalent parameterizations representing the same object. These two input–output relationships can be represented by a shift-operator (respectively δ-operator) state-space model as follows:

$$z x_t^{(1)} = A_z x_t^{(1)} + B_z u_t \quad (2.6a)$$

$$y_t = C_z x_t^{(1)} \quad (2.6b)$$

and

$$\delta x_t^{(2)} = A_\delta x_t^{(2)} + B_\delta u_t \quad (2.7a)$$

$$y_t = C_\delta x_t^{(2)}. \quad (2.7b)$$

The following relationships relate the internal and external representations:

$$H(z) = C_z(zI - A_z)^{-1}B_z, \qquad H_\delta(\delta) = C_\delta(\delta I - A_\delta)^{-1}B_\delta. \quad (2.8)$$

For future use, we shall introduce the notion of a realization set S_ρ. We define:

$$S_\rho \triangleq \left\{(A_\rho, B_\rho, C_\rho): H_\rho(\rho) = C_\rho(\rho I - A_\rho)^{-1}B_\rho\right\} \quad (2.9)$$

where $\rho = z$ or δ, and $H_z(z) = H(z)$. Hence, if $(A_\rho, B_\rho, C_\rho) \in S_\rho$, $(T^{-1}A_\rho T, T^{-1}B_\rho, C_\rho T) \in S_\rho$ if and only if T is nonsingular.

Substituting (2.2) in (2.7), it is straightforward to establish that the following relationship exists between the state-space realizations $(A_z, B_z, C_z) \in S_z$ and $(A_\delta, B_\delta, C_\delta) \in S_\delta$:

$$A_z = \Delta A_\delta + I, \quad B_z = \Delta B_\delta, \quad C_z = C_\delta. \quad (2.10)$$

This means that if $(A_\delta, B_\delta, C_\delta) \in S_\delta$, one can find a corresponding realization $(A_z, B_z, C_z) \in S_z$ and vice-versa by the one-to-one mapping (2.10).

III. Sensitivity of Delta Operator Parameterizations

Consider the generalized state-space realization

$$\rho x_t = A_\rho x_t + B_\rho u_t \quad (3.1a)$$

$$y_t = C_\rho x_t \quad (3.1b)$$

where ρ is z or δ [see (2.6) and (2.7)], and where (A_ρ, B_ρ, C_ρ) is an infinite precision implementation of a transfer function $H_\rho(\rho)$, $\rho = z$ or δ. Assume that B_0 bits are available and denote by $A_\rho^*, B_\rho^*, C_\rho^*$ the implemented version of A_ρ, B_ρ, C_ρ where the coefficients have been truncated to B_0 bits. The actually implemented state-space model is then given by (3.1) with (A_ρ, B_ρ, C_ρ) replaced by $(A_\rho^*, B_\rho^*, C_\rho^*)$. It follows that the actual transfer function $H_\rho^*(\rho) = C_\rho^*(\rho I - A_\rho^*)^{-1}B_\rho^*$ and the ideal $H_\rho(\rho) = C_\rho(\rho I - A_\rho)^{-1}B_\rho$ will differ, and hence the output of the actually implemented filter to any input sequence will deviate from the output of the ideal filter. One way to evaluate this error is to compute a measure of the sensitivity of the transfer function $H(\rho)$ to errors on the matrices A_ρ, B_ρ, C_ρ. Here we first introduce a commonly used definition for the sensitivity measure of the state-space implementation of a transfer function in the generalized operator ρ. We then specialize these expressions to the case of shift and δ-operator representations.

A. Sensitivity Measure

Definition 3.1: Let $M \in C^{n\times m}$ be a matrix and let $f(M) \in C$ be a scalar complex function of M, differentiable w.r.t. all the elements of M. We then denote

$$\frac{\partial f}{\partial M} = S, \quad \text{where the } (i,j)\text{th element of } S \text{ is } s_{ij} \triangleq \frac{\partial f}{\partial m_{ij}}. \quad (3.2)$$

Definition 3.2: Let $f(z) \in C^{n\times m}$ be any complex matrix valued function of the complex variable z. We then define the l_p-norm of $f(z)$ as

$$\|f\|_p \triangleq \left(\frac{1}{2\pi}\int_0^{2\pi} \|f(e^{j\omega})\|_F^p \, d\omega\right)^{1/p} \quad (3.3)$$

where $\|f(e^{j\omega})\|_F$ is the Frobenius norm of the matrix $f(e^{j\omega})$:

$$\|f(e^{j\omega})\|_F = \left(\sum_{i=1}^n \sum_{k=1}^m |f_{ik}(e^{j\omega})|^2\right)^{1/2} = (\mathrm{tr}\,(f^T(e^{-j\omega})f(e^{j\omega})))^{1/2}. \quad (3.4)$$

The absolute sensitivity measure of the transfer function $H(z)$ w.r.t. the parameters in the realization A_ρ, B_ρ, C_ρ is then de-

fined as follows in [4]:

$$M_{a,\rho} \triangleq \left\| \frac{\partial H_\rho}{\partial A_\rho} \right\|_1^2 + \left\| \frac{\partial H_\rho}{\partial B_\rho} \right\|_2^2 + \left\| \frac{\partial H_\rho}{\partial C_\rho} \right\|_2^2 \tag{3.5}$$

where

$$\frac{\partial H_\rho}{\partial A_\rho} = \left(\rho I - A_\rho^T \right)^{-1} C_\rho^T B_\rho^T \left(\rho I - A_\rho^T \right)^{-1}$$

$$\frac{\partial H_\rho}{\partial B_\rho} = \left(\rho I - A_\rho^T \right)^{-1} C_\rho^T, \quad \frac{\partial H_\rho}{\partial C_\rho} = B_\rho^T \left(\rho I - A_\rho^T \right)^{-1}. \tag{3.6}$$

It then easily follows from (2.2), (2.10), and (3.6) that:

$$\frac{\partial H_\delta}{\partial A_\delta} = \Delta \frac{\partial H_z}{\partial A_z}, \quad \frac{\partial H_\delta}{\partial B_\delta} = \Delta \frac{\partial H_z}{\partial B_z}, \quad \frac{\partial H_\delta}{\partial C_\delta} = \frac{\partial H_z}{\partial C_z}. \tag{3.7}$$

Therefore, (3.5) specializes, for $\rho = z$ and $\rho = \delta$, to:

$$M_{a,z} \triangleq \left\| \frac{\partial H_z}{\partial A_z} \right\|_1^2 + \left\| \frac{\partial H_z}{\partial B_z} \right\|_2^2 + \left\| \frac{\partial H_z}{\partial C_z} \right\|_2^2 \tag{3.8}$$

and

$$M_{a,\delta} \triangleq \left\| \frac{\partial H_\delta}{\partial A_\delta} \right\|_1^2 + \left\| \frac{\partial H_\delta}{\partial B_\delta} \right\|_2^2 + \left\| \frac{\partial H_\delta}{\partial C_\delta} \right\|_2^2$$

$$= \Delta^2 \left\| \frac{\partial H_z}{\partial A_z} \right\|_1^2 + \Delta^2 \left\| \frac{\partial H_z}{\partial B_z} \right\|_2^2 + \left\| \frac{\partial H_z}{\partial C_z} \right\|_2^2. \tag{3.9}$$

We have therefore proved the following result.

Theorem 3.1: Consider two realizations (A_z, B_z, C_z) in S_z and $(A_\delta, B_\delta, C_\delta)$ in S_δ of the same transfer function, related by (2.10). Then $M_{a,\delta} < M_{a,z}$ if and only if $\Delta < 1$.

Proof: The proof follows from (3.8) and (3.9). END

Comment 3.1: Theorem 3.1 shows that $M_{a,\delta}$ can be made smaller than $M_{a,z}$ provided Δ can be chosen smaller than 1. We should note, however, that the value of Δ influences the range of the coefficients appearing in $(A_\delta, B_\delta, C_\delta)$ as is clear from (2.10). Therefore, the games we can play with Δ are limited by dynamic range considerations. To show this, let us consider the two following examples. In a fixed-point arithmetic implementation, the absolute values of all implemented coefficients are constrained to be within some interval, say [0.001, 1]. If a realization in shift operator is given by

$$A_z = \begin{pmatrix} 0.99 & -0.01 \\ 0.01 & 0.99 \end{pmatrix}, \quad B_z = \begin{pmatrix} 0.02 \\ 0.10 \end{pmatrix}, \quad C_z = (0.50 \quad -0.67),$$

the maximal absolute value of the elements in $(A_z - I)$ is 0.01. So, the maximal absolute value of the elements in $(A_z - I)$ and B_z is 0.1. Therefore, according to (2.10), the choices of Δ that will keep the coefficients of A_δ, B_δ within the required range are any value between 0.1 and 1. Hence, by choosing a δ-operator implementation with $\Delta = 0.1$ we can significantly reduce the sensitivity w.r.t. a shift-operator implementation while satisfying the dynamic range constraint.

Assume now that B_z and C_z are as before, but

$$A_z = \begin{pmatrix} 0.2314 & -0.0127 \\ 0.0231 & 0 \end{pmatrix}.$$

The dynamic range constraint will then force $\Delta = 1$ and hence the δ-operator realization will have the same sensitivity performance as the shift-operator realization. From these two examples, one can see that for a well scaled realization (for example, its largest absolute value is smaller than one) the choice of Δ is related to the poles of the system. See also Comment 3.4 below and the numerical example in Section IV for an illustration of this issue.

B. Optimal Realizations

One of the problems that has attracted attention of finite wordlength experts has been to minimize $M_{a,z}$ over all equivalent state-space realizations $\{A_z, B_z, C_z\}$ in S_z, i.e., over all possible shift-operator state-space realizations. As it turns out, the direct minimization of (3.8) is mathematically intractable. The problem was solved by Thiele [6], who first replaced $M_{a,z}$ of (3.8) by an upper bound $\overline{M}_{a,z}$ obtained by the Cauchy–Schwartz inequality:

$$M_{a,z} \le \overline{M}_{a,z} \triangleq \left\| \frac{\partial H_z}{\partial B_z} \right\|_2^2 \left\| \frac{\partial H_z}{\partial C_z} \right\|_2^2 + \left\| \frac{\partial H_z}{\partial B_z} \right\|_2^2 + \left\| \frac{\partial H_z}{\partial C_z} \right\|_2^2. \tag{3.10}$$

We note that

$$\left\| \frac{\partial H_z}{\partial B_z} \right\|_2^2 = \frac{1}{2\pi} \int_0^{2\pi} \operatorname{tr} \left[(e^{j\omega} I - A_z^T)^{-1} C_z^T C_z (e^{-j\omega} I - A_z)^{-1} \right] d\omega$$

$$= \operatorname{tr} \left[\sum_{i=0}^{\infty} (A_z^T)^i C_z^T C_z A_z^i \right] = \operatorname{tr} W_o \tag{3.11}$$

and, similarly,

$$\left\| \frac{\partial H_z}{\partial C_z} \right\|_2^2 = \operatorname{tr} W_c. \tag{3.12}$$

Here W_o and W_c are, respectively, the observability and controllability Gramians of (A_z, B_z, C_z). The upper bound $\overline{M}_{a,z}$ can be expressed as

$$\overline{M}_{a,z} = \operatorname{tr} W_o \operatorname{tr} W_c + \operatorname{tr} W_o + \operatorname{tr} W_c. \tag{3.13}$$

Thiele first characterized the set of realizations (A_z, B_z, C_z) that minimize $\overline{M}_{a,z}$ [6], and then showed that that set also minimizes $M_{a,z}$, and that $\overline{M}_{a,z} = M_{a,z}$ for those optimal realizations [12]. His results can be summarized as follows.

Theorem 3.2 [6], [12]:

i) $$\min_{S_z} \overline{M}_{a,z} = \min_{S_z} M_{a,z} = \left(\sum_1^n \sigma_i \right)^2 + 2 \sum_1^n \sigma_i \tag{3.14}$$

where σ_i, $i = 1, \cdots, n$ are the Hankel singular values of the transfer function $H(z)$ defined by

$$\sigma_i \triangleq [\lambda_i(W_c W_o)]^{1/2}. \tag{3.15}$$

ii) The set of optimizing realizations is characterized by

$$S_z^{\text{opt}} = \{(A_z, B_z, C_z) : W_c = W_o\}. \tag{3.16}$$

Proof: See [6] and [12].

These singular values are invariants of the transfer function, i.e., they are state-space realization independent. We now establish two new results. First we give an expression for the minimizing value of $M_{a,\delta}$ over all $(A_\delta, B_\delta, C_\delta)$ in S_δ. Then we characterize the optimal set S_δ^{opt} by relating the optimal realizations in delta form to the optimal realizations in shift form.

Similarly to the procedure used by Thiele, we first replace $M_{a,\delta}$ by an upper bound $\overline{M}_{a,\delta}$ using the same Cauchy–Schwartz inequality:

$$M_{a,\delta} \le \overline{M}_{a,\delta} \triangleq \left\| \frac{\partial H_\delta}{\partial B_\delta} \right\|_2^2 \left\| \frac{\partial H_\delta}{\partial C_\delta} \right\|_2^2 + \left\| \frac{\partial H_\delta}{\partial B_\delta} \right\|_2^2 + \left\| \frac{\partial H_\delta}{\partial C_\delta} \right\|_2^2. \quad (3.17)$$

Theorem 3.3: i) The minimal value of $M_{a,\delta}$ over all equivalent realizations $(A_\delta, B_\delta, C_\delta)$ in S_δ is

$$\min_{S_\delta} M_{a,\delta} = \min_{S_\delta} \overline{M}_{a,\delta} = \Delta^2 \left(\sum_1^n \sigma_i \right)^2 + 2\Delta \sum_1^n \sigma_i. \quad (3.18)$$

ii) The set of optimal realizations is characterized by

$$S_\delta^{\text{opt}} = \{(A_\delta, B_\delta, C_\delta) : W_c = \Delta^2 W_o\} \quad (3.19)$$

where W_c and W_o are the Gramians of the corresponding z-operator realization obtained from (2.10).

Proof: For every $(A_\delta, B_\delta, C_\delta)$ there exists a corresponding triple (A_z, B_z, C_z) defined by (2.10); it has a controllability Gramian W_c and an observability Gramian W_o. We now denote $\overline{W}_o = \Delta^2 W_o$. Therefore, by (3.7), (3.11), (3.12), (3.17) $\overline{M}_{a,\delta}$ can be expressed as

$$\begin{aligned} \overline{M}_{a,\delta} &= \Delta^2 \operatorname{tr}(W_c) \operatorname{tr}(W_o) + \Delta^2 \operatorname{tr}(W_o) + \operatorname{tr}(W_c) \\ &= \operatorname{tr}(\overline{W}_o) \operatorname{tr}(W_c) + \operatorname{tr}(\overline{W}_o) + \operatorname{tr}(W_c). \end{aligned} \quad (3.20)$$

Denote $u_i \triangleq [\lambda_i(W_c \overline{W}_o)]^{1/2} = [\lambda_i(\Delta^2 W_c W_o)]^{1/2} = \Delta\sigma_i$. It then follows by the same proof as that of Thiele [6] (an alternative proof can be found in [8]) that the minimizing value of $\overline{M}_{a,\delta}$ is

$$\min_{S_\delta} \overline{M}_{a,\delta} = \left(\sum_1^n u_i \right)^2 + 2 \sum_1^n u_i, \quad (3.21)$$

and that this value is achieved if and only if $W_c = \overline{W}_o$ (i.e., $W_c = \Delta^2 W_o$), where W_c and W_o are given in (3.11) and (3.12). Using the same procedure as used in [12], one can show that $\min_{S_\delta} M_{a,\delta} = \min_{S_\delta} \overline{M}_{a,\delta}$. END

Theorem 3.4: Let $S_\delta^{\text{opt}} = \{A_\delta^{\text{opt}}, B_\delta^{\text{opt}}, C_\delta^{\text{opt}}\}$ denote the subset of S_δ that minimizes $M_{a,\delta}$ and let $S_z^{\text{opt}} = \{A_z^{\text{opt}}, B_z^{\text{opt}}, C_z^{\text{opt}}\}$ denote the subset of S_z that minimizes $M_{a,z}$. Then to each $(A_z^{\text{opt}}, B_z^{\text{opt}}, C_z^{\text{opt}}) \in S_z^{\text{opt}}$ there corresponds a $(A_\delta^{\text{opt}}, B_\delta^{\text{opt}}, C_\delta^{\text{opt}}) \in S_\delta^{\text{opt}}$ such that

$$\begin{aligned} A_\delta^{\text{opt}} &= \Delta^{-1}(A_z^{\text{opt}} - I), \quad B_\delta^{\text{opt}} = \Delta^{-1/2} B_z^{\text{opt}}, \\ C_\delta^{\text{opt}} &= \Delta^{-1/2} C_z^{\text{opt}}. \end{aligned} \quad (3.22)$$

Proof: Consider a member $(A_z^{\text{opt}}, B_z^{\text{opt}}, C_z^{\text{opt}})$ of S_z^{opt}. Then the corresponding Gramians satisfy $W_c = W_o$. Let $(A_\delta, B_\delta, C_\delta)$ be obtained from $(A_z^{\text{opt}}, B_z^{\text{opt}}, C_z^{\text{opt}})$ by (2.10):

$$A_\delta = \Delta^{-1}(A_z^{\text{opt}} - I), \quad B_\delta = \Delta^{-1} B_z^{\text{opt}}, \quad C_\delta = C_z^{\text{opt}}. \quad (3.23)$$

We know by the proof of Theorem 3.3 that optimality in S_δ requires $W_c = \Delta^2 W_o$. Now apply a similarity transformation $T = \Delta^{-1/2} I$ to $(A_z^{\text{opt}}, B_z^{\text{opt}}, C_z^{\text{opt}})$ and define

$$A_z^1 = T^{-1} A_z^{\text{opt}} T, \quad B_z^1 = T^{-1} B_z^{\text{opt}}, \quad C_z^1 = C_z^{\text{opt}} T. \quad (3.24)$$

The Gramians of the realization (A_z^1, B_z^1, C_z^1) are $W_c^1 = \Delta W_c$ and $W_o^1 = \Delta^{-1} W_o$. Since $W_c = W_o$, it follows that $W_c^1 = \Delta^2 W_o^1$, and hence, by Theorem 3.3 ii), the δ-realization corresponding to (A_z^1, B_z^1, C_z^1) is optimal in S_δ [see (3.19)]. It now follows from (3.23), (3.24) and $T = \Delta^{-1/2} I$ that this δ-realization is expressed in terms of $(A_z^{\text{opt}}, B_z^{\text{opt}}, C_z^{\text{opt}})$ by (3.22). END

Comment 3.2: Theorem 3.4 is important in that it shows that the search for optimal realizations $(A_\delta, B_\delta, C_\delta)$ in δ-form does not require a new construction. The results of Thiele [6] that characterize the optimal shift operator state variable forms also completely characterize the optimal delta operator forms via (3.22).

Comment 3.3: For nonoptimal realizations, $\overline{M}_{a,\rho}$ is much easier to compute than $M_{a,\rho}$ and is therefore a reasonable measure of comparison between different realizations. A detailed discussion can be found in [12], [13].

Comment 3.4: The balanced realization (A_b, B_b, C_b) is one of the optimal realizations in shift operator. From (2.10) one sees that the choices of Δ depend on the diagonal elements of $(A_b - I)$. When the largest absolute value of these elements is less than the largest absolute value of the elements of A_b, we will be able to choose Δ smaller than one. This is the case when the poles of a system are clustered around $z = +1$, i.e., when fast sampling is used (see [3]). In this case, the diagonal elements of A_b will be near 1, and hence the diagonal elements of $(A_b - I)$ will be much smaller than 1, which yields the possibility of choosing Δ smaller than 1. We refer the reader to [14] for a further discussion of this. The following example also illustrates this point.

IV. Numerical Example

We now illustrate our previous results and calculations on the optimal sensitivity measure with the following example, already used in [9]. Consider a system described in shift operator implementation by the following control canonical from:

$$A_{c,z} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.4538 & -1.5562 & \underline{1.9749} \end{bmatrix},$$

$$B_{c,z} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad C_{c,z} = \begin{bmatrix} 0.0232 \\ \underline{0.0230} \\ 0.0792 \end{bmatrix}^T.$$

The poles are at 0.6579 and $0.6585 \pm j0.5061$. The smallest and largest numbers (in magnitude) are underlined, as they will be in the other realizations.

The balanced form in S_z, one of the optimal realizations minimizing $M_{a,z}$, is given by

$$A_z^{\text{opt}} = \begin{bmatrix} \underline{0.8236} & 0.3999 & -0.0165 \\ -0.3999 & 0.5935 & -0.3425 \\ \underline{-0.0165} & -0.3425 & 0.5577 \end{bmatrix}$$

$$B_z^{\text{opt}} = \begin{bmatrix} 0.4424 \\ 0.3799 \\ 0.1671 \end{bmatrix} \quad C_z^{\text{opt}} = \begin{bmatrix} 0.4424 \\ -0.3799 \\ 0.1671 \end{bmatrix}^T.$$

The largest number (in magnitude) in the triplet $(A_z^{\text{opt}} - I, B_z^{\text{opt}}, C_z^{\text{opt}})$ is 0.4423. Therefore, we choose $\Delta = 2^{-1}$. The corresponding companion and optimal realizations in S_δ are

$$A_{c,\delta} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1.0203 & \underline{-2.4258} & -2.0503 \end{bmatrix},$$

$$B_{c,\delta} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad C_{c,\delta} = \begin{bmatrix} 1.0040 \\ 0.7265 \\ \underline{0.1586} \end{bmatrix}^T$$

and

$$A_\delta^{\text{opt}} = \begin{bmatrix} -0.3527 & 0.7999 & -0.0329 \\ -0.7999 & -0.8130 & 0.6849 \\ \underline{-0.0329} & -0.6849 & \underline{-0.8846} \end{bmatrix}$$

$$B_\delta^{\text{opt}} = \begin{bmatrix} 0.6256 \\ 0.5373 \\ 0.2363 \end{bmatrix} \quad C_\delta^{\text{opt}} = \begin{bmatrix} 0.6256 \\ -0.5373 \\ 0.2363 \end{bmatrix}^T.$$

We note that the coefficient ranges of $(A_z^{\text{opt}}, B_z^{\text{opt}}, C_z^{\text{opt}})$ and $(A_\delta^{\text{opt}}, B_\delta^{\text{opt}}, C_\delta^{\text{opt}})$ are roughly the same: in both cases the coefficients are between 2^{-6} and 1. The ratios of |max.coefficient|/|min.coefficient| are, respectively, $(0.8236/0.0165) \approx 50$ for the shift form and $(0.8846/0.0329) \approx 27$ for the δ-form. The optimal values of the sensitivity measures are, respectively, $M_{a,z}^{\text{opt}} = 4.7560 = \overline{M}_{a,z}^{\text{opt}}$ and $M_{a,\delta}^{\text{opt}} = 1.8886 = \overline{M}_{a,\delta}^{\text{opt}}$. We have also computed $\overline{M}_a$ for the shift operator and delta operator control canonical forms. These are, respectively, $\overline{M}_{a,z}^{cc} = 81.9891$ and $\overline{M}_{a,\delta}^{cc} = 5.1605$.

These theoretical results will now be confirmed by a numerical simulation on the same example. For both the optimal z-form realization $(A_z^{\text{opt}}, B_z^{\text{opt}}, C_z^{\text{opt}})$ and the optimal δ-form realization $(A_\delta^{\text{opt}}, B_\delta^{\text{opt}}, C_\delta^{\text{opt}})$ presented above, we compute the corresponding frequency response $H_{wl}^p(\omega)$ obtained when the coefficients are implemented in fixed point with p significant bits, with p ranging from 5 to 30. We compare this with the ideal frequency response $H_{id}(\omega)$ implemented with infinite precision, by computing the worst deviation over the frequency range, i.e., the H_∞ error:

$$R \triangleq \log\left[\sup_{\omega \in (0, 2\pi)} |H_{id}(\omega) - H_{wl}^p(\omega)|\right].$$

The results for the example described above are shown in Fig. 1 in which R_z^{opt} and R_δ^{opt} denote the optimal realizations $(A_z^{\text{opt}}, B_z^{\text{opt}}, C_z^{\text{opt}})$ and $(A_\delta^{\text{opt}}, B_\delta^{\text{opt}}, C_\delta^{\text{opt}})$, respectively. It clearly shows the superiority of the optimal δ-form realization over the optimal z-form realization whatever the number of bits.

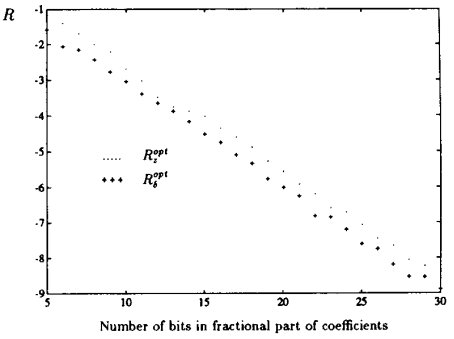

Fig. 1.

V. Conclusions

Our aim in this note has been to compare shift operator and delta operator state space parameterizations in terms of the effects of finite wordlength errors on the actual transfer function. We have first established the relationships between shift- and δ-operator parameterizations in terms of transfer function as well as state-space realization. Using a commonly adopted sensitivity measure, we have then found the optimal realization set in δ-operator. The relationship between this optimal realization set and that in shift operator has been established. It has been shown that the parameterizations in δ-operator yield a superior sensitivity performance over those in shift operator as long as the design parameter Δ can be chosen less than 1. This requirement is almost always satisfied when fast sampling is used.

The results presented in this note relate to the sensitivity of the transfer function with respect to coefficient errors in shift or delta operator implementations, respectively. Another important point of comparison between shift and delta operator realizations is their behavior with respect to roundoff errors on the signals. This has been examined in [14], with similar conclusions: delta operator realizations will typically yield a smaller roundoff noise gain than shift operator realizations when the poles are clustered around $z = 1$.

References

[1] V. Peterka, "Control of uncertain processes: Applied theory and algorithms," *Kybernetika*, vol. 22, pp. 1–102, 1986.

[2] R. H. Middleton and G. C. Goodwin, "Improved finite wordlength characteristics in digital control using delta operators," *IEEE Trans. Automat. Contr.*, vol. AC-31, pp. 1015–1021, Nov. 1986.

[3] ——, *Digital Estimation and Control: A Unified Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1990.

[4] V. Tavsanoglu and L. Thiele, "Optimal design of state-space digital filters by simultaneous minimization of sensitivity and roundoff noise," *IEEE Trans. Circuits Syst.*, vol. CAS-31, no. 10, pp. 884–888, Oct. 1984.

[5] W. J. Lutz and S. Louis Hakimi, "Design of multiinput multioutput systems with minimum sensitivity," *IEEE Trans. Circuits Syst.*, vol. 35, no. 9, pp. 1114–1122, Sept. 1988.

[6] L. Thiele, "Design of sensitivity and roundoff noise optimal state-space discrete systems," *Int. J. Circuit Theory Appl.*, vol. 12, pp. 39–46, 1984.

[7] G. Li and M. Gevers, "Optimal finite precision implementation of a state-estimate feedback controller," *IEEE Trans. Circuits Syst.*, vol. 37, no. 12, pp. 1487–1498, Dec. 1990.

[8] ——, *Parameterizations in Control, Estimation and Filtering Problems: Accuracy Aspects, Communication and Control Engineering Series*. London: Springer-Verlag, Jan. 1993.

[9] S. Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," *IEEE Trans. Acoustics, Speech Signal Processing*, vol. ASSP-25, pp. 273–281, Aug. 1977.

[10] D. Williamson and K. Kadiman, "Optimal finite wordlength linear quadratic regulation," *IEEE Trans. Automat. Contr.*, vol. 34, no. 12, pp. 1218–1228, Dec. 1989.

[11] D. Williamson, "Delay replacement in direct form structures," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 36, no. 4, pp. 453–460, Apr. 1988.

[12] L. Thiele, "On the sensitivity of linear state-space systems," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 502–510, May 1986.

[13] G. Li, B. D. O. Anderson, M. Gevers, and J. E. Perkins, "Optimal FWL design of state-space digital systems with weighted sensitivity minimization and sparseness consideration," *IEEE Trans. Circuits, Syst.*, vol. CAS-39, pp. 365–377, May 1992.

[14] G. Li and M. Gevers, "Roundoff noise minimization using Delta-operator realizations," *IEEE Trans. Acoustics, Speech, Signal Processing*, to appear in Feb. 1993.