WP-12-1 - 3:40

Proceedings of the 29th Conference
on Decision and Control
Honolulu, Hawaii • December 1990

COMPARATIVE STUDY OF FINITE WORDLENGTH EFFECTS IN SHIFT AND DELTA OPERATOR PARAMETRIZATIONS

Li Gang and Gevers Michel
Laboratoire d'Automatique, Dynamique et Analyse des Systèmes
Bâtiment Maxwell, 3 place du Levant
B-1348 Louvain-la-Neuve - Belgium

ABSTRACT

This paper 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 parametrizations. Both the absolute sensitivity, naturally connected to fixed point arithmetic, and the relative sensitivity, naturally connected to floating point arithmetic, are analyzed. In both cases, but particularly the latter, the delta operator parametrizations are shown to produce better sensitivity properties.

1.INTRODUCTION

In the last few years, both Peterka [1] and Middleton and Goodwin [2] have promoted the use of the delta operator as opposed to the shift operator in estimation and control applications. This promotional effort has culminated in the publication of a book where much of the present digital control and estimation theory has been reformulated in a delta operator framework [3]. Two major advantages are claimed for the delta operator formulation : 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 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 parametrization or a delta-operator parametrization. This is the object of the present paper. This problem is of course of interest when the coefficients of the state-space model are implemented in finite wordlength (FWL), which causes the transfer function of the actual model to deviate from the ideal (infinite precision) transfer function.

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]). In [7] this study has been extended to the effect of FWL errors on the closed loop transfer fuction in the case of a pole-placement control strategy while in [10] the effect on LQG regulators has been examined. This has led to a commonly accepted measure for the sensitivity of a transfer function w.r.t. the coefficients of (A,B,C) (see [4] - [6]), and to the search for optimal realizations (A^{opt}, B^{opt}, C^{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 [12]. All of these results relate solely to shift operator state space representations.

Here we first show in Section 2 that shift and delta operator representations can be embedded as special cases of more general polynomial parametrizations. The relationships between shift and delta operator parametrizations, both in transfer function and in state-space form, are established in Section 3. We use a definition of delta operator that is slightly more flexible than that commonly used by Middleton and Goodwin. Section 4 is concerned with the study of a commonly accepted absolute sensitivity measure and its upper bound. Our main new results of this Section consist in computing an expression for the best achievable absolute sensitivity of all equivalent delta state space realizations. 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 absolute sensitivity upper bound achievable with the delta operator state space models is normally smaller than that achievable with the shift operator state space models. For floating point computations, it is more natural to search for parametrizations that make the relative sensitivity small. In Section 5, we derive expressions for the relative sensivity for both shift and delta state space models, together with reasonable and tractable upper bounds. We show that if the sampling rate has been chosen reasonably, then the upper bound of the relative sensitivity is always smaller for delta parametrizations than for shift parametrizations.

2. GENERALIZED POLYNOMIAL PARAMETRIZATIONS

Throughout this paper 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 :

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

Here z can be considered as the complex variable of the z - transform. However, it can also be looked upon as a time-domain operator :

$$zf(t) \overset{\Delta}{=} f(t+1), \text{ forward shift} \qquad (2.2a)$$

$$z^{-1}f(t) \overset{\Delta}{=} f(t-1), \text{ backward shift} \qquad (2.2b)$$

Using vector notations, we can rewrite H(z) as

$$H(z) = \frac{\bar{b}^T\bar{Z}}{\bar{a}^T\bar{Z}} \qquad (2.3)$$

where

$$\bar{a} \overset{\Delta}{=} (1 \;\; a_1 \;\; a_2 \;\dots\; a_n)^T \qquad (2.4a)$$

$$\bar{b} \overset{\Delta}{=} (0 \;\; b_1 \;\; b_2 \;\dots\; b_n)^T \qquad (2.4b)$$

$$\bar{Z} \overset{\Delta}{=} (z^n \;\; z^{n-1} \;\dots\; z \;\; 1)^T \qquad (2.4c)$$

Now consider a nxn nonsingular matrix T, whose first row is (1 0 ... 0) , but that is otherwise arbitrary. Then H(z) can also be represented as

$$H(z) = \frac{\bar{b}^T T^T T^{-T}\bar{Z}}{\bar{a}^T T^T T^{-T}\bar{Z}} = \frac{\bar{\beta}^T\bar{P}(z)}{\bar{\alpha}^T\bar{P}(z)} \qquad (2.5)^+$$

where

$$\bar{\alpha} = T\bar{a} \overset{\Delta}{=} (1 \;\; \alpha_1 \dots \alpha_n)^T \qquad (2.6a)$$

$$\bar{\beta} = T\bar{b} \overset{\Delta}{=} (0 \;\; \beta_1 \dots \beta_n)^T \qquad (2.6b)$$

$$\bar{P}(z) = T^{-T}\bar{Z} \overset{\Delta}{=} (p_0(z) \;\; p_1(z) \dots p_{n-1}(z) \;\; p_n(z))^T \qquad (2.6c)$$

with $p_0(z)$ a monic polynomial of degree n and $p_i(z)$, i = 1, ..., n polynomials of degree less than or equal to n-1. The $p_i(z)$, i = 0, 1, ..., n can be thought of as basis functions in which the polynomials $\bar{a}^T Z$ and $\bar{b}^T Z$ are expressed. This shows that a given transfer function can be represented in an infinite number of equivalent (at least in infinite precision) parametrizations. When this transfer function model is to be implemented in finite precision, this observation can be exploited to improve the numerical properties such as sensitivity and roundoff noise. This will be fully explored in [8], but here we shall concentrate on a comparison of the numerical properties of two special cases of the representation (2.5), namely the shift operator representation (2.1) and the δ-operator representation popularized by Middleton and Goodwin [3]. We shall introduce the following definition for the δ-operator. We take

$$\delta \overset{\Delta}{=} \frac{z-1}{\Delta} \qquad (2.7)$$

Here Δ is any positive number, not necessarily the sampling period T_S of the discrete time system as in [3]. The choice of a value for Δ and its role in improving some numerical properties will be discussed later.

3. RELATIONSHIP BETWEEN SHIFT OPERATOR AND δ-OPERATOR REPRESENTATIONS

In this section we shall establish a number of equivalence relationships between the coefficients of shift operator and δ-operator representations, both in input-output form and in state-space form.
With the definition (2.7) for δ, the transfer function H(z) of (2.3) can be reexpressed in δ-form as follows :

$$H(z) = \frac{\sum_1^n b_i z^{n-i}}{z^n + \sum_1^n a_i z^{n-i}} = \frac{\sum_1^n \beta_i \delta^{n-i}}{\delta^n + \sum_1^n \alpha_i \delta^{n-i}} \overset{\Delta}{=} H_\delta(\delta) \qquad (3.1)$$

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 :

$$\bar{\beta} = \begin{pmatrix} 0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_n \end{pmatrix} = T \begin{pmatrix} 0 \\ b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{pmatrix}, \bar{\alpha} = \begin{pmatrix} 1 \\ \alpha_1 \\ \cdot \\ \cdot \\ \cdot \\ \alpha_n \end{pmatrix} = T \begin{pmatrix} 1 \\ a_1 \\ \cdot \\ \cdot \\ \cdot \\ a_n \end{pmatrix} \qquad (3.2)$$

where

$$T = \begin{pmatrix} 1 & 0 & \cdots & & 0 \\ t_{21} & t_{22} & 0 & \cdots & 0 \\ t_{31} & t_{32} & t_{33} & \ddots & \cdot \\ \cdot & \cdot & & \ddots & \cdot \\ \cdot & \cdot & & \ddots & 0 \\ t_{n+1,1} & t_{n+1,2} & & & t_{n+1,n+1} \end{pmatrix} \qquad (3.3a)$$

with

$$t_{ij} = C_{n+1-j}^{i-j}\, \Delta^{-(i-1)}, i \geq j ; C_m^i = \frac{m!}{(m-i)!\, i!} . \qquad (3.3b)$$

Comparing with (2.5)-(2.6), we note that the δ-operator corresponds with the following choice of polynomial basis functions :

$$p_i(z) \overset{\Delta}{=} \Delta^n \left(\frac{z-1}{\Delta}\right)^{n-i} = \Delta^n \delta^{n-i} \quad i = 0, 1, 2, \dots n , \qquad (3.4)$$

where the normalizing factor Δ^n is there to force the monicity of $p_0(z)$. Going back to (3.1), we observe that H(z) and $H_\delta(\delta)$ are two different but equivalent parametrizations representing the same object. Assuming that u(.) and y(.) are, respectively, the input and output of the filter H(z), they correspond to the following equivalent time-domain representations :

$$y_{t+n} + a_1 y_{t+n-1} + \dots + a_n y_t = b_1 u_{t+n-1} + \dots + b_n u_t \qquad (3.5)$$

$$\delta^n y_t + \alpha_1 \delta^{n-1} y_t + \dots + \alpha_{n-1} \delta y_t + \alpha_n y_t = \beta_1 \delta^{n-1} u_t + \dots + \beta_{n-1} \delta u_t + \beta_n u_t \qquad (3.6)$$

where

$$\delta y_t \overset{\Delta}{=} \frac{y_{t+1} - y_t}{\Delta} . \qquad (3.7)$$

These two input-output relationships can also be represented by a shift-operator (resp. δ-operator) state-space model as follows :

$$\begin{cases} z x_t^{(1)} = A_z x_t^{(1)} + B_z u_t \\ y_t = C_z x_t^{(1)} \end{cases} \qquad (3.8)$$

and

$$\begin{cases} \delta x_t^{(2)} = A_\delta x_t^{(2)} + B_\delta u_t \\ y_t = C_\delta x_t^{(2)} \end{cases} \qquad (3.9)$$

The following relationships relate the internal and external representations :

$$H(z) = C_z (zI - A_z)^{-1} B_z , \quad H_\delta(\delta) = C_\delta (\delta I - A_\delta)^{-1} B_\delta . \qquad (3.10)$$

We shall for future use introduce the notion of a realization set S_ρ. We define :

$$S_\rho \overset{\Delta}{=} \{(A_\rho, B_\rho, C_\rho) : H(\rho) = C_\rho(\rho I - A_\rho)^{-1} B_\rho\} \qquad (3.11)$$

where ρ = z or δ. 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 (2.7) in (3.9), 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 $(A_\delta, B_\delta, C_\delta) \in S_\delta$:

$$\begin{cases} A_z = \Delta . A_\delta + I \\ B_z = \Delta . B_\delta \\ C_z = C_\delta \end{cases} \qquad (3.12)$$

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 (3.12).

4. ABSOLUTE SENSITIVITY OF GENERALIZED POLYNOMIAL PARAMETRIZATIONS

In any practical implementation of an input-output relationship, the coefficients can only be implemented in finite precision, with the number of bits determined by the available hardware. This means that the finite wordlength (FWL) errors on the implementation of these coefficients introduce errors on the actually computed transfer function. The magnitude of these errors can be measured by what is called a sensitivity measure. Here we first introduce a commonly used definition for the sensitivity measure of the state-space implementation of a transfer function for generalized polynomial implementations. We then specialize these expressions to the case of shift and δ-operator representations.

Consider the generalized state space realization

$$\begin{cases} \rho x_t = A_\rho x_t + B_\rho u_t \\ y_t = C_\rho x_t \end{cases} \qquad (4.1)$$

where ρ is z or δ (see (3.8)-(3.9)),and where (A_ρ, B_ρ, C_ρ) is an infinite precision implementation of a transfer function

$$H(\rho) = C_\rho(\rho I - A_\rho)^{-1}B_\rho, \quad \rho = z \text{ or } \delta. \tag{4.2}$$

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

$$\begin{cases} \rho\tilde{x}_t = A_\rho^*\tilde{x}_t + B_\rho^* u_t \\ \tilde{y}_t = C_\rho^*\tilde{x}_t \end{cases} \tag{4.3}$$

It follows that $H_p(\rho) = C_\rho(\rho I - A_\rho)^{-1} B_\rho$ and $H_\rho^*(\rho) = C_\rho^*(\rho I - A_\rho^*)^{-1}B_\rho^*$ will differ. One way to evaluate this error is to compute a measure of the sensitivity of the transfer function $H_\rho(\rho)$ to errors on the matrices A_ρ, B_ρ, C_ρ. We now define such a sensitivity measure.

4.1. Absolute sensitivity measure

Definition 4.1

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

$$\frac{\partial f}{\partial M} = S \text{, with } s_{ij} \overset{\Delta}{=} \frac{\partial f}{\partial m_{ij}} \tag{4.4}$$

where s_{ij} denotes the $(i,j)^{th}$ element of the matrix S.

Definition 4.2

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

$$\| f \|_p \overset{\Delta}{=} \left(\frac{1}{2\pi}\int_0^{2\pi} \| f(e^{j\omega}) \|_F^p \, d\omega \right)^{1/p} \tag{4.5}$$

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

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 defined as follows in [4],

$$M_{a,\rho} = \left| \frac{\partial H}{\partial A_\rho} \right|_1^2 + \left| \frac{\partial H}{\partial B_\rho} \right\|_2^2 + \left| \frac{\partial H}{\partial C_\rho^T} \right|_2^2 . \tag{4.6}$$

The word absolute is used to express the fact that this sensitivity measure expresses the effect on the transfer function of an absolute error on any coefficient of A_ρ, B_ρ, C_ρ. It is an obvious sensitivity measure in the case of fixed-point implementations. It is easy to see that :

$$\begin{aligned} \frac{\partial H}{\partial A_\rho} &= (\rho I - A_\rho^T)^{-1} C_\rho^T B_\rho^T (\rho I - A_\rho^T)^{-1} \\ \frac{\partial H}{\partial B_\rho} &= (\rho I - A_\rho^T)^{-1} C_\rho^T \\ \frac{\partial H}{\partial C_\rho} &= B_\rho^T (\rho I - A_\rho^T)^{-1} \end{aligned} \tag{4.7}$$

4.2. An upper bound for the absolute sensitivity measure

One of the purposes of the sensitivity measure (4.6) is that it should enable one to compute a realization $(A_\rho^0, B_\rho^0, C_\rho^0)$ that minimizes $M_{a,\rho}$ over the equivalence class $S_\rho = \{T^{-1}A_\rho T, T^{-1}B_\rho, C_\rho T\}$ of realizations. As it turns out, it is extremely hard to perform this optimization of $M_{a,\rho}$. However, the problem becomes feasible if $M_{a,\rho}$ is replaced by the following reasonable upper bound, which follows from the Cauchy-Schwarz inequality :

$$\bar{M}_{a\rho} \overset{\Delta}{=} \left| \frac{\partial H}{\partial B_\rho} \right|_2^2 \left| \frac{\partial H}{\partial C_\rho} \right|_2^2 + \left| \frac{\partial H}{\partial B_\rho} \right|_2^2 + \left| \frac{\partial H}{\partial C_\rho} \right|_2^2 . \tag{4.8}$$

For $\rho = z$, we can compute an interesting alternative expression for $\bar{M}_{a,\rho}$ on noting that

$$\begin{aligned} \left| \frac{\partial H}{\partial B_z} \right|_2^2 &= \frac{1}{2\pi}\int_0^{2\pi} \mathrm{tr}[(e^{j\omega} I - A_z^T)^{-1} C_z^T C_z (e^{-j\omega} I - A_z)^{-1}] d\omega \\ &= \frac{1}{2\pi j}\oint_{|z|=1} \mathrm{tr}[(zI - A_z^T)^{-1} C_z^T C_z (z^{-1}I - A_z)^{-1}] z^{-1} dz \\ &= \mathrm{tr}\left[\sum_{i=0}^{\infty} (A_z^T)^i C_z^T C_z A_z^i \right] = \mathrm{tr}\, W_0 . \end{aligned} \tag{4.9}$$

Similarly

$$\left| \frac{\partial H}{\partial C_z} \right|_2^2 = \mathrm{tr}\, W_c . \tag{4.10}$$

Here W_0 and W_c are, respectively, the observability and controllability Gramians of (A_z, B_z, C_z). The upper bound for shift-operator state-space forms, $\bar{M}_{a,z}$, can then be reexpressed as

$$\bar{M}_{a,z} = \mathrm{tr}(W_c)\, \mathrm{tr}(W_0) + \mathrm{tr}(W_c) + \mathrm{tr}(W_0) . \tag{4.11}$$

To compare $\bar{M}_{a,z}$ with the upper bound for delta-operator state-space forms, $\bar{M}_{a,\delta}$, we first compute the relationships between the respective Gramians. It follows from (4.7), (3.12) and (2.7) that

$$\begin{aligned} \frac{\partial H}{\partial B_\delta} &= (\delta I - A_\delta^T)^{-1} C_\delta^T = (zI - I - \Delta A_\delta^T)^{-1} \Delta C_\delta^T \\ &= (zI - A_z)^{-1} \Delta C_z^T = \Delta \frac{\partial H}{\partial B_z} \end{aligned} \tag{4.12}$$

$$\begin{aligned} \frac{\partial H}{\partial C_\delta} &= B_\delta^T (\delta I - A_\delta^T)^{-1} = \Delta B_\delta^T (zI - I - \Delta A_\delta^T)^{-1} \\ &= B_z^T (zI - A_z^T)^{-1} = \frac{\partial H}{\partial C_z} \end{aligned} \tag{4.13}$$

Therefore the sensitivity $\bar{M}_{a,\delta}$ can be expressed as

$$\bar{M}_{a,\delta} = \Delta^2 \mathrm{tr}(W_c)\, \mathrm{tr}(W_0) + \mathrm{tr}(W_c) + \Delta^2 \mathrm{tr}(W_0) \tag{4.14}$$

4.3. Optimal realizations

One of the problems that has attracted attention of finite wordlength experts has been to minimize the upper bound $\bar{M}_{a,z}$ 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. This problem has been solved by Thiele [6], who characterized the set of optimizing realizations

$$S_z^{opt} = \{(A_z, B_z, C_z) : W_c = W_0\} . \tag{4.15}$$

He also showed that

$$\min_{S_z} \bar{M}_{a,z} = \left(\sum_1^n \sigma_i \right)^2 + 2\sum_1^n \sigma_i = \bar{M}_{a,z}^{opt} \tag{4.16}$$

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

$$\sigma_i \overset{\Delta}{=} [\lambda_i(W_c W_0)]^{1/2} . \tag{4.17}$$

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 $\bar{M}_{a,\delta}$ 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. The proofs are available in the journal version of this paper.

Theorem 4.1

The minimal value of $\bar{M}_{a,\delta}$ over all equivalent realizations (A_δ, B_δ, C_δ) in S_δ is

$$\bar{M}_{a,\delta}^{opt} = \Delta^2\left(\sum_1^n \sigma_i\right)^2 + 2\Delta\sum_1^n \sigma_i \tag{4.18}$$

It should be pointed out that recently Thiele has shown that the optimal realizations in S_z^{opt}not only minimize the upper bound $\bar{M}_{a,z}$but also the actual sensitivity measure $M_{a,z}$[12]. The same holds for S_δ^{opt}.

Theorem 4.2

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

$$\begin{cases} A_\delta^{opt} = \Delta^{-1}(A_z^{opt} - I) \\ B_\delta^{opt} = \Delta^{-1/2} B_z^{opt} \\ C_\delta^{opt} = \Delta^{-1/2} C_z^{opt} \end{cases} \tag{4.19}$$

Expressions (4.16) and (4.18) allow a comparison between the best achievable sensitivity upper bounds in shift and delta state variable realizations, respectively. Clearly the delta form realization will yield a better bound if Δ can be chosen smaller than 1. Now, it appears from many examples that for stable systems the number Δ can be chosen less than 1 while preserving the same range for the coefficients of $(A_\delta^{opt}, B_\delta^{opt}, C_\delta^{opt})$ as the range of $(A_z^{opt}, B_z^{opt}, C_z^{opt})$.

4.4 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 :

$$A_{c,z} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.4538 & -1.5562 & \underline{1.9749} \end{bmatrix} \quad B_{c,z} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad C_{c,z} = \begin{bmatrix} 0.0232 \\ \underline{0.0230} \\ 0.0792 \end{bmatrix}^T$$

The poles are at 0.6579 and 0.6585 ± j 0.5061. The smallest and largest numbers (in magnitude) are underlined, as they will be in the other realizations. The internally balanced form in S_z is one of the optimal realizations minimizing (4.11). It is given by

$$A_z^{opt} = \begin{bmatrix} \underline{0.8236} & 0.3999 & -0.0165 \\ -0.3999 & 0.5935 & -0.3425 \\ -\underline{0.0165} & -0.3425 & 0.5577 \end{bmatrix} \quad B_z^{opt} = \begin{bmatrix} 0.4424 \\ 0.3799 \\ 0.1671 \end{bmatrix} \quad C_z^{opt} = \begin{bmatrix} 0.4424 \\ -0.3799 \\ 0.1671 \end{bmatrix}$$

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

$$A_{c,\delta} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1.0203 & -\underline{2.4258} & -2.0503 \end{bmatrix}, B_{c,\delta} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, C_{c,\delta} = \begin{bmatrix} 1.0040 \\ 0.7265 \\ \underline{0.1586} \end{bmatrix}^T$$

$$A_\delta^{opt} = \begin{bmatrix} -0.3527 & 0.7999 & -0.0329 \\ -0.7999 & -0.8130 & 0.6849 \\ -\underline{0.0329} & -0.6849 & -\underline{0.8846} \end{bmatrix} \quad B_\delta^{opt} = \begin{bmatrix} 0.6256 \\ 0.5373 \\ 0.2363 \end{bmatrix} \quad C_\delta^{opt} = \begin{bmatrix} 0.6256 \\ -0.5373 \\ 0.2363 \end{bmatrix}^T$$

The optimal values of the sensitivity upper bounds are, respectively,

$$\bar{M}_{a,z}^{opt} = 4.756 \quad \text{and} \quad \bar{M}_{a,\delta}^{opt} = 1.8886 . \tag{4.20}$$

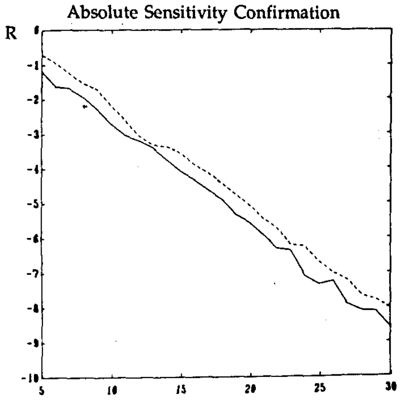
To illustrate the fact that the optimal realizations, both in shift operator form and in delta operator form, yield much smaller sensitivities than non-optimal realizations, we have also computed $\bar{M}_a$ for the shift operator and delta operator control canonical forms. These are, respectively,

$$\bar{M}_{a,z}^{c} = 81.9891 \quad \text{and} \quad \bar{M}_{a,\delta}^{c} = 5.1605 . \tag{4.21}$$

These theoretical results will now be confirmed by a numerical simulation on the same example. For both the optimal z-form realization $((A_z^{opt}, B_z^{opt}, C_z^{opt}))$ and the optimal δ-form realization $(A_\delta^{opt}, B_\delta^{opt}, C_\delta^{opt})$ presented above, we compute the corresponding frequency response $H_{fwl}^p(\omega)$ 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 :

$$R \overset{\Delta}{=} \log\left[\sup_{\omega \in (0,2\pi)} | H_{id}(\omega) - H_{fwl}^p(\omega) |\right]. \tag{4.22}$$

The results, for the example described above, are shown in Figure 1.



Number of bits in fractional part of coefficients

Fig. 1

----- The optimal realization in S_z

—— The optimal realization in S_δ

Fig. 1 clearly shows the superiority of the optimal δ-form realization over the optimal z-form realization whatever the number of bits.

5. RELATIVE SENSITIVITY OF POLYNOMIAL PARAMETRIZATIONS

The absolute sensitivity is a natural quantity to be used to compute the effects of absolute coefficient errors on the output of a filter, i.e. it is naturally connected to FWL implementations in fixed point. In floating point implementations it is more natural to study the effect of relative errors in the coefficients. This leads to a definition of a relative sensitivity measure.

5.1 Relative sensitivity measure

Definition 5.1

Let M and f(M) be as in Definition 4.1. We then denote

$$\frac{\delta f}{\delta M} = P, \text{ with } p_{ij} \overset{\Delta}{=} \frac{\partial f}{\partial m_{ij/m_{ij}}} = m_{ij}\frac{\partial f}{\partial m_{ij}} , \tag{5.1}$$

where p_{ij} denotes the $(i,j)^{th}$ element of the matrix P.

The relative sensitivity measure of the transfer function H(z) w.r.t. the parameters in the realization A_ρ, B_ρ, C_ρ is then defined as follows,

$$M_{r,\rho} \overset{\Delta}{=} \| \frac{\delta H}{\delta A_\rho} \|_1^2 + \| \frac{\delta H}{\delta B_\rho} \|_2^2 + \| \frac{\delta H}{\delta C_\rho} \|_2^2 . \qquad (5.2)$$

Just as for the absolute sensitivity measure, M_ρ is not easily computable. Therefore we replace it, similarly, by an upper bound. To justify this bound we introduce the following technical result.

Lemma 5.1

Let f(M) be as in Definitions 4.1 and 5.1. Then

$$\| \frac{\delta f}{\delta M} \|_F^2 \le \| \frac{\partial f}{\partial M} \|_F^2 \| M \|_F^2 \qquad (5.3)$$

Proof : The result follows directly as follows

$$\| \frac{\delta f}{\delta M} \|_F^2 = \sum_{i,j} \left(\frac{\partial f}{\partial m_{ij}} m_{ij} \right)^2$$

$$\le \sum_{i,j} \left(\frac{\partial f}{\partial m_{ij}} \right)^2 \sum_{i,j} (m_{ij})^2 = \| \frac{\delta f}{\delta M} \|_F^2 \| M \|_F^2 . \qquad (5.4)$$

The inequality follows from the fact that the expression on the right hand side contains all the terms of the left hand side plus additional positive terms.

♦

This result is now used to establish an upper bound for $M_{r,\rho}$:

$$M_{r,\rho} \le \| A_\rho \|_F^2 \| \frac{\partial H}{\partial A_\rho} \|_1^2 + \| B_\rho \|_F^2 \| \frac{\partial H}{\partial B_\rho} \|_2^2 + \| C_\rho \|_F^2 \| \frac{\partial H}{\partial C_\rho} \|_2^2$$

$$\le \| A_\rho \|_F^2 \| \frac{\partial H}{\partial B_\rho} \|_2^2 \| \frac{\partial H}{\partial C_\rho} \|_2^2 + \| B_\rho \|_F^2 \| \frac{\partial H}{\partial B_\rho} \|_2^2 + \| C_\rho \|_F^2 \| \frac{\partial H}{\partial C_\rho} \|_2^2$$

$$= \bar{M}_{r,\rho} . \qquad (5.5)$$

For $\rho = z$, we get

$$\bar{M}_{r,z} = \| A_z \|_F^2 \operatorname{tr}(W_c) \operatorname{tr}(W_0) + \| B_z \|_F^2 \operatorname{tr}(W_c) + \| C_z \|_F^2 \operatorname{tr}(W_0) \qquad (5.6)$$

To compute the upper bound for $r = \delta$ and compare it with $\bar{M}_{r,z}$, we use (4.12)-(4.13) and we note from (3.12) that

$$\| A_\delta \|_F = \Delta^{-1} \| A_z - I \|_F$$

$$\| B_\delta \|_F = \Delta^{-1} \| B_z \|$$

$$\| C_\delta \|_F = \| C_z \|_F \qquad (5.7)$$

Therefore :

$$\bar{M}_{r,\delta} = \| A_z - I \|_F^2 \operatorname{tr}(W_c) \operatorname{tr}(W_0) + \| B_z \|_F^2 \operatorname{tr}(W_0) + \| C_z \|_F^2 \operatorname{tr}(W_c) \qquad (5.8)$$

A major difference between the expressions (4.14) and (5.8) for the absolute and relative sensitivity upper bounds is that Δ does not appear in the latter; this is of course due to the fact that the computation of relative sensitivities introduces a normalization (see (5.1)) in which Δ disappears.

Just as in (4.14) it is important to note that the matrices appearing in (5.8) are all computed from the shift operator realization (A_z, B_z, C_z), and that $\bar{M}_{r,\delta}$is the upper bound for the corresponding (A_δ, B_δ, C_δ) realization obtained from (A_z, B_z, C_z) via (3.12). This allows for the following comparison :

$$\bar{M}_{r,\delta} - \bar{M}_{r,z} = (\| A_z - I \|_F^2 - \| A_z \|_F^2) \operatorname{tr}(W_c) \operatorname{tr}(W_0) . \qquad (5.9)$$

A common practice, strongly recommended by Middleton and Goodwin [3], is to choose the sampling frequency f_s between 10 and 50 times the pass band of the system's transfer function. In such case, the poles lie within the shaded region depicted in Fig.2. For reasons that we let the reader guess, we shall call this the MG region.

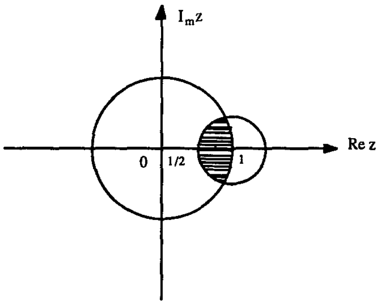


Fig.2. The MG region.

We now show that, if the sampling period is chosen according to Middleton and Goodwin's prescription, then $\bar{M}_{r,\delta}$ is smaller than $\bar{M}_{r,z}$

Theorem 5.1

Let the poles of a system (A_z, B_z, C_z) be in the MG region depicted in Fig.2. Then

$$\bar{M}_{r,\delta} < \bar{M}_{r,z} \qquad (5.10)$$

Proof : See the journal version of this paper.

This result has important implications. It shows that if the sampling period is chosen "reasonably" (i.e. the poles are in the MG region) and if a shift operator state space realization is chosen to optimize the relative sensitivity upper bound $\bar{M}_{r,z}$, then the corresponding delta operator realization obtained from (3.12) will always yield a smaller sensitivity upper bound $\bar{M}_{r,\delta}$. The minimization of $\bar{M}_{r,z}$ and $\bar{M}_{r,\delta}$ w.r.t. to all similarity transformations is much harder than the corresponding minimization of $\bar{M}_{a,z}$ and $\bar{M}_{a,\delta}$, and, to our knowledge, this problem has not been solved yet.

5.2. Numerical examples

We now illustrate our assertions with two numerical examples.

The computations of $\bar{M}_{r,z}$ and $\bar{M}_{r,\delta}$ for the four realizations described in Section 4.4 yield the following results :
- for the control canonical forms in z and δ :

$$\bar{M}_{r,z} = 308.8346, \quad \bar{M}_{r,\delta} = 40.0618$$

- for the forms that optimize $\bar{M}_{a,z}$ and $\bar{M}_{a,\delta}$:

$$\bar{M}_{r,z} = 4.7428, \quad \bar{M}_{r,\delta} = 4.0420 .$$

The next example is Example 5.2 from [11]. In that paper, the system is described in transfer function form. We note that all the poles are in the MG region. Again, we have computed the two control canonical forms, as well as the z- and δ-operator realizations that minimize $\bar{M}_a$. For these four realizations we have computed the relative sensitivity upper bound. This yields the following results :

- for the control canonical forms in z and δ :

$$\bar{M}_{r,z} = 1.2836\text{x}10^{11}, \quad \bar{M}_{r,\delta} = 58.1231$$

- for the forms that optimize $\bar{M}_{a,z}$ and $\bar{M}_{a,\delta}$:

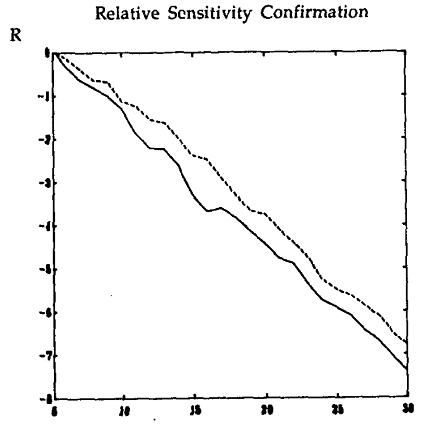
$$\bar{M}_{r,z} = 13.6537, \quad \bar{M}_{r,\delta} = 1.3417 .$$

These results illustrate two facts : the forms that optimize the absolute sensitivity upper bound also yield good relative sensitivity property, and the δ-operator realizations yield better relative sensitivity behaviour than the corresponding z-operator realizations.

In order to validate the upper bound $\bar{M}_r$ of the relative sensitivity measure, we have also computed the quantity

$$R = \log\left[\sup_{\omega\in(0,2\pi)} |H_{id}(\omega) - H^{p}_{fwl}(\omega)|\right]$$

for the last two realizations, respectively, where $H_{id}(\omega)$ is the ideal (or infinite precision) frequency response corresponding to the given state space model, and $H^{p}_{fwl}(\omega)$ is the frequency response obtained when the coefficients are implemented in floating point with p bits for the mantissa, p ranging from 5 to 30. We call R_Z the z-operator state space form that optimizes $\bar{M}_{a,z}$ and R_δ the δ-operator state space form that optimizes $\bar{M}_{a,z}$. The results are presented in Fig. 3. They confirm the superiority of the δ-operator model.



Number of Bits in Mantissa
Fig. 3.

---- The realization R_Z in S_Z which minimizes $\bar{M}_{a,z}$

—— The realization R_δ in S_δ which minimizes $\bar{M}_{a,\delta}$

6. CONCLUSIONS

Our aim in this paper has been to compare shift operator and delta operator state space parametrizations in terms of the effects of both absolute and relative finite wordlength errors on the actual transfer function. We have therefore defined an absolute and a relative sensitivity measure, and we have computed their expressions for these two types of parametrizations. Given that these expressions are difficult to handle, we have replaced them by reasonable and more tractable upper bounds.

In terms of the relative sensitivity (relevant for floating point computations) our results are clear-cut : they show that delta operator parametrizations will always yield a smaller upper bound than shift operator models for reasonable choices of sampling periods. As for the absolute sensitivity (relevant for fixed point computations), we have computed the set of optimizing realizations in both classes of parametrizations, and shown clear-cut comparison cannot be made, but all examples indicate that our design parameter Δ can always be chosen less than 1, in which case the delta parametrizations again achieve better upper bounds.

REFERENCES

[1] V. Peterka, "Control of uncertain processes : applied theory and algorithms", Kybernetika, vol.22, 1986, pp.1-102.
[2] R.H. Middleton and G.C. Goodwin,"Improved finite wordlength characteristics in digital control using delta operators", IEEE Trans. Auto. Control, vol.AC-31, Nov.1986, pp.1015-1021.
[3] R.H. Middleton and G.C. Goodwin,"Digital estimation and control : a unified approach", Prentice Hall, 1990.
[4] V. Tavsanoglu and L. Thiele (1984), "Optimal Design of State-Space Digital Filters by Simultaneous Minimization of Sensitivity and Roundoff Noise", IEEE Trans. on Circuits ans Systems, Vol. CAS-31, No 10 - Oct., pp.884-888.
[5] W.J. Lutz and S. Louis Hakimi (1988), "Design of Multi-Input Muiti-Output Systems with Minimum Sensitivity", IEEE. Trans. on Circuits and Systems, Vol.-35, No 9 - Sept., pp.1114-1122.
[6] L. Thiele (1984), "Design of Sensitivity and Roundoff Noise Optimal State-Space Discrete Systems", Int. J. Circuit Theory Appl., Vol-12, pp.39-46.
[7] Li Gang and M. Gevers, "Optimal finite precision implementation of a state-estimate feedback controller", to appear in IEEE Trans. on Circuits and Systems.
[8] Li Gang and M. Gevers, "Polynomial operators and equivalent computation structures with some applications", in preparation.
[9] S.Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering", IEEE Trans. on Acoustics, Speech and Signal Processing, vol. ASSP-25, Aug.1977, pp.273-281.
[10] D. Williamson and K. Kadiman, "Optimal Finite Wordlength Linear Quadratic Regulation", IEEE Trans. on Auto. Control, Vol. 34, No 12 - Dec. 1989, pp.1218-1228.
[11] D. Williamson, "Delay Replacement in Direct Form Structures", IEEE Trans. on Acoustics, Speech and Signal processing, Vol.36, No 4 - April 1988, pp.453-460.
[12] L. Thiele, "On the Sensitivity of Linear State-Space Systems", IEEE Trans. on Circuits ans Systems, Vol. CAS-33, No.5, May 1986.