Corrections

C47

SENSITIVITY AND ROUNDOFF NOISE OPTIMIZATION OF A STATE-ESTIMATE FEEDBACK CONTROLLER

Li GANG and Michel GEVERS
Laboratoire d'Automatique et d'Analyse des Systèmes
Louvain University, Bâtiment Maxwell
B-1348 Louvain-la-Neuve, BELGIUM

Abstract. Expressions are derived for the sensitivity and the roundoff noise gain of the closed loop transfer function of a system when the state estimate feedback controller is implemented with a finite word length and when the computations are performed in finite precision. The set of state space models minimizing either the sensitivity or the roundoff noise gain is computed.

Keywords. Finite wordlength effects, roundoff noise, sensitivity.

1. INTRODUCTION

The deterioration of the performances of a realization of a digital filter due to finite wordlength (FWL) can be separated into two effects : one is due to the finite wordlength implementation of the coefficients of the filter, the other is due to roundoff of the signals after every arithmetical operation. The first effect can be measured by a global sensitivity measure of the filter transfer function w.r.t. all the parameters, the other by the roundoff noise gain. In [1]-[2], expressions have been obtained for the roundoff noise gain of a state-variable implementation of a given discrete-time transfer function, and in [3] a global sensitivity measure of the transfer function w.r.t. the parameters of the state space model was proposed, and a reasonable upper bound was computed. It was also shown that, under a dynamic range constraint on the states, the roundoff noise and this sensitivity bound could be simultaneously optimized w.r.t. all equivalent state-space realizations.

Here we solve a more complicated problem using the same philosophy : we study the effects of finite wordlength and roundoff errors in the digital implementation of a pole placement state-estimate feedback controller. We first derive expressions for the roundoff noise gain and for a global sensitivity measure of the closed loop transfer function w.r.t. the parameters of the observer-controller. We then give a constructive procedure for the computation of the optimal realization sets, that minimize, respectively, the sensitivity and the roundoff noise gain of the closed loop system w.r.t. all similar state-variable observer-controller realizations.

The theoretical results have been tested on a numerical example. The sensitivity and the roundoff noise gain of the closed loop transfer function have been computed for a companion form realization of the observer, a particular δ-operator realization (see [4]) and the optimal realization (i.e. the realization that minimizes the roundoff noise gain of the closed loop system transfer function). The sensitivity of the optimal form can be an order of magnitude better than that of the companion form, and is comparable to that of the δ-operator form. The roundoff noise gain can be several orders of magnitude better than that of the companion form, and is better than that of the δ-operator form.
We should note that our sensitivity measure, in line with those used in [1]-[3], is based on implementation of unscaled parameters in fixed point arithmetic. The δ-operator realizations may well prove to have better performance when floating point arithmetic is used.

2. SENSITIVITY AND ROUNDOFF NOISE GAIN: PRELIMINARIES

Our aim in this paper is to study the effect of various state-estimate feedback implementations on the sensitivity and the roundoff noise gain of a closed loop system. To set up the notations and introduce the problem, we briefly review in this section the concepts of sensitivity measure, roundoff noise gain and dynamic range constraint for the finite precision state-variable implementation of a given filter.

Consider a discrete scalar transfer function :

$$H(z) = \frac{\sum_{i=0}^{n} b_i z^{-i}}{1 + \sum_{i=1}^{n} a_i z^{-i}} \qquad (2.1)$$

and a minimal state-space realization of H(z) :

$$x(k+1) = A\,x(k) + B\,u(k)$$
$$y(k) = C\,x(k) + D\,u(k) \qquad (2.2)$$

with A in $\mathbb{R}^{n\times n}$, B in $\mathbb{R}^{n}$, C^T in $\mathbb{R}^{n}$ and D in $\mathbb{R}$. The transfer function can be expressed in terms of state matrices as

$$H(z) = C(zI-A)^{-1}B + D \qquad (2.3)$$

If the coefficients in A, B, C, D are implemented in finite wordlength (FWL), the transfer function H(z) computed from (2.3) will deviate from its required value. The amount of this deviation can be measured by the sensitivity of the system transfer function H(z) w.r.t. the coefficients of the matrices A, B, C. Here we present a sensitivity measure proposed in [3], which has proved to be operational. It is based on an implementation of the unscaled parameters in fixed point; alternative implementations using scaled parameters have been discussed in [4].

Definition 2.1.

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

$$\frac{\partial f}{\partial M} = S \quad \text{with } s_{ij} \overset{\Delta}{=} \frac{\partial f}{\partial m_{ij}} \qquad (2.4)$$

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

Definition 2.2

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

$$\| f \|_p \triangleq \left(\frac{1}{2\pi} \int_0^{2\pi} \| f(e^{j\omega}) \|_F^p \, d\omega \right)^{1/p} \qquad (2.5)$$

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

$$\| f(e^{j\omega}) \|_F = \left(\sum_{i=1}^{n} \sum_{k=1}^{m} | f_{ik}(e^{j\omega}) |^2 \right)^{1/2} \qquad (2.6a)$$

$$= (\mathrm{tr}(f^T(e^{-j\omega}) f(e^{j\omega})))^{1/2} \qquad (2.6b)$$

The overall sensitivity measure of the transfer function H(z) w.r.t. the parameters in the realization A, B, C is then defined as follows in [3] :

$$M_S = \| \frac{\partial H}{\partial A} \|_1^2 + \| \frac{\partial H}{\partial B} \|_2^2 + \| \frac{\partial H}{\partial C^T} \|_2^2 \qquad (2.7)$$

Using the Cauchy-Schwartz inequality, it can easily be shown (see [3]) that an upper bound for M_S is given by

$$M = \| \frac{\partial H}{\partial B} \|_2^2 \, \| \frac{\partial H}{\partial C} \|_2^2 + \| \frac{\partial H}{\partial B} \|_2^2 + \| \frac{\partial H}{\partial C^T} \|_2^2 \qquad (2.8a)$$

$$= \mathrm{tr}W_0 \mathrm{tr}W_c + \mathrm{tr}W_0 + \mathrm{tr}W_c \qquad (2.8b)$$

where W_0 and W_c are, respectively, the observability and controllability Gramians of the realization (A,B,C).

A similarity transformation x = Tz transforms { A, B, C, W_c, W_0} into { $T^{-1}AT$, $T^{-1}B$, CT, $T^{-1}W_c T^{-T}$, $T^T W_0 T$}+ . An obvious problem is then to search for a choice of coordinates (i.e. a similarity transformation T) that minimizes the sensitivity measure M_S. Instead, one solves the easier problem of minimizing the upper bound M : see [5].

Limited wordlength effects on the signals cause another source of error on the output y(k) of the realization (2.2) which is known as roundoff noise : this is due to the fact that the signals are rounded off after each arithmetic operation. Assuming that the roundoff residue sequence can be modeled as zero mean white noise, then the roundoff noise gain G of the realization (2.2) can be shown to be (see [1], [2]) :

$$G = \mathrm{tr}W_0 \qquad (2.9)$$

where W_0 is the observability Gramian. The problem of minimizing the roundoff noise gain G over all equivalent minimal state space realizations of H(z) can therefore be formulated as follows : given an arbitrary minimal realization (A, B, C) find :

$$\min_T G = \min_T \mathrm{tr}\tilde{W}_0 \qquad (2.10)$$

where $\tilde{W}_0 = T^T W_0 T$. As such, the solution appears to depend only on the choice of C and A. In fact, the problem (2.10) does not make much sense unless a scaling of the states is introduced.

In order to maintain the amplitudes of the states within an acceptable range, and hence to reduce the probability of overflow, a l_2-norm scaling is introduced : a similarity tranformation T is performed such that in the new coordinate system

$$(\tilde{W}_c)_{ii} = (T^{-1}W_c T^{-T})_{ii} = 1 \qquad i = 1,\ldots,n \qquad (2.11)$$

i.e. the controllability Gramian has its diagonal elements all equal to unity. This has the effect of giving an equal probability of overflow to all components of the state. The minimum is achieved by a set of optimal realizations, all of which satisfy the dynamic range constraint. A constructive procedure for computing this optimal realization set has been given by Hwang [2].

3. FINITE PRECISION ASPECTS IN A CLOSED-LOOP COMPENSATOR : PROBLEM FORMULATION

The results published so far on finite precision implementations have turned around the following questions : given a filter (i.e. a transfer function) that must be implemented in finite precision, find a realization that minimizes either a sensitivity measure, or the roundoff noise gain of the filter or both. These results concern the FWL implementation of a given filter.

In this section we want to study the effects of finite wordlength implementation and finite precision arithmetic on the performances (sensitivity and roundoff noise gain) of a state-estimate feedback controller. We consider an open loop system specified by its transfer function $H_0(z)$: it is the given plant. Even though the plant is not implemented in a computer, it will be useful to think of $H_0(z)$ as being implemented by an infinite precision state-variable realization (A_0, B_0, C_0) in some coordinate system :

$$H_0(z) = C_0(zI - A_0)^{-1}B_0 \qquad (3.1)$$

Once a controller design and observer design strategy have been chosen (e.g. pole placement or LQ, Luenberger observer or Kalman filter), there results an (infinite precision) control gain K_0 and observer gain J_0. Their corresponding FWL implementations in the output feedback controller will be called K and J, respectively. For example K_0 is such that λ_i $(A_0 - B_0K_0)$ = set of desired closed loop poles, and K is the FWL implementation of K_0.

The block diagram of the plant $H_0(z)$ and its state-estimate feedback controller C(z) is given in Fig. 3.1.

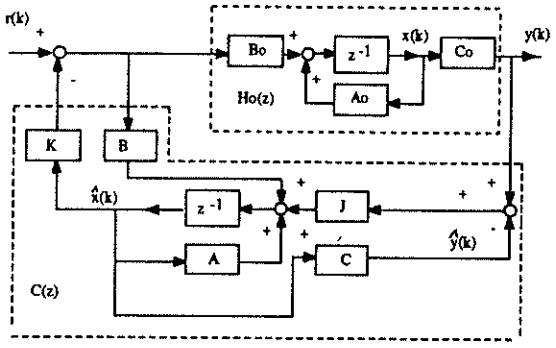

Fig.3-1.

We recall that (A,B,C,J,K) are implemented in FWL. In particular :

$$C(zI-A)^{-1}B \neq C_0(zI-A_0)^{-1}B_0 \qquad (3.2)$$

What will be of interest here are the effects of the FWL implementation of (A,B,C,K,J) and of the finite precision of the arithmetic calculations on the closed loop transfer function (from r to y) and on the closed loop roundoff noise gain.

The closed loop tranfer function $H_c(z)$ is a function of the actual system and of the compensator C(z) :

+ We denote $(T^{-1})^T$ by T^{-T}.

$$H_c(z) = F[H_0(z), (A,B,C,K,J)] \quad (3.3)$$

where (A,B,C,K,J) are the FWL implementations of (A_0,B_0, C_0, K_0, J_0). Under a similarity transformation, (A_0,B_0, C_0, K_0, J_0) is transformed into ($T^{-1}A_0T$, $T^{-1}B_0$, C_0T, K_0T, $T^{-1}J_0$) yielding a different FWL implementation.

Our task in the remainder of this paper is as follows : for a given $H_0(z)$, a desired set of closed loop poles and a desired set of observer poles, find

a) a computable measure of the sensitivity of the closed loop transfer function $H_c(z)$ w.r.t. the parameters of the realization (A,B,C,K,J);
b) the roundoff noise gain of the closed loop realization;
c) the realization set (i.e. the set of realizations (A,B,C,K,J)) that minimizes the sensitivity measure, the roundoff noise gain, or both, subject to a dynamic range constraint on the states of the observer.

4. SENSITIVITY MEASURE OF THE CLOSED LOOP SYSTEM

The state-equations of the closed loop system are

$$\begin{bmatrix} x(k+1) \\ \hat{x}(k+1) \end{bmatrix} = \begin{bmatrix} A_0 & -B_0K \\ JC_0 & A-BK-JC \end{bmatrix} \begin{bmatrix} x(k) \\ \hat{x}(k) \end{bmatrix} + \begin{bmatrix} B_0 \\ B \end{bmatrix} r(k) \quad (4.1)$$

$$y(k) = [C_0 \quad 0] \begin{bmatrix} x(k) \\ \hat{x}(k) \end{bmatrix} \quad (4.2)$$

We denote :

$$\bar{A} \triangleq \begin{bmatrix} A_0 & -B_0K \\ JC_0 & A-BK-JC \end{bmatrix} \quad \bar{B} \triangleq \begin{bmatrix} B_0 \\ B \end{bmatrix} \quad \bar{C} \triangleq [C_0 \quad 0] \quad (4.3)$$

The closed-loop transfer function is

$$H_c(z) = \bar{C}(zI-\bar{A})^{-1}\bar{B} \quad (4.4)$$

Using the notations introduced in Definition 2.1, we now compute the sensitivities of $H_c(z)$ w.r.t. A,B,C,K,J, evaluated at the exact (i.e. infinite precision) values A_0,B_0,C_0,K_0,J_0. After lengthy manipulations, the following expressions are obtained :

$$\frac{\partial H_c}{\partial A}(z) = -H_c^o(z)G_o(z)F_K^T(z) \quad (4.5a)$$

$$\frac{\partial H_c}{\partial B}(z) = -H_c^o(z)[1-H_k(z)]G_o(z) \quad (4.5b)$$

$$\frac{\partial H_c}{\partial C^T}(z) = -H_c^o(z)H_o(z)F_K(z) \quad (4.5c)$$

$$\frac{\partial H_c}{\partial K}(z) = -H_c^o(z)F_K(z) \quad (4.5d)$$

$$\frac{\partial H_c}{\partial J}(z) = 0 \quad (4.5e)$$

where $H_c(z)$ is the desired (or infinite precision) closed loop transfer function

$$H_c^o(z) = C_0(zI-A_0+B_0K_0)^{-1}B_0 \quad (4.6)$$

and where

$$G_o(z) \triangleq [zI-(A_0-J_0C_0)^T]^{-1}K_0^T \quad (4.7a)$$

$$F_K(z) \triangleq [zI-(A_0-B_0K_0)]^{-1}B_0 \quad (4.7b)$$

$$H_K(z) \triangleq K_0(zI-(A_0-B_0K_0)]^{-1}B_0 \quad (4.7c)$$

$$H_o(z) \triangleq K_0[zI-(A_0-J_0C_0)]^{-1}J_0 \quad (4.7d)$$

Note that (4.5e) does not mean that $H_c(z)$ is not a function of J. It means that the sensitivity of $H_c(z)$ w.r.t. J becomes nil when it is evaluated at the exact (A_0,B_0,C_0,K_0,J_0). The expressions (4.5a) to (4.5d) contain a common factor, which is precisely the desired closed loop transfer function. We define "normalized sensitivities" as follows for X = A,B,C or K :

$$\frac{\delta H_c}{\delta X} \triangleq \frac{1}{H_c^o(z)} \cdot \frac{\partial H_c(z)}{\partial X} = \frac{\partial \ln H_c(z)}{\partial X} \Big| A_0,B_0,C_0,K_0,J_0 \quad (4.8)$$

The normalized sensitivity $\delta H_c/\delta X$ is like the sensitivity of the Bode plot of $H_c(e^{j\omega})$ w.r.t. X.
With these definitions we get :

$$\frac{\delta H_c}{\delta A}(z) = -G_o(z)F_K^T(z) \quad (4.9a)$$

$$\frac{\delta H_c}{\delta B}(z) = -[1-H_K(z)]G_o(z) \quad (4.9b)$$

$$\frac{\delta H_c}{\delta C^T}(z) = -H_o(z)F_K(z) \quad (4.9c)$$

$$\frac{\delta H_c}{\delta K}(z) = -F_K(z) \quad (4.9d)$$

$$\frac{\delta H_c}{\delta J}(z) = 0 \quad (4.9e)$$

We now define the sensitivity of $\ln H_c(z)$ w.r.t. the parameters of A,B,C,K,J as (see (2.9) for comparison) :

$$M_S = \left\| \frac{\partial \ln H_c}{\partial A} \right\|_1^2 + \left\| \frac{\partial \ln H_c}{\partial B} \right\|_2^2 + \left\| \frac{\partial \ln H_c}{\partial C} \right\|_2^2 + \left\| \frac{\partial \ln H_c}{\partial K} \right\|_2^2 + \left\| \frac{\partial \ln H_c}{\partial J} \right\|_2^2 \quad (4.10)$$

An upper bound for this sensitivity is given by

$$M = \| G_o \|_2^2 \| F_K \|_2^2 + \| (1-H_K)G_o \|_2^2 + \| H_oF_K \|_2^2 + \| F_K \|_2^2 \quad (4.11)$$

M can be rewritten as

$$M = \text{tr } W_{oo} \text{ tr } W_{cc} + \text{tr } W_3 + \text{tr } W_4 + \text{tr } W_{cc} \quad (4.12)$$

where

$$W_{oo} = \frac{1}{2\pi j} \oint_{|z|=1} G_o(z)\, G_o^T(z^{-1})\, z^{-1} dz \quad (4.13.a)$$

$$W_{cc} = \frac{1}{2\pi j} \oint_{|z|=1} F_K(z)\, F_K^T(z^{-1})\, z^{-1} dz \quad (4.13.b)$$

and W_3 are W_4 are defined similarly. It follows from (4.7) and (4.13) that W_{oo} and W_{cc} are, respectively, the observability Gramian of the state observer and the controllability Gramian of the feedback controller. It is easy to compute the effect of similarity transformations on the Gramians W_{oo}, W_{cc}, W_3 and W_4 appearing in (4.12). The optimal sensitivity realization problem can then be stated as follows : given a particular realization A,B,C,K,J and the corresponding Gramians W_{oo}, W_{cc}, W_3, W_4, find the (set of) nonsingular transformations T such that

$$M = \text{tr}(T^TW_{oo}T)\text{tr}(T^{-1}W_{cc}T^{-T}) + \text{tr}(T^TW_3T) + \text{tr}(T^{-1}W_4T^{-T}) + \text{tr}(T^{-1}W_{cc}T^{-T}) \quad (4.14)$$

is minimized :

$$\min_{T} M \qquad (4.15)$$
$$\det T \neq 0$$

In the next section, we characterize the set of optimal transformations, i.e. the solution set of the problem (4.15).

5. MINIMIZATION OF THE CLOSED LOOP SENSITIVITY

We now solve the problem (4.15) with M defined by (4.14). First notice that the problem can be reformulated as follows :

$$M = \mathrm{tr}(T^T M_1^0 T)\mathrm{tr}(T^{-1} M_2^0 T^{-T}) + \mathrm{tr}(T^T M_3^0 T) + \mathrm{tr}(T^T M_3^0 T) + \mathrm{tr}(T^{-1} M_4^0 T^{-T}) \qquad (5.1)$$

where

$$M_1^0 = W_{oo},\ M_2^0 = W_{cc},\ M_3^0 = W_3,\ M_4^0 = W_4 + W_{cc} \qquad (5.2)$$

These are four positive definite matrices; the supercript 0 denotes the fact that they correspond to an arbitrary initial state space realization. First notice that there exists a nonsingular matrix T_0 such that :

$$T_0^T M_1^0 T_0 = \Sigma \qquad (5.3a)$$
$$T_0^{-1} M_2^0 T_0^{-T} = \Sigma \qquad (5.3b)$$

where

$$\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, ..., \sigma_n),\ \sigma_i > 0 \qquad (5.4)$$

This transformation matrix is unique up to a signature matrix. It transforms M_3^0 and M_4^0 into

$$M_3 \overset{\Delta}{=} T_0^T M_3^0 T_0,\quad M_4 \overset{\Delta}{=} T_0^{-1} M_4^0 T_0^{-T} \qquad (5.5)$$

Now let

$$T = T_0 T_1 \qquad (5.6)$$

where T_1 is nonsingular. Then T_1 can be written (see [2]) :

$$T_1 = R_1 \Pi R_0^T. \qquad (5.7)$$

where R_1 and R_0 are orthogonal matrices, and

$$\Pi = \mathrm{diag}(x_1^{1/2}, x_2^{1/2}, ..., x_n^{1/2}),\ x_i > 0. \qquad (5.8)$$

Therefore M can be rewritten as

$$M = \mathrm{tr}(R_1^T \Sigma R_1 \Pi^2)\ \mathrm{tr}(R_1^T \Sigma R_1 \Pi^{-2}) + \mathrm{tr}(R_1^T M_3 R_1 \Pi^2) + \mathrm{tr}(R_1^T M_4 R_1 \Pi^{-2}) \qquad (5.9)$$

$$= \sum_{i=1}^{n}(x_i k_{ii}) \sum_{i=1}^{n}(x_i^{-1} k_{ii}) + \sum_{i=1}^{n}(x_i q_{ii} + x_i^{-1} p_{ii}) \qquad (5.10)$$

where k_{ii}, q_{ii} and p_{ii} are the diagonal elements of

$$K \overset{\Delta}{=} R_1^T \Sigma R_1 = \{k_{ij}\} \qquad i,j = i, ...,n \qquad (5.11a)$$

$$Q \overset{\Delta}{=} R_1^T M_3 R_1 = \{q_{ij}\} \qquad i,j = i, ...,n \qquad (5.11b)$$

$$P \overset{\Delta}{=} R_1^T M_4 R_1 = \{p_{ij}\} \qquad i,j = 1, ..., n \qquad (5.11c)$$

Next we notice that the following constraints apply to K, Q and P :

$$\sum_{i=1}^{n} k_{ii} = \mathrm{tr}K = \mathrm{tr}\Sigma \overset{\Delta}{=} S_0 \qquad (5.12a)$$

$$\sum_{i=1}^{n} q_{ii} = \mathrm{tr}Q = \mathrm{tr}M_3 \overset{\Delta}{=} S_1 \qquad (5.12b)$$

$$\sum_{i=1}^{n} p_{ii} = \mathrm{tr}P = \mathrm{tr}M_4 \overset{\Delta}{=} S_2 \qquad (5.12c)$$

The optimization problem (4.15) can therefore be reformulated as follows :

$$\min M \text{ w.r.t. } \{x_i\}, \{k_{ii}\}, \{q_{ii}\}, \{p_{ii}\},\ i = 1,...,n$$
$$\text{subject to (5.11)–(5.12).} \qquad (5.13)$$

The solution of (5.13) has been derived in [6].
The optimal solution set, i.e. the set of similarity transformations T that minimize M in (5.1), is defined by

$$T = T_0 R_1 \Pi R_0^T \qquad (5.14)$$

where T_0 is (almost uniquely) defined by (5.3), R_0 is an arbitrary orthogonal matrix, Π is given by :

$$\Pi = \left(\frac{S_2}{S_1}\right)^{1/4} I \qquad (5.15)$$

and R_1 is any orthogonal matrix satisfying

$$(R_1^T M_3 R_1)_{ii} = \frac{S_1}{S_2}(R_1^T M_4 R_1)_{ii} \qquad i=1,..., n \qquad (5.16)$$

where $S_1 = \mathrm{tr}\, M_3$ and $S_2 = \mathrm{tr}\, M_4$. The existence of such R_1 is proved in [6].

6. ROUNDOFF NOISE GAIN OF THE CLOSED LOOP SYSTEM

Recall that the closed loop system is described by (4.1)-(4.2). We now consider the case where the estimated state $\hat{x}(k)$ is rounded off to b_0 bits before multiplication in (4.1), and we denote by $Q[x(k)]$ the quantized value of a vector x(k), rounded off to the first b_0 bits. The model (4.1)-(4.2) is then replaced by

$$\begin{bmatrix} x^*(k+1) \\ \hat{x}^*(k+1) \end{bmatrix} = \begin{bmatrix} A_0 & -B_0K \\ JC_0 & A-BK-JC \end{bmatrix} \begin{bmatrix} x^*(k) \\ Q[\hat{x}^*(k)] \end{bmatrix} + \begin{bmatrix} B_0 \\ B \end{bmatrix} r(k)$$

$$y^*(k) = [C_0\ 0] \begin{bmatrix} x^*(k) \\ Q[\hat{x}^*(k)] \end{bmatrix} \qquad (6.1)$$

Here we neglect the effect of roundoff on the signal r(k). Denote

$$E(k) \overset{\Delta}{=} \begin{bmatrix} x(k) - x^*(k) \\ \hat{x}(k) - \hat{x}^*(k) \end{bmatrix},\quad e(k) \overset{\Delta}{=} \begin{bmatrix} 0 \\ \hat{x}^*(k) - Q[\hat{x}^*(k)] \end{bmatrix} \qquad (6.2a)$$

$$\Delta y(k) \overset{\Delta}{=} y(k) - y^*(k) \qquad (6.2b)$$

It then follows from (6.1), (6.2) that

$$E(k+1) = \bar{A}E(k) + \bar{A}e(k) \qquad (6.3a)$$

$$\Delta y(k) = \bar{C}E(k) + \bar{C}e(k) \qquad (6.3b)$$

with $\bar{A}$ and $\bar{C}$ defined in (4.3). We assume that r(k) is sufficiently exciting so that e(k) can be modeled as a uniformly distributed zero mean uncorrelated random vector with

variance $\sigma^2 I$. It then follows from (6.3) that the roudoff noise gain of the closed loop system is

$$G \overset{\Delta}{=} \frac{1}{\sigma^2} \lim_{k\to\infty} E[\Delta^2 y(k)] = \frac{1}{\sigma^2} \mathrm{tr}[\bar{W}_o R] \tag{6.4}$$

where

$$R \overset{\Delta}{=} E[e(k)e^T(k)] = \begin{bmatrix} 0 & 0 \\ 0 & \sigma^2 I \end{bmatrix} \tag{6.5}$$

$$\bar{W}_o \overset{\Delta}{=} \sum_{k=0}^{\infty} (\bar{A}^T)^k \bar{C}^T \bar{C} \bar{A}^k \tag{6.6}$$

$\bar{W}_o$ is the observability Gramian of the closed loop system. It can be shown that this roundoff noise gain of the closed loop system is approximately given by

$$G \cong \mathrm{tr}\left[\frac{1}{2\pi j} \oint_{|z|=1} H_c^o(z) H_c^o(z^{-1}) G_o(z) G_o^T(z^{-1}) z^{-1} dz \right]$$

$$= \| H_c^o(z) G_o(z) \|_2^2 \tag{6.7}$$

where the approximation sign is there to indicate that the finite precision quantities have been replaced by infinite precision quantities in the computation of G. The roundoff noise of the output of the closed loop system can therefore be interpreted as white noise with variance σ^2 passing first through the observer dynamics, then filtered by the desired closed loop system $H_c^o(z)$.

7. MINIMIZATION OF THE ROUNDOFF NOISE GAIN UNDER DYNAMIC RANGE CONSTRAINT

In this section, we characterize the set of all state observer realizations that minimize the roundoff noise gain G of the closed loop system subject to an l_2-scaling on the observer states, which is meant to guarantee an equal probability of overflow. By the same token, we will give a constructive procedure for the computation of a realization that minimizes this roundoff noise gain.

Let $\bar{W}_c$ be the controllability Gramian of the closed loop system :

$$\bar{W}_c = \sum_{k=0}^{\infty} \bar{A}^k \bar{B} \bar{B}^T (\bar{A}^T)^k \tag{7.1a}$$

$$= \frac{1}{2\pi j} \oint_{|z|=1} \bar{F}(z) \bar{F}^T(z^{-1}) z^{-1} dz \tag{7.1b}$$

$$= \begin{pmatrix} \bar{W}_c(1,1) & \bar{W}_c(1,2) \\ \bar{W}_c(2,1) & \bar{W}_c(2,2) \end{pmatrix} \tag{7.1c}$$

where $\bar{A}$, $\bar{B}$ are defined by (4.3) and

$$\bar{F}(z) = (zI - \bar{A})^{-1} \bar{B} = \begin{bmatrix} f_1(z) \\ f_2(z) \end{bmatrix} \tag{7.2}$$

Here $f_1(z)$ and $f_2(z)$ are the first n and last n components of $\bar{F}(z)$. Imposing a l_2-scaling on the observer states $\hat{x}$ corresponds with finding a coordinate basis for $\bar{A}, \bar{B}, \bar{C}$ in which

$$(\bar{W}_c(2,2))_{i,i} = 1 \text{ for } i = 1,2,\ldots,n \tag{7.3}$$

Replacing again (A,B,C,K,J) in $\bar{A}$ by the infinite precision quantities (A_0,B_0,C_0,K_0,J_0) and calling the resulting matrix $\bar{A}_o$ yields

$$f_2(z) = (zI - A_0 + B_0 K_0)^{-1} B_0 = F_K(z) \text{ (see 4.7b)} \tag{7.4}$$

Therefore :

$$\bar{W}_c(2,2) = \frac{1}{2\pi j} \oint_{|z|=1} F_K(z) F_K^T(z^{-1}) z^{-1} dz \tag{7.5a}$$

$$= W_{cc} \tag{7.5b}$$

with W_{cc} as defined by (4.13b). The minimization of the roundoff noise gain subject to the dynamic range constraint can therefore be formulated as follows :

$$\min_T \mathrm{tr}(T^T W T) \tag{7.6a}$$

$$\det T \neq 0$$

subject to

$$(T^{-1} W_{cc} T^{-T})_{ii} = 1 \qquad i = 1,\ldots,n \tag{7.6b}$$

where

$$W = \frac{1}{2\pi j} \oint_{|z|=1} H_c^o(z) H_c^o(z^{-1}) G_o(z) G_o^T(z^{-1}) z^{-1} dz \tag{7.7a}$$

$$W_{cc} = \frac{1}{2\pi j} \oint_{|z|=1} F_K(z) F_K^T(z^{-1}) z^{-1} dz \tag{7.7b}$$

To solve this problem, we follow the procedure of [2]. Given an arbitrary initial realization $(A^0, B^0, C^0, K^0, J^0)$ and the corresponding Gramians W^0 and W_{cc}^0 defined by (7.7), we first compute a square root factor of W_{cc}^0 :

$$W_{cc}^0 = T_0 T_0^T \tag{7.8}$$

Notice that T_0 is not unique. We denote by σ_i the singular values of the product $W_{cc}^0 W$. The optimal realization set S_{opt} is obtained from the initial realization $(A^0, B^0, C^0, K^0, J^0)$ by the set of similarity transformations

$$T_{opt} = T_0 T_1 = T_0 R_1 \Pi R_0^T \tag{7.9}$$

where T_0 is defined by (7.8) and R_1, Π and R_0 by

$$\Pi = \mathrm{diag}(\Pi_i), \quad \Pi_i = \left(\frac{\sum_{m=1}^{n} \sigma_m}{n\sigma_i} \right)^{1/2} \qquad i = 1,\ldots,n \tag{7.9a}$$

where R_0 and R_1 are orthogonal matrices $(R_0 R_0^T = R_0^T R_0 = 1)$ satisfying :

$$(R_0 \Pi^{-2} R_0^T)_{ii} = 1 \qquad i = 1, \ldots, n \tag{7.9b}$$

$$R_1^T W^1 R_1 = \Sigma^2 = \mathrm{diag}(\sigma_i^2) \quad i = 1,\ldots,n \tag{7.9c}$$

The existence of such R_0 has been proved in [2], where it was also shown that R_0 is not unique. However, it is unclear how to parametrize the freedom in R_0. An important remark is that, with T_{opt} defined by (7.9), the upper bound M on the sensitivity is independent of R_0 : see (4.14).

8. A NUMERICAL EXAMPLE

In this section we present an example that illustrates the typical improvement in accuracy obtained by the optimal realization in comparison with two other widely used realizations, a

companion form and a δ-form. We refer to Middleton and Goodwin (1987) for a presentation and a thorough discussion of δ-form realizations.

Let the system to be controlled be given by

$$H_o(z) = \frac{0.0022\ (z+1)^2}{(z-0.9588)\ (z-0.9231)\ (z-0.8763)}$$

Let the desired closed loop poles be $\lambda_i(A_0 - B_0K_0) = 0.9067, 0.7523, 0.6231$ and let the poles of the observer be $\lambda_i(A_0 - J_0C_0) = 0.4532, 0.5761, 0.8437$.

We compare the sensitivity and the roundoff noise gain, computed by using the formulae (4.12) and (6.4), for a control canonical realization, a delta form realization and the realizations that minimize the sensitivity of the closed loop transfer function and its roundoff noise gain, respectively. For this third order system, the control canonical realization takes the form

$$A_c = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 \end{bmatrix} \qquad B_c = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \qquad C_c = [c_1\ c_2\ c_3]$$

The delta realization (with $\delta = \frac{z-1}{T_S}$, where T_S is the sampling period) takes the form

$$A_\delta = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ -d_0 & -d_1 & 1-d_2 \end{bmatrix} \qquad B_\delta = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \qquad C_\delta = [e_1\ e_2\ e_3]$$

with $T_S = 1$.The optimal realizations are typically fully parametrized.

Consider first the case where there are no scaling constraints on the states of the realizations, and where the closed loop sensitivity is minimized using the procedure of section 5. We then obtain the following values for the closed loop sensitivities of the optimal, control canonical and δ realizations, respectively (The corresponding roundoff noise gains are also indicated) :

M_{opt}	$= 12.4557$	$G = 2.4265$
M_c	$= 1.444 \times 10^4$	$G_c = 3.3268$
M_δ	$= 1.9268 \times 10^3$	$G_\delta = 0.6616$

We now consider the case where the roundoff noise gain of the closed loop system is optimized with l_2 scaling using the procedure of section 7. We compare the roundoff noise gain of the optimal structure with that of the l_2 scaled control canonical form and delta form, noting that these are obtained from the unscaled realizations by a suitable diagonal transormation. Their canonical structure is thereby not preserved, except for the zeroes which remain in the same positions. We then obtain the following values for the roundoff noise gain of the l_2 scaled optimal, control canonical and δ realizations, respectively (The corresponding sensitivities are also given) :

$G^{(s)}_{opt} = 0.3811$	$M = 37.3962$
$G^{(s)}_c = 1.5006 \times 10^3$	$M^{(s)}_c = 1.1914 \times 10^4$
$G^{(s)}_\delta = 1.0305$	$M^{(s)}_\delta = 26.4696$

Comments

We notice that the sensitivity of the realization that minimizes the sensitivity is several orders of magnitude smaller than the sensitivities of the companion form realization and of the δ-form realization, while the corresponding roundoff noise gains are of the same order of magnitude, with a certain superiority for the δ-form realization.

On the other hand, when it comes to the objective of minimizing the roundoff noise gain under dynamic range constraint, superiority of the optimal realization over that derived from a δ-form is only marginal, both of them being several orders of magnitude better than the l_2 scaled companion form realization.

These results are rather typical of other examples we have studied.

9. CONCLUSIONS

We have derived expressions for the sensitivity and the roundoff noise gain of a closed-loop transfer function w.r.t. the parameters of a state-variable realization when the regulator is a state-estimate feedback regulator implemented in finite wordlength with the computations also being performed in finite arithmetic. We have then computed the set of optimal realizations, i.e. the set of realizations that optimize either the sentivity, or the roundoff noise gain under dynamic range constraint.

We have illustrated with a numerical example the typical accuracy gains that can be achieved by the optimal realizations as compared to a companion form in the shift operator or a companion form in the δ operator. We should note that, in line with the results of [1]-[3] for open loop transfer functions, our computations have all been performed using fixed point arithmetic without scaling of the parameters. Extensions using either a scaling of the parameters or floating point arithmetic are presently being studied. We also note that, in our numerical example, the optimal realization shows to yield superior performance when compared to a δ-form implemented in companion form (i.e. δ-companion form). This particular δ-form is clearly not optimal among all possible δ-forms. A further extension, recently suggested by Goodwin [7], is to compute the optimal realizations for both sensitivity and roundoff noise, among the set of all δ-operator state variable representations. These optimal δ-operator realizations might well prove to have better performance with floating point arithmetic than the optimal shift operator realizations. These questions are the subject of continuing investigations.

REFERENCES

[1] C.T. Mullis and R.A. Roberts (1976), "Filter structures which minimize roundoff noise in fixed-point digital filters", Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing, pp. 505-508.

[2] S.Y. Hwang (1977), "Minimum Uncorrelated Unit Noise in State-Space Digital Filtering" IEEE Trans. on Acoust. Speech and Signal Processing, Vol. ASSP-25, No 4 - Aug., pp.273-281.

[3] 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 and Systems, Vol. CAS-31, No 10 - Oct., pp. 884-888.

[4] R.H. Middleton and G.C. Goodwin (1987), "Digital Estimation and Control : A Unified Approach" Part II. Department of Electrical and Computer Engineering, University of Newcastle, N.S.W. Australia.

[5] L. Thiele (1984), "Design of Sensitivity and Roundoff Noise Optimal State-Space Discrete Systems", Int. J. Circuit Theory Appl., Vol-12, pp 39-46.

[6] Li Gang and M. Gevers, "Optimal finite precision implementation of a state-estimate feedback controller", submitted for publication.

[7] G.C. Goodwin (1989), personal communication.