Optimal Finite Precision Implementation of a State-Estimate Feedback Controller

GANG LI AND MICHEL GEVERS, FELLOW, IEEE

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. This allows one to compare the sensitivity and noise gain over all equivalent state variable implementations of the observer plus controller. The set of state-space models minimizing either the sensitivity or the roundoff noise gain is computed. Simulations compare the performance of the optimal implementations with respect to companion forms and a particular delta model.**

I. INTRODUCTION

MUCH attention has recently been paid to the problem of optimal design of finite word length (FWL) state-variable implementations of linear time invariant filters [1]–[4]. The problem of characterizing all state-space realizations that minimize the roundoff noise gain has been solved in [1], [2]. 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 and easily computable upper bound for this measure was proposed. It was shown that, under a dynamic range constraint, this sensitivity bound and the roundoff noise gain are simultaneously optimized and the set of optimizing structures was characterized. In [4], the scalar sensitivity measure proposed in [3] was further extended to continuous time multivariable filters, and the set of state-variable models minimizing the same upper bound was again characterized. The sensitivity measures computed in [1]–[4] are all based on implementation of unscaled parameters in fixed point arithmetic.

Most of the results described so far have a signal processing flavor, i.e., the objective is to minimize the FWL effects of the filter coefficients implementation and the roundoff error effects of the finite arithmetic on the output of the filter. By comparison, the effects of FWL and roundoff errors in the digital implementation of controllers has received much less attention. One notable exception is the so-called δ-operator implementation proposed by Middleton and Goodwin [5], [6]. One of the advantages of δ-forms as opposed to shift polynomial forms is to avoid the clustering of poles around one at high sampling rates, and hence to increase the numerical accuracy of the calculations in finite precision.

Here we study the problem of state estimate feedback design, where the state-variable observer and the feedback gain are implemented in finite precision and where the arithmetic computations are rounded off. One contribution of our paper is to compute the sensitivity of the closed-loop transfer function w.r.t. the parameters of the observer–controller, and the roundoff noise gain of the closed-loop system as a function of these parameters. Another is to compute the optimal realization sets that minimize either this sensitivity or this roundoff noise gain under the usual dynamic range constraint. Our sensitivity measure is the same as that proposed in [3], and our results are therefore an extension of [3] and [4] to the feedback control problem. The design procedure is illustrated with a numerical example; the optimal realization is compared with the companion form and a δ-form realization.

The paper is organized as follows. In Section II, we introduce the definitions of sensitivity, roundoff noise gain, and dynamic range constraint for the state-variable realization of a filter, and we recall the most relevant results of [1]–[4]. In Section III, we formulate the optimal finite precision realization problem in the case where the model is used to compute a feedback compensator. Our first new contribution is in Section IV, where we derive a computable expression for the sensitivity of the transfer function of a closed-loop system w.r.t. the parameters of a state-space realization. In Section V we give a procedure for the computation of the realization set that optimizes this closed-loop sensitivity. The roundoff noise gain of the closed-loop system is computed in Section VI, while the set of realizations that minimize this closed-loop roundoff noise gain under dynamic range constraint is derived in Section VII. Finally, a computational example in Section VIII illustrates the differences in performance that can be obtained using an optimal realization compared with more commonly used realizations such as a companion form or a particular delta-operator realization.

Manuscript received September 27, 1989; revised June 13, 1990. The results presented in this paper have been obtained within the framework of the Belgian Program on Concerted Research Actions and on the Interuniversity Attraction Poles initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. This paper was recommended by Associate Editor S. B. Haley.

The authors are with the Laboratoire D'automatique, de Dynamique et D'analyse des Systemes, Louvain University, B-1348, Louvain-la-neuve, Belgium.

IEEE Log Number 9039220.

0098-4094/90/1200-1487$01.00

II. 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.

A. Sensitivity Measure

Consider a discrete scalar transfer function:

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

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

$$\begin{aligned} x(k+1) &= Ax(k)+Bu(k) \\ y(k) &= Cx(k)+Du(k) \end{aligned} \tag{2.2}$$

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

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

If the coefficients in A,B,C,D are implemented in FWL, the output sequence $y(k)$ will deviate from its required value, and 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. There are of course several ways of defining an overall sensitivity measure. Here we present a 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 [5] and [6].

Definition 2.1: Let $M\in\mathbb{R}^{n\times m}$ 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 define

$$\frac{\partial f}{\partial M}=S \text{ with } s_{ij}\triangleq\frac{\partial f}{\partial m_{ij}} \tag{2.4}$$

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

With these notations it is easy to show that

$$\begin{aligned} S_A(z) &\triangleq \frac{\partial H(z)}{\partial A}=G(z)F^T(z) \\ S_B(z) &\triangleq \frac{\partial H(z)}{\partial B}=G(z) \\ S_C(z) &\triangleq \frac{\partial H(z)}{\partial C^T}=F(z) \end{aligned} \tag{2.5}$$

where

$$\begin{aligned} F(z) &\triangleq (zI-A)^{-1}B=[f_1(z)\cdots f_n(z)]^T \\ G^T(z) &\triangleq C(zI-A)^{-1}=[g_1(z)\cdots g_n(z)]. \end{aligned} \tag{2.6}$$

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

$$\|f\|_p\triangleq\left(\frac{1}{2\pi}\int_0^{2\pi}\|f(e^{j\omega})\|_F^p\,d\omega\right)^{1/p} \tag{2.7}$$

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

$$\|f(e^{j\omega})\|_F=\left(\sum_{i=1}^{n}\sum_{k=1}^{m}|f_{ik}(e^{j\omega})|^2\right)^{1/2} \tag{2.8a}$$

$$=\left(\operatorname{tr}\left(f^T(e^{-j\omega})f(e^{j\omega})\right)\right)^{1/2}. \tag{2.8b}$$

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]:

$$M_S=\left\|\frac{\partial H}{\partial A}\right\|_1^2+\left\|\frac{\partial H}{\partial B}\right\|_2^2+\left\|\frac{\partial H}{\partial C^T}\right\|_2^2. \tag{2.9}$$

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

$$M=M_A+M_B+M_C \tag{2.10}$$

where

$$\begin{aligned} M_B=\|G\|_2^2 &= \frac{1}{2\pi}\int_0^{2\pi}G^T(e^{-j\omega})G(e^{j\omega})\,d\omega \\ &= \operatorname{tr}\left(\frac{1}{2\pi j}\oint_{|z|=1}G(z)G^T(z^{-1})z^{-1}\,dz\right) \\ &= \operatorname{tr}\left(\sum_{k=0}^{\infty}(A^T)^kC^TCA^k\right)=\operatorname{tr}W_o \end{aligned} \tag{2.11a}$$

$$\begin{aligned} M_C=\|F\|_2^2 &= \frac{1}{2\pi}\int_0^{2\pi}F^T(e^{-j\omega})F(e^{j\omega})\,d\omega \\ &= \operatorname{tr}\left(\frac{1}{2\pi j}\oint_{|z|=1}F(z)F^T(z^{-1})z^{-1}\,dz\right) \\ &= \operatorname{tr}\left(\sum_{k=0}^{\infty}A^kBB^T(A^T)^K\right)=\operatorname{tr}W_c \end{aligned} \tag{2.11b}$$

$$M_A=\|G\|_2^2\|F\|_2^2=\operatorname{tr}W_o\operatorname{tr}W_c. \tag{2.11c}$$

It follows that

$$M=\operatorname{tr}W_o\operatorname{tr}W_c+\operatorname{tr}W_o+\operatorname{tr}W_c \tag{2.12}$$

where W_o and W_c are the observability Gramian and the controllability Gramian, respectively.

A similarity transformation $x=Tz$ transforms $\{A,B,C,W_c,W_o\}$ into $\{T^{-1}AT,T^{-1}B,CT,T^{-1}W_cT^{-T},T^TW_oT\}$*. An obvious problem is then to search for a choice of

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

coordinates (i.e., a similarity transformation T) that minimizes the sensitivity measure M_S. The problem of minimizing the upper bound M is much easier. It was shown in [7] that

$$\begin{aligned} M &= \operatorname{tr}(T^T W_o T)\operatorname{tr}(T^{-1}W_c T^{-T}) \\ &\quad + \operatorname{tr}(T^T W_o T) + \operatorname{tr}(T^{-1}W_c T^{-T}) \\ &\geq \left(\sum_{i=1}^{n}\sigma_i\right)^2 + 2\left(\sum_{i=1}^{n}\sigma_i\right) \end{aligned} \tag{2.13}$$

for all nonsingular T, where σ_i, $i=1,\cdots,n$ are the Hankel singular values of $H(z)$ defined by

$$\sigma_i \triangleq [\lambda_i(W_c W_o)]^{1/2} = [\lambda_i(T^{-1}W_c T^{-T} T^T W_o T)]^{1/2}. \tag{2.14}$$

Equality is obtained in (2.13) if and only if

$$T^{-1}W_c T^{-T} = \tilde{W}_c = \tilde{W}_o = T^T W_o T. \tag{2.15}$$

It was later shown by Thiele [11] that M_S also obeys

$$M_S \geq \left(\sum_{i=1}^{n}\sigma_i\right)^2 + 2\left(\sum_{i=1}^{n}\sigma_i\right)$$

and that equality is obtained if and only if (2.15) holds. In other words, realizations satisfying (2.15) minimize not only the upper bound but also the sensitivity measure M_S itself.

B. Roundoff Noise Gain

Limited word length effects on the signals cause another source of error on the output $y(k)$ of the realization (2.2) that 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])

$$G = \operatorname{tr} W_o \tag{2.16}$$

where W_o is the observability Gramian defined in (2.11a).

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

$$\min_T G = \min_T \operatorname{tr} \tilde{W}_o \tag{2.17}$$

where $\tilde{W}_o = T^T W_o T$. As such, the solution appears to depend only on the choice of C and A. In fact, (2.17) does not make much sense unless a scaling of the states is introduced.

C. Dynamic Range Constraint

In order to maintain the amplitudes of the states within an acceptable range, and hence to reduce the probability of overflow, an l_2-norm scaling is introduced as follows. First notice that with $x(0)=0$, we have

$$x(k) = \sum_{j=0}^{k-1} A^j B u(k-j-1) \tag{2.18}$$

so that the impulse response elements of the system from $u(\cdot)$ to $x(\cdot)$ are given by

$$f(j) \triangleq A^j B = [f_1(j)\cdots f_n(j)]^T, \qquad j=0,1,\cdots. \tag{2.19}$$

To ensure equal probability of overflow between all components of $x(\cdot)$, the scalar impulse response sequences $f_i(j)$, $j=0,1,\cdots$; $i=1,\cdots,n$ are equally scaled. Classically, an l_2-norm scaling is introduced:

$$\sum_{j=0}^{\infty} f_i^2(j) = 1, \qquad i=1,\cdots,n. \tag{2.20}$$

It is easy to see that this scaling is achieved by a similarity transformation T such that in the new coordinate system

$$(\tilde{W}_c)_{ii} = (T^{-1}W_c T^{-T})_{ii} = 1, \qquad i=1,\cdots,n \tag{2.21}$$

i.e., the controllability Gramian has its diagonal elements all equal to unity.

D. Relationship Between Sensitivity Measure and Roundoff Noise Gain

It follows from (2.12), (2.16), and (2.21) that, under the dynamic range constraint

$$\begin{aligned} M &= (1+n)\operatorname{tr} W_o + n \\ &= (1+n)G + n. \end{aligned} \tag{2.22}$$

Since M and G are both positive quantities, it follows from (2.22) that, under the dynamic range constraint, the sensitivity bound M and the roundoff noise gain G are simultaneously minimized. The minimum achievable gain is given by (see [1], [2])

$$G_{\min} = \frac{1}{n}\left(\sum_{i=1}^{n}\sigma_i\right)^2. \tag{2.23}$$

This 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]. Starting from an arbitrary (A,B,C), he constructs the set of similarity transformations yielding optimal realizations.

III. Finite Precision Aspects in a Closed-Loop Compensator: Problem Formulation

The results published so far on finite precision implementations (which we have summarized in Section II) 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. The results typically have a signal processing flavor, i.e., they concern the implementation of a given filter.

In this section we want to study the effects of FWL implementation and finite precision arithmetic on the

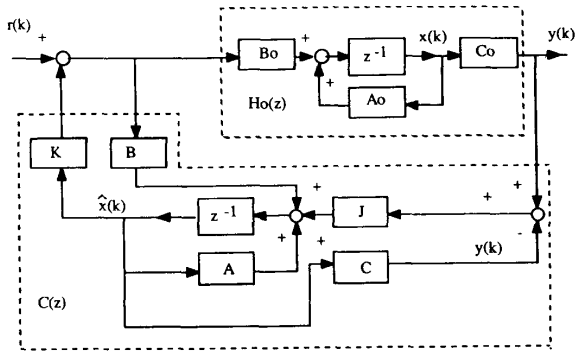


Fig. 1.

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:

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

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 implementation in the output feedback controller will be called K and J, respectively. For example, K_0 is such that $\lambda_i(A_0 - B_0K_0) =$ set of desired closed-loop poles, and K is the FWL implementation of K_0.

The state-variable observer and feedback controller contain an FWL implementation (A, B, C) of (A_0, B_0, C_0). The block diagram of the plant $H_0(z)$ and its state-estimate feedback controller $C(z)$ is given in Fig. 1.

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

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

We shall not discuss these issues here. 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 transfer function $H_c(z)$ is a function of the actual system and of the compensator $C(z)$:

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

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.

IV. Sensitivity Measure of the Closed-Loop System

We assume that a given coordinate space has been chosen for the state-space representation of $H_0(z)$. We then call (A_0, B_0, C_0) the exact (infinite precision) implementation of $H_0(z)$ in that coordinate space, and (A, B, C) the corresponding FWL implementation that will be used to construct a state estimate. We also denote by K and J the FWL implementations of the gains K_0 and J_0 that correspond to the design choices for the state feedback gain and the observer gain, respectively.

The state equations of the closed-loop system are then (see Fig. 1)

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

$$y(k) = \begin{bmatrix} C_0 & 0 \end{bmatrix} \begin{bmatrix} x(k) \\ \hat{x}(k) \end{bmatrix}. \tag{4.2}$$

We denote

$$\bar{A} \triangleq \begin{bmatrix} A_0 & -B_0K \\ JC_0 & A - BK - JC \end{bmatrix} \bar{B} \triangleq \begin{bmatrix} B_0 \\ B \end{bmatrix} \bar{C} \triangleq \begin{bmatrix} C_0 & 0 \end{bmatrix}. \tag{4.3}$$

The closed-loop transfer function is

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

We note immediately that, because of the FWL effects (i.e., $(A, B, C) \neq (A_0, B_0, C_0)$), the observer dynamics do not cancel in the closed-loop transfer function.

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 follow-

ing 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 $H_c^o(z)$ 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

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

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

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

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

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:

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

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)

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

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

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

where

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

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

and W_3 and 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 (see (2.11) for comparison). Finally, $H_K(z)$ and $H_o(z)$ are the return differences of the controller and observer, respectively.

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

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

is minimized:

$$\min_{T:\det T\neq 0} M. \quad (4.15)$$

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

V. 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:

$$M = \operatorname{tr}(T^TM_1^0T)\operatorname{tr}(T^{-1}M_2^0T^{-T}) + \operatorname{tr}(T^TM_3^0T) + \operatorname{tr}(T^{-1}M_4^0T^{-T}) \quad (5.1)$$

where

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

These are four positive definite matrices; the superscript 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:

$$T_0^T M_1^0 T_0 = \Sigma \tag{5.3a}$$

$$T_0^{-1} M_2^0 T_0^{-T} = \Sigma \tag{5.3b}$$

where

$$\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \cdots, \sigma_n),\ \sigma_i > 0. \tag{5.4}$$

This transformation to a balanced realization can be performed using a numerically well-conditioned algorithm due to Laub [10]. It is unique up to a signature matrix (see [8]). It transforms M_3^0 and M_4^0 into

$$M_3 \triangleq T_0^T M_3^0 T_0$$

$$M_4 \triangleq T_0^{-1} M_4^0 T_0^{-T}. \tag{5.5}$$

Now let

$$T = T_0 T_1 \tag{5.6}$$

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

$$T_1 = R_1 \Pi R_0^T \tag{5.7}$$

where R_1 and R_0 are orthogonal matrices, and

$$\Pi = \operatorname{diag}\left(x_1^{1/2}, x_2^{1/2}, \cdots, x_n^{1/2}\right), \qquad x_i > 0. \tag{5.8}$$

Therefore, M can be rewritten as

$$M = \operatorname{tr}\left(R_1^T \Sigma R_1 \Pi^2\right) \operatorname{tr}\left(R_1^T \Sigma R_1 \Pi^{-2}\right) + \operatorname{tr}\left(R_1^T M_3 R_1 \Pi^2\right) + \operatorname{tr}\left(R_1^T M_4 R_1 \Pi^{-2}\right) \tag{5.9}$$

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

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

$$K \triangleq R_1^T \Sigma R_1 = \{k_{ij}\}, \qquad i,j = i, \cdots, n \tag{5.11a}$$

$$Q \triangleq R_1^T M_3 R_1 = \{q_{ij}\}, \qquad i,j = i, \cdots, n \tag{5.11b}$$

$$P \triangleq R_1^T M_4 R_1 = \{p_{ij}\}, \qquad i,j = 1, \cdots, n. \tag{5.11c}$$

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

$$\sum_{i=1}^{n} k_{ii} = \operatorname{tr} K = \operatorname{tr} \Sigma \triangleq S_0 \tag{5.12a}$$

$$\sum_{i=1}^{n} q_{ii} = \operatorname{tr} Q = \operatorname{tr} M_3 \triangleq S_1 \tag{5.12b}$$

$$\sum_{i=1}^{n} p_{ii} = \operatorname{tr} P = \operatorname{tr} M_4 \triangleq S_2. \tag{5.12c}$$

It follows from (5.7)–(5.9) that the minimization of M w.r.t. T_1 reduces to the minimization of M w.r.t. R_1 (orthogonal) and $x_1, \cdots, x_n$. It further follows from (5.10) and (5.11) that the only way in which R_1 affects M is via the diagonal elements k_{ii}, p_{ii}, and q_{ii}, $i = 1, \cdots, n$ of K, Q, and P, whose sums are, respectively, constrained by (5.12). The values of S_0, S_1, S_2 are independent of R_1. The optimization problem (4.15) can therefore be reformulated as follows:

$$\min M \text{ as in (5.10) w.r.t. } \{x_i\}, \{k_{ii}\}, \{q_{ii}\}, \{p_{ii}\}, \quad i = 1, \cdots, n \text{ subject to (5.11) and (5.12).} \tag{5.13}$$

We comment immediately that

a) R_0 does not affect M. Therefore, if an optimal R_1 and Π can be found, then a set of optimizing T will result from (5.6) and (5.7) for any orthogonal R_0.
b) The $\{x_i\}$ are independent of $\{k_{ii}\}$, $\{q_{ii}\}$, and $\{p_{ii}\}$; the only constraint on $\{x_i\}$ is $x_i > 0$ for all i.
c) The only constraint on R_1 (besides its orthogonality) is through (5.11) and (5.12).

To solve (5.13), we use Lagrange's method. We define

$$L \triangleq M + \lambda_0 \left(S_0 - \sum_1^n k_{ii}\right) + \lambda_1 \left(S_1 - \sum_1^n q_{ii}\right) + \lambda_2 \left(S_2 - \sum_1^n p_{ii}\right) \tag{5.14}$$

where λ_i are constants for $i = 0, 1, 2$. Taking the derivatives of L w.r.t. the $\{x_i\}$, $\{k_{ii}\}$, $\{q_{ii}\}$, and $\{p_{ii}\}$ yields

$$-\sum_{i=1}^{n} (x_i k_{ii}) x_j^{-2} k_{jj} + k_{jj} \sum_{i=1}^{n} \left(x_i^{-1} k_{ii}\right) + q_{jj} - x_j^{-2} p_{jj} = 0 \tag{5.15a}$$

$$x_j \sum_{i=1}^{n} \left(x_i^{-1} k_{ii}\right) + \sum_{i=1}^{n} (x_i k_{ii}) x_j^{-1} - \lambda_0 = 0 \tag{5.15b}$$

$$x_j - \lambda_1 = 0 \tag{5.15c}$$

$$x_j^{-1} - \lambda_2 = 0 \tag{5.15d}$$

for $j = 1, 2, \cdots, n$, plus the additional constraints (5.12). It follows from (5.15c) and (5.15d) that

$$x_j = \lambda_1, \qquad j = 1, \cdots, n, \text{ and } \lambda_1 \lambda_2 = 1. \tag{5.16}$$

Inserting (5.16) into (5.15a) yields

$$q_{jj} - \lambda_2^2 p_{jj} = 0. \tag{5.17}$$

This, together with (5.12b) and (5.12c) and the constraints $x_j > 0$ for all j, yields

$$x_j = \lambda_1 = \frac{1}{\lambda_2} = \sqrt{\frac{S_2}{S_1}}. \tag{5.18}$$

It then follows from (5.15b) and (5.12a) that

$$\lambda_0 = 2 \sum_{i=1}^{n} k_{ii} = 2 S_0. \tag{5.19}$$

The optimum value of M is then given by

$$M_{\min} = S_0^2 + 2\sqrt{S_1 S_2}. \tag{5.20}$$

Note that it only depends on the traces of Σ, M_3, and M_4.

We can now characterize the optimal solution set, i.e., the set of similarity transformations T that minimize M in (5.1). This set is defined by

$$T = T_0 R_1 \Pi R_0^T \tag{5.21}$$

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

$$\Pi = \left(\frac{S_2}{S_1}\right)^{1/4} I \tag{5.22}$$

and R_1 is any orthogonal matrix satisfying

$$\left(R_1^T M_3 R_1\right)_{ii} = \frac{S_1}{S_2}\left(R_1^T M_4 R_1\right)_{ii}, \qquad i = 1, \cdots, n \tag{5.23}$$

where $S_1 = \operatorname{tr} M_3$ and $S_2 = \operatorname{tr} M_4$. The existence of such R_1 is proved in Appendix A.

Comment: The set of realizations (5.21) with the properties (5.22) and (5.23) minimize the upper bound M of the sensitivity (5.1), yielding the optimal value $M_{\min}$ of (5.20). Note that they do not necessarily minimize the closed-loop sensitivity M_S itself, given by (4.10). Hence the pleasing result of Thiele [11] for realizations of filters (see Section II-A) cannot be extended to the present closed-loop situation. The reason for this discrepancy is that in the case of (2.12) and (2.13), the class of realizations that minimizes the sum of the last two terms also minimizes the first term, whereas in (4.11) the last three terms contain quantities that are not included in the first term. For this reason, it will be useful to check whether realizations that optimize the sensitivity upper bound are also close to optimal w.r.t. the actual sensitivity. This will be verified by simulation in Section VIII.

VI. Roundoff Noise Gain of the Closed-Loop System

Recall that the closed-loop system is described by (4.1) and (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) and (4.2) is then replaced by

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

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

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

$$E(k) \triangleq \begin{bmatrix} x(k) - x^*(k) \\ \hat{x}(k) - \hat{x}^*(k) \end{bmatrix},$$

$$e(k) \triangleq \begin{bmatrix} 0 \\ \hat{x}^*(k) - Q[\hat{x}^*(k)] \end{bmatrix} \tag{6.2a}$$

$$\Delta y(k) \triangleq y(k) - y^*(k). \tag{6.2b}$$

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

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

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

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 $\sigma^2 I$, with

$$\sigma^2 = \frac{1}{12} 2^{-2b_0}. \tag{6.4}$$

It then follows from (6.3) that the steady state noise output of the closed-loop system is

$$\begin{aligned} G\sigma^2 &\triangleq \lim_{k\to\infty} E\left[\Delta^2 y(k)\right] \\ &= \bar{C} \sum_{k=0}^{\infty} \bar{A}^k R (\bar{A}^T)^k \bar{C}^T \\ &= \operatorname{tr}\left[\bar{W}_o R\right] \end{aligned} \tag{6.5}$$

where G is defined as the roundoff noise gain,

$$R \triangleq E\left[e(k)e^T(k)\right] = \begin{bmatrix} 0 & 0 \\ 0 & \sigma^2 I \end{bmatrix} \tag{6.6}$$

$$\bar{W}_o \triangleq \sum_{k=0}^{\infty} (\bar{A}^T)^k \bar{C}^T \bar{C} \bar{A}^k. \tag{6.7}$$

$\bar{W}_o$ is the observability Gramian of the closed-loop system. The roundoff noise gain of the closed-loop system is

$$G = \frac{1}{\sigma^2} \operatorname{tr}\left[\bar{W}_o R\right]. \tag{6.8}$$

We now compute a more usable expression for G. First notice that

$$\bar{W}_o R = \frac{1}{2\pi j} \oint_{|z|=1} \bar{G}(z) \bar{G}^T(z^{-1}) z^{-1} dz \cdot R \tag{6.9}$$

where

$$\bar{G}^T(z) = \bar{C}(zI - \bar{A})^{-1} \tag{6.10a}$$

$$= \left[g_1^T(z)\, g_2^T(z)\right]. \tag{6.10b}$$

Here $g_1(z)$ and $g_2(z)$ are, respectively, the first n and the last n components of the vector $\bar{G}(z)$. It follows from (5.6), (5.8), and (5.9) that

$$G = \operatorname{tr}\left[\frac{1}{2\pi j} \oint_{|z|=1} g_2(z) g_2^T(z^{-1}) z^{-1} dz\right]. \tag{6.11}$$

In the computation of G, we shall replace the FWL quantities A, B, C, K, J in $\bar{A}$ by their infinite precision equivalents A_0, B_0, C_0, K_0, J_0. We denote by $\bar{A}_0$ the corresponding matrix $\bar{A}$. To compute the inverse of $(zI - \bar{A}_0)$ we use the following trick:

$$(zI - \bar{A}_0) = TMT^{-1} \tag{6.12}$$

where

$$M = \begin{bmatrix} zI - A_0 + B_0K_0 & B_0K_0 \\ 0 & zI - A_0 + JC_0 \end{bmatrix} = \begin{bmatrix} M_1 & M_3 \\ 0 & M_2 \end{bmatrix} \tag{6.13a}$$

$$T = \begin{bmatrix} I & 0 \\ I & I \end{bmatrix}. \tag{6.13b}$$

Therefore,

$$(zI-\bar{A}_0)^{-1}=TM^{-1}T^{-1}=T\begin{bmatrix} M_1^{-1} & -M_1^{-1}M_3M_2^{-1} \\ 0 & M_2^{-1} \end{bmatrix}T^{-1}. \tag{6.14}$$

For the computation of $g_2(z)$, we only need the (1,2) element of $(zI-\bar{A}_0)^{-1}$. It easily follows from (6.10), (6.13), and (6.14) that

$$\begin{aligned} g_2^T(z) &= -C_0(zI-A_0+B_0K_0)^{-1} \\ &\quad \cdot B_0K_0(zI-A_0+J_0C_0)^{-1} \\ &= -H_c^o(z)G_o^T(z) \end{aligned} \tag{6.15}$$

with $H_c^o(z)$ defined by (4.6) and $G_o(z)$ by (4.7a). Finally, the roundoff noise gain of the closed-loop system is given by

$$\begin{aligned} G &\cong \mathrm{tr}\left[\frac{1}{2\pi j}\oint_{|z|=1} H_c^o(z)H_c^o(z^{-1})G_o(z)G_o^T(z^{-1})z^{-1}\,dz\right] \\ &= \|H_c^o(z)G_o(z)\|_2^2 \end{aligned} \tag{6.16}$$

where the approximate 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 $H_c^o(z)$.

VII. 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 (see (6.11)) 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 an l_2-scaling on the observer states $\hat{x}$ corresponds with finding a coordinate basis for $\bar{A},\bar{B},\bar{C}$ in which

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

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 $\bar{A}_0$, and using the same trick as before for the computation of the inverse of $(zI-\bar{A}_0)$, yields (see (4.7b))

$$f_2(z)=(zI-A_0+B_0K_0)^{-1}B_0=F_K(z). \tag{7.4}$$

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:

$$\min_{T:\det T\neq 0} \mathrm{tr}(T^TWT) \tag{7.6a}$$

subject to

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

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:

$$W_{cc}^0=T_0T_0^T. \tag{7.8}$$

Notice that T_0 is not unique. We apply the transformation T_0 to (A^0,B^0,C^0,K^0,J^0) to produce (A^1,B^1,C^1,K^1,J^1); the Gramians W_{cc}^0 and W^0 are transformed into

$$W_{cc}^1=T_0^{-1}W_{cc}^0T_0^{-T}=I \tag{7.9a}$$

$$W^1=T_0^TW^0T_0. \tag{7.9b}$$

We now denote by σ_i the singular values of the product $W_{cc}^0W^0$, and we note that these are invariant under similarity transformations:

$$\sigma_i=\sqrt{\lambda_i(W_{cc}^0W^0)}>0, \qquad i=1,\cdots,n. \tag{7.10}$$

The set of optimal realizations is then obtained from (A^1,B^1,C^1,K^1,J^1) by any similarity transformation T_1 constructed as follows:

$$T_1=R_1\Pi R_0^T \tag{7.11}$$

where

$$\Pi=\mathrm{diag}(\Pi_i),\ \Pi_i=\left(\frac{\sum_{m=1}^{n}\sigma_m}{n\sigma_i}\right)^{1/2}, \qquad i=1,\cdots,n \tag{7.12a}$$

and where R_0 and R_1 are orthogonal matrices ($R_0R_0^T = R_0^TR_0 = I$) satisfying

$$\left(R_0\Pi^{-2}R_0^T\right)_{ii} = 1, \qquad i = 1,\cdots,n \tag{7.12b}$$

$$R_1^TW^1R_1 = \Sigma^2 = \operatorname{diag}\left(\sigma_i^2\right), \qquad i = 1,\cdots,n. \tag{7.12c}$$

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

$$T_{\text{opt}} = T_0T_1 = T_0R_1\Pi R_0^T \tag{7.13}$$

where T_0 is defined by (7.8) and R_1, Π, and R_0 by (7.12).

We now discuss the construction of T_{opt} in more detail in order to investigate the possible degrees of freedom in the choices of T_0, R_1, Π, or R_0, which might be used to achieve other desirable goals, such as minimizing the bound M on the sensitivity of the closed-loop system.

First notice that the transformation T_0R_1 brings the initial Gramians W^0 and W_{cc}^0 into input balanced form, i.e.,

$$(T_0R_1)^{-1}W_{cc}^0(T_0R_1)^{-T} = I \tag{7.14a}$$

$$(T_0R_1)^TW^0(T_0R_1) = \Sigma^2. \tag{7.14b}$$

If the σ_i are arranged in nonincreasing order (and we shall henceforth assume this), then this form is unique (see, e.g., [8]). Since Π is also unique, this means that the only degree of freedom in T_{opt} is R_0 (see (7.13)). The only constraints on R_0 are

$$R_0R_0^T = R_0^TR_0 = I \tag{7.15a}$$

$$\left(R_0\Pi^{-2}R_0^T\right)_{ii} = 1 \text{ with } \Pi = \operatorname{diag}(\Pi_i),\ \sum_1^n \Pi_i^{-2} = n. \tag{7.15b}$$

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 (7.13), the upper bound M on the sensitivity is independent of R_0; see (4.14). It follows that even though the transformations T_{opt} that optimize the roundoff noise gain of the closed-loop system are not unique, they all yield the same value for M, the upper bound on the sensitivity of the closed-loop system.

VIII. 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 [5]–[6] 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 formulas (4.12) and (6.16), 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

$$A_c = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 \end{bmatrix}$$

$$B_c = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$C_c = [c_1 \quad c_2 \quad c_3].$$

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

$$A_\delta = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ -d_0 & -d_1 & 1-d_2 \end{bmatrix} \tag{8.1a}$$

$$B_\delta = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \tag{8.2b}$$

$$C_\delta = [e_1 \quad e_2 \quad e_3] \tag{8.2c}$$

with $d_0 = 0.0004$, $d_1 = 0.0178$, $d_2 = 0.2418$, $e_1 = 0.0087$, $e_2 = 0.0087$, and $e_3 = 0.0022$ for the case $T_S = 1$. The realizations that minimize the closed-loop sensitivity upper bound are typically fully parametrized. One of these optimal realizations is given by

$$A_{\text{opt}} = \begin{bmatrix} 0.9318 & -0.0543 & -0.1010 \\ 0.0197 & 1.0170 & 0.1274 \\ 0.1486 & 0.0542 & 0.8094 \end{bmatrix}$$

$$B_{\text{opt}} = \begin{bmatrix} -0.5102 \\ -0.1290 \\ 0.0459 \end{bmatrix}$$

$$C_{\text{opt}} = [0.1285 \quad -0.5252 \quad -0.0006].$$

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 V. 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):

$$M_{\text{opt}} = 12.4557 \qquad G = 2.4265$$

$$M_c = 1.444 \times 10^4 \qquad G_c = 3.3268$$

$$M_\delta = 1.9268 \times 10^3 \qquad G_\delta = 0.6616.$$

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 VII. 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 transformation. Their canonical structure is thereby not preserved, except for the zeros, 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):

$$G_{\text{opt}}^{(s)} = 0.3811 \qquad M = 37.3962$$

$$G_c^{(s)} = 1.5006 \times 10^3 \qquad M_c^{(s)} = 1.1914 \times 10^4$$

$$G_\delta^{(s)} = 1.0305 \qquad M_\delta^{(s)} = 26.4696.$$

Comments:

1) 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.

2) 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.

3) The comparison with the δ-form has been added for curiosity, but it is somewhat unfair because it is a nonoptimal δ-form. The actual δ-form is the form (8.1) and all its similarity transformations. Rather than implementing it in a control canonical (i.e., companion) implementation, one could optimize the sensitivity and/or roundoff noise of the δ-form over the equivalence class of all representations that are similar to (8.1). This requires modifying the formulas for the sensitivity and roundoff noise gain and has been studied in [12].

The numerical values of the sensitivities given above are those of the upper bounds given by formula (4.11). In order to validate the quality of this upper bound, the following computation has been performed on all three realizations. The fractional part of the coefficients of each realization has been truncated to p bits, with p ranging from 1 to 35. This corresponds with corrupting each coefficient with an error of the same range. For a given p, the degradation between the ideal frequency response $H_{id}(\omega)$ and the FWL frequency response $H_p(\omega)$ has been measured by the following quantity:

$$R_p = \log_{10}\left[\max_\omega |H_{id}(\omega) - H_p(\omega)|\right].$$

Here $H_p(\omega)$ denotes the FWL frequency response computed from a realization in which the fractional part of each coefficient is truncated to p bits. The results of these computations are shown in Fig. 2. They confirm

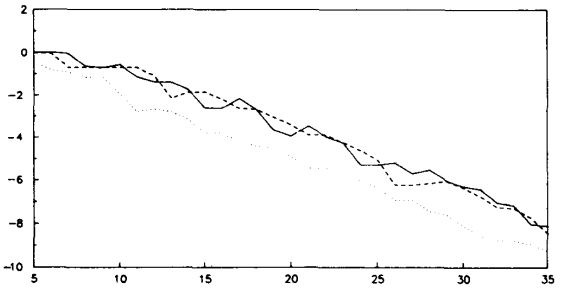


Fig. 2. Sensitivity comparison of three realizations. The dashed line indicates the particular δ-form, the solid line the companion form, and the dotted line the optimal realization.

that the realization that minimizes the upper bound of the sensitivity does in fact yield an actual sensitivity that is better than those of the other two realizations.

IX. Conclusions

The contribution of this paper has been twofold. The first was to derive computable 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 compensator implemented in FWL with the computations also being performed in finite arithmetic. The expressions clearly show the dependence of both the sensitivity and the roundoff noise on the coordinate basis of the state variable realization. The second contribution was to compute the set of optimal realizations, i.e., the set of realizations that optimize either the sensitivity 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 our numerical example, the optimal realization shows 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, suggested by Goodwin [9], is to compute the optimal realizations for both sensitivity and roundoff noise, among the set of all δ-operator state variable representations. A comparison between optimal δ-operator realizations and optimal shift operator realizations in the simpler case of filter realization (rather than closed-loop controller realization) has been performed in [12].

In line with the results of [1]–[4] for open-loop transfer functions, our sensitivity computations have all been performed using absolute sensitivity measures, which are relevant for the case of fixed point arithmetic without scaling of the parameters. In practice the calculations are most often performed using either floating point arithmetic or fixed point with scaling. We note that our results

concerning the minimization of the closed-loop roundoff noise gain do include an l_2-scaling. In addition, it is well known [13], [14] that optimal structures that minimize output roundoff noise due to FWL effects on fixed point implemented signals (with the filter coefficients in infinite precision) also tend to yield low sensitivity. Our numerical example confirms this. Therefore, we believe that our results on l_2-scaled optimal roundoff noise gain realizations also have relevance in terms of obtaining low sensitivities for fixed point arithmetic with scaling. Clearly, as pointed out by a reviewer, some form of fixed scaling should be introduced to take full advantage of the available word length. An extension of our results to the case of relative sensitivities (pertinent to fixed point arithmetic with scaling or floating point arithmetic) for closed-loop systems is beyond the scope of this paper. However, for open-loop systems a first step has been obtained in [12] with the computation of a reasonable upper bound of the relative sensitivity of a state-variable realization of a filter.

Appendix A

Lemma A: Given any two positive definite matrices A and B, both of dimension $n \times n$, there exists an orthogonal matrix U such that

$$\rho^2 (U^T A U)_{ii} = (U^T B U)_{ii}, \qquad i = 1, \cdots, n \tag{A.1}$$

where

$$\rho^2 = \frac{\operatorname{tr} B}{\operatorname{tr} A}. \tag{A.2}$$

Proof: Without loss of generality we can assume that A is diagonal: $A = \operatorname{diag}(a_1, \cdots, a_n)$, with $a_i > 0$. If this were not the case, we could write $U = U_0 U_1$, where U_0 is an orthogonal matrix such that $U_0^T A U_0$ is diagonal. We now construct U by a series of Givens transformations. We denote by $U(i,j)$ an $n \times n$ Givens matrix defined as follows for $i < j$:

$$U(i,j) = \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & c & & & s & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & -s & & & & c & \\ & & & & & & & & 1 \end{bmatrix} \tag{A.3}$$

with $c \triangleq \cos\theta$, $s \triangleq \sin\theta$, $\theta \in [0, \pi]$, and all other elements other than those indicated equal to zero.

We first study the effect of a basic Givens transformation on a 2×2 matrix; the generalization to $n \times n$ matrices will be straightforward. Denote

$$A = \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix}$$
$$B = \begin{bmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{bmatrix}$$
$$U = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \tag{A.4}$$

With $c = \cos\theta$ and $s = \sin\theta$, we now show that, for arbitrary $a_1 > 0$, $a_2 > 0$, b_{11}, b_{12}, and b_{22} such that B is positive definite, and $\rho^2 > 0$ as defined in (A.2), we can find θ such that $\rho^2 (U^T A U)_{11} = (U^T B U)_{11}$, i.e.,

$$b_{11} c^2 - 2 b_{12} cs + b_{22} s^2 = \rho^2 (a_1 c^2 + a_2 s^2). \tag{A.5}$$

We denote

$$\epsilon_i \triangleq b_{ii} - \rho^2 a_i, \qquad i = 1,2 \tag{A.6}$$

and assume that $\epsilon_1 \epsilon_2 \leq 0$. In the case of 2×2 matrices A and B, this assumption is satisfied since $\epsilon_1 + \epsilon_2 = 0$; in the $n \times n$ case, it will also be satisfied as we shall see later. Using $s^2 = 1 - c^2$, it follows that (A.5) is equivalent with $(\epsilon_1 - \epsilon_2) c^2 - 2 b_{12} cs + \epsilon_2 = 0$, or, equivalently,

$$\frac{1}{2}(\epsilon_1 - \epsilon_2) \cos 2\theta - b_{12} \sin 2\theta + \frac{1}{2}(\epsilon_1 + \epsilon_2) = 0. \tag{A.7}$$

Now define θ_0 through

$$\cos\theta_0 = \frac{1}{2}(\epsilon_1 - \epsilon_2) \Big/ \sqrt{\frac{1}{4}(\epsilon_1 - \epsilon_2)^2 + b_{12}^2}. \tag{A.8}$$

It follows that

$$\sin\theta_0 = b_{12} \Big/ \sqrt{\frac{1}{4}(\epsilon_1 - \epsilon_2)^2 + b_{12}^2}. \tag{A.9}$$

(A.7) can then be rewritten as

$$\cos(2\theta + \theta_0) = -\frac{\frac{1}{2}(\epsilon_1 + \epsilon_2)}{\sqrt{\frac{1}{4}(\epsilon_1 - \epsilon_2)^2 + b_{12}^2}}. \tag{A.10}$$

A necessary and sufficient condition for (A.10) to have a solution θ^* for θ is that

$$\frac{1}{4}(\epsilon_1 - \epsilon_2)^2 + b_{12}^2 \geq \frac{1}{4}(\epsilon_1 + \epsilon_2)^2.$$

Equivalently,

$$b_{12}^2 \geq \epsilon_1 \epsilon_2 = (b_{11} - \rho^2 a_1)(b_{22} - \rho^2 a_2). \tag{A.11}$$

This is always satisfied by our assumption that $\epsilon_1 \epsilon_2 \leq 0$.

We have shown that with one Givens transformation operating on 2×2 matrices A and B, we can satisfy one relation (A.1). We now show that for $n \times n$ matrices A and B satisfying the assumptions of Lemma A, we can satisfy (A.1) with U obtained as a product of Givens matrices. Denote

$$\epsilon_i \triangleq b_{ii} - \rho^2 a_i, \qquad i = 1, \cdots, n \tag{A.12}$$

where $\rho^2 = trB/trA$. If $\epsilon_i = 0$ for $i = 1, \cdots, n$, then $U = I$ in (A.1) and the Lemma is proved. Therefore, assume

that $\epsilon_i \neq 0$ for some i. Since $\sum_1^n \epsilon_k = 0$, it follows that there exists j such that $\epsilon_i \epsilon_j < 0$. Assuming $i < j$, take $U(i,j)$ as in (A.3). Then with proper choice of c (and hence s) as just discussed, we can obtain

$$\left(U^T(i,j)BU(i,j)\right)_{ii} = \rho^2\left(U^T(i,j)AU(i,j)\right)_{ii}. \quad \text{(A.13)}$$

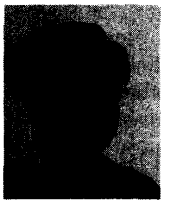
In addition, $U(i,j)$ only changes the ith and jth diagonal elements of A and B. Therefore, U in (A.1) can be obtained as the product of at most $n-1$ Givens transformations.

Acknowledgment

The authors would like to thank Graham Goodwin, Jean Meinguet, Paul Van Dooren, and Vincent Wertz, as well as the reviewers, for helpful discussions and suggestions during the preparation of this manuscript.

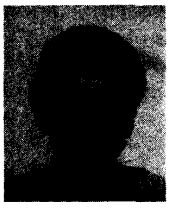
References

[1] C. T. Mullis and R. A. Roberts, "Filter structures which minimize roundoff noise in fixed-point digital filters," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 505–508, 1976.
[2] S. Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 273–281, Aug. 1977.
[3] 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, pp. 884–888, Oct. 1984.
[4] W. J. Lutz and S. Louis Hakimi, "Design of multi-input multi-output systems with minimum sensitivity," *IEEE Trans. Circuits Syst.*, vol. 35, pp. 1114–1122, Sept. 1988.
[5] R. H. Middleton and G. C. Goodwin, *Digital Control and Estimation: A Unified Approach*. Englewood Cliffs, NJ: Prentice Hall, 1990.
[6] R. H. Middleton and G. C. Goodwin, "Improved finite word length characteristics in digital control using delta operators," *IEEE Trans. on Automat. Contr.*, vol. 31, pp. 1015–1021, Nov. 1986.
[7] L. Thiele, "Design of sensitivity and roundoff noise optimal state-space discrete systems," *Int. J. Circuit Theory Appl.*, vol. 12, pp. 39–46, 1984.
[8] Darrell Williamson, "Roundoff noise minimization and pole-zero sensitivity in fixed-point digital filters using residue feedback," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1210–1220, Oct. 1986.
[9] G. C. Goodwin, personal communication, 1989.
[10] A. J. Laub, "Computations of balancing transformations," in *Proc. 1980 JACC*, vol. 1, San Francisco, CA, 1980.
[11] L. Thiele, "On the sensitivity of linear state-space systems," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 502–510, May 1986.
[12] G. Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parametrizations," to be published.
[13] L. B. Jackson, "Roundoff noise bounds derived from coefficient sensitivities for digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-23, 481–485, Aug. 1976.
[14] M. Kawamata and T. Higuchi, "A unified approach to the optimal synthesis of fixed-point state-space digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 911–920, 1985.



Gang Li received the B.S. degree in electrical engineering from Beijing Institute of Technology, Beijing, China, in 1982, and the M.S. degree in telecommunications from Louvain University, Belgium, in 1988. He is currently working towards the Ph.D. degree at Louvain University.

His research interests include digital system design, optimal control and filtering, and numerical problems in estimation and control theory.



Michel Gevers (S'66–M'72–SM'86–F'90) received the electrical engineering degree from Louvain University, Belgium, in 1968, and the Ph.D. degree from Stanford University, California, in 1972.

He is now full professor and head of the Control Group at Louvain University. His research interests are in system identification, adaptive estimation and control, multivariable systems, optimal control and filtering, and nonlinear systems. He is Associate Editor of *Mathematics of Control, Signals, and Systems*, and a co-author, with R. R. Bitmead and V. Wertz, of *Adaptive Optimal Control—The Thinking Man's GPC*, Prentice Hall, 1990.