2-D State-Space Digital Filters with Fewer Multipliers

WU-SHENG LU, MEMBER, IEEE, AND ANDREAS ANTONIOU, FELLOW, IEEE

Abstract —It is shown that as many as $[m(m-1)+n(n-1)]/2$ multiplications can be eliminated in a local state-space realization of a 2-D digital filter of the order (m, n) by applying an appropriate transformation from the class of orthogonal similarity transformations. Further, it is demonstrated that the class of similarity transformations can be enlarged so that one can either introduce $m+n$ free parameters in the sensitivity function, which may be chosen to reduce the sensitivity of the filter or to eliminate $m+n$ additional multiplications while keeping the filter free of overflow oscillations. A numerical example is then given to illustrate the various techniques.

I. Introduction

IT HAS been known for some time that minimum-norm state-space realizations of 1-D digital filters have certain desirable properties when the effects of finite wordlength are taken into consideration [1]–[3], e.g., low roundoff noise and freedom from overflow oscillations. Owing to these advantages and the continued interest in multidimensional digital-signal processing, the 2-D minimum-norm state-space realization has received particular attention and a number of contributions have been published on this subject [4]–[6].

Lodge and Fahmy [6] have shown that if the system matrix of a 2-D state-space digital filter satisfies the 2-D Lyapunov equation, then the Euclidean norm of the system matrix of a minimum-norm realization is strictly less than one and, therefore, such a realization is free of overflow oscillations. In such a case, entries in the state-space representation are highly unlikely to be either zero or one and, consequently, $(m+n)(m+n+2)+1$ multipliers are almost always needed in the implementation of a filter of the order (m, n).

More economical realizations requiring only $2(mn+m+n)+1$ multiplications can be achieved by using the method reported by Kung *et al.* in [13]. However, the norm of the system matrix is always greater than one and the advantages of freedom from overflow oscillations and low roundoff noise do not apply in general.

Recently, Aboulnasr and Fahmy [8] have suggested using an orthogonal similarity transformation in order to reduce the number of the multiplications while preserving the norm of the system matrix. Specifically, through the use of the singular value decomposition (SVD) technique, they proved [8] that a suitably chosen orthogonal transformation can introduce $r_1 = n(m-1)$ zero entries (when $m \geqslant n$) in the system matrix and hence r_1 multiplications can be eliminated.

In this paper, we show that as many as $r_2 = [m(m-1)+n(n-1)]/2$ multiplications can be eliminated in a minimum-norm realization by applying an appropriate transformation from the class of orthogonal similarity transformations where $r_2 \geqslant r_1$ and $r_2 \gg r_1$ if $|m-n|$ is large. Further, through the use of a broader class of similarity transformations it is demonstrated that one can either introduce $m+n$ free parameters in the sensitivity function of the digital filter, which can be adjusted to reduce the sensitivity or to eliminate $m+n$ additional multiplications. These improvements are brought about without changing the norm of the system matrix and, therefore, improved realizations are achieved that are free of overflow oscillations. This paper concludes with a numerical example which illustrates our approach.

II. Reduction in the Number of Multiplications

The 2-D digital filter considered in this paper is represented by Roesser's local state-space model [7]

$$\begin{bmatrix} x^v(i+1, j) \\ x^h(i, j+1) \end{bmatrix} = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix} \begin{bmatrix} x^v(i, j) \\ x^h(i, j) \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} u(i, j) \equiv Ax + bu \tag{1a}$$

$$y(i, j) = [c_1 \quad c_2] \begin{bmatrix} x^v(i, j) \\ x^h(i, j) \end{bmatrix} + du(i, j) \equiv cx + du \tag{1b}$$

where $x^v \in R^m$, $x^h \in R^n$, and (m, n) will be referred to as the order of the filter. Notice that if (1) is a minimal realization of the transfer function, then (m, n) is also the order of the transfer function [13]. It is assumed in the rest of this paper that the realization represented by (1a) and (1b) is a minimum-norm realization of a stable 2-D quarter-plane digital filter, i.e., $\|A\| < 1$ where $\|A\|$ denotes the Euclidean norm of A defined as the square root of the maximal eigenvalue of A^tA (A^t denotes the transpose of A). Our goal is to seek an appropriate similarity transformation $Q = Q_1 \oplus Q_2$ ($\oplus$ means direct sum) such that the resulting realization $(QAQ^{-1}, Qb, cQ^{-1}, d)$ has a maximum number of zero entries while preserving the norm of the system matrix.

Manuscript received April 24, 1985; revised January 10, 1986. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

The authors are with the Department of Electrical Engineering, University of Victoria, Victoria, B.C., Canada, V8W 2Y2.

IEEE Log Number 8609297.

In this section we restrict our attention to the class of orthogonal transformations, in which case $\|QAQ^{-1}\| = \|A\|$.

Let

$$\bar{A} = QAQ^t = \begin{bmatrix} \bar{A}_1 & \bar{A}_2 \\ \bar{A}_3 & \bar{A}_4 \end{bmatrix},$$

$$\bar{b} = Qb = \begin{bmatrix} \bar{b}_1 \\ \bar{b}_2 \end{bmatrix} \quad \text{and} \quad \bar{c} = cQ^t = [\bar{c}_1 \quad \bar{c}_2]$$

where

$$\bar{A}_1 = Q_1 A_1 Q_1^t, \quad \bar{b}_1 = Q_1 b_1, \quad \bar{c}_1 = c_1 Q_1^t$$

and

$$\bar{A}_4 = Q_2 A_4 Q_2^t, \quad \bar{b}_2 = Q_2 b_2, \quad \bar{c}_2 = c_2 Q_2^t.$$

Viewing $S_1 \equiv (A_1, b_1, c_1)$ as the representation of a 1-D single-input subsystem, S_1 is said to be reachable if, and only if, its reachability matrix

$$F_1 = [b_1 \quad A_1 b_1 \cdots A_1^{m-1} b_1]$$

is of full rank, i.e.,

$$\det F_1 \neq 0. \tag{2}$$

Furthermore, subsystem S_1 is said to be observable if, and only if, its observability matrix

$$Y_1 = \begin{bmatrix} c_1 \\ c_1 A_1 \\ \vdots \\ c_1 A_1^{m-1} \end{bmatrix}$$

is of full rank, namely

$$\det Y_1 \neq 0. \tag{3}$$

Let

$$F_2 = [b_2 \quad A_4 b_2 \cdots A_4^{n-1} b_2], \quad \text{and} \quad Y_2 = \begin{bmatrix} c_2 \\ c_2 A_4 \\ \vdots \\ c_2 A_4^{n-1} \end{bmatrix}$$

be the reachability and the observability matrices of subsystem $S_2 \equiv (A_4, b_2, c_2)$. The reachability and the observability of S_2 can be characterized by the nonsingularities of F_2 and Y_2. It is worthwhile to observe that in a digital-filter context the resulting subsystems S_1 and S_2 rarely fail the reachability test (2) and the observability test (3) simultaneously. We, therefore, assume that both subsystems S_1 and S_2 are reachable. The case where S_1 or S_2 is neither reachable nor observable will be dealt with subsequently.

Since subsystem S_1 is assumed to be reachable, F_1 is nonsingular and, therefore, its QR decomposition (QRD) gives (see Theorem A.2 of the Appendix)

$$Q_1 F_1 = R_1 \tag{4}$$

where Q_1 is an $m \times n$ orthogonal matrix and R_1 is an upper triangular matrix, i.e.,

$$R_1 = \begin{bmatrix} r_{11} & & * \\ & r_{12} & \\ & & \ddots & \\ 0 & & & r_{1m} \end{bmatrix} \tag{5}$$

with $r_{1i} \neq 0$, $1 \leqslant i \leqslant m$. Denoting

$$\bar{A}_1 = Q_1 A_1 Q_1^t, \quad \bar{b}_1 = Q_1 b, \quad \text{and} \quad \bar{c}_1 = cQ_1^t$$

equations (4) and (5) imply that

$$\bar{F}_1 \equiv [\bar{b}_1 \quad \bar{A}_1 \bar{b}_1 \cdots \bar{A}_1^{m-1} \bar{b}_1] = \begin{bmatrix} r_{11} & & * \\ & r_{12} & \\ & & \ddots & \\ 0 & & & r_{1m} \end{bmatrix} \tag{6}$$

which immediately gives

$$\bar{b}_1 = \begin{bmatrix} r_{11} \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{7}$$

Matrix $\bar{A}_1$ can be computed by using the Cayley–Hamilton theorem. We can write

$$\begin{aligned} \bar{A}_1 \bar{F}_1 &= \bar{A}_1 [\bar{b}_1 \quad \bar{A}_1 \bar{b}_1 \cdots \bar{A}_1^{m-1} \bar{b}_1] \\ &= [\bar{A}_1 \bar{b}_1 \quad \bar{A}_1^2 \bar{b}_1 \cdots \bar{A}_1^m \bar{b}_1] \\ &= [\bar{b}_1 \quad \bar{A}_1 \bar{b}_1 \cdots \bar{A}_1^{m-1} \bar{b}_1] \\ &\quad \cdot \begin{bmatrix} 0 & 0 & & 0 & -a_1 \\ 1 & 0 & & 0 & -a_2 \\ 0 & 1 & & 0 & -a_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & & 1 & -a_m \end{bmatrix} \end{aligned} \tag{8}$$

where elements a_i can be determined by calculating

$$\det(\lambda I - A_1) = \lambda^m + a_m \lambda^{m-1} + \cdots + a_1.$$

Since both $\bar{F}_1$ and $\bar{F}_1^{-1}$ are upper triangular, we have

$$\begin{aligned} \bar{A}_1 &= \bar{F}_1 \begin{bmatrix} 0 & 0 & & 0 & -a_1 \\ 1 & 0 & & 0 & -a_2 \\ 0 & 1 & & 0 & -a_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & & 1 & -a_m \end{bmatrix} \bar{F}_1^{-1} \\ &= \begin{bmatrix} * & * & \cdots & * & * \\ \frac{r_{12}}{r_{11}} & * & \cdots & * & * \\ & \frac{r_{13}}{r_{12}} & & \vdots & \vdots \\ & & \ddots & & \\ 0 & & & \frac{r_{1m}}{r_{1,m-1}} & * \end{bmatrix}. \end{aligned} \tag{9}$$

Similarly, since subsystem S_2 is also assumed to be reachable, matrix F_2 is nonsingular. Hence there exists an $n \times n$ orthogonal matrix Q_2 such that

$$\bar{b}_2 = Q_2 b_2 = \begin{bmatrix} r_{21} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{10}$$

$$\bar{A}_4 = Q_2 A_4 Q_2^t = \begin{bmatrix} * & * & \cdots & * & * \\ \frac{r_{22}}{r_{21}} & * & \cdots & * & * \\ & \frac{r_{23}}{r_{22}} & & \vdots & \vdots \\ & & \ddots & & \\ 0 & & & \frac{r_{2n}}{r_{2,n-1}} & * \end{bmatrix} \tag{11}$$

where $r_{2j} \neq 0$ $(1 \leqslant j \leqslant n)$ are given by QRD of matrix F_2.

If S_2 is observable but not reachable, one can apply QRD to the transpose of the observability matrix Y_2 to obtain the desired orthogonal transformation matrix as

$$\bar{Q}_2 Y_2^t = \bar{R}_2$$

i.e.,

$$Y_2 \bar{Q}_2^t = \bar{R}_2^t \tag{12}$$

where $\bar{Q}_2$ is orthogonal and $\bar{R}_2^t$ is lower triangular, i.e.,

$$\bar{R}_2^t = \begin{bmatrix} \bar{r}_{21} & & \\ & \bar{r}_{22} & 0 \\ & * & \ddots & \\ & & & \bar{r}_{2n} \end{bmatrix} \tag{13}$$

where $\bar{r}_{2j} \neq 0$, $1 \leqslant j \leqslant n$. Now let

$$\bar{A}_4 = \bar{Q}_2 \bar{A}_4 \bar{Q}_2^t, \quad \bar{b}_2 = \bar{Q}_2 b_2, \quad \text{and} \quad \bar{c}_2 = c_2 \bar{Q}_2^t.$$

From (12) and (13), we obtain

$$\bar{Y}_2 = \begin{bmatrix} \bar{c}_2 \\ \bar{c}_2 \bar{A}_4 \\ \vdots \\ \bar{c}_2 \bar{A}_4^{n-1} \end{bmatrix} = \begin{bmatrix} \bar{r}_{21} & & \\ & \bar{r}_{22} & 0 \\ & * & \ddots & \\ & & & \bar{r}_{2n} \end{bmatrix} \tag{14}$$

which gives

$$\bar{c}_2 = [\bar{r}_{21} \quad 0 \cdots 0]. \tag{15}$$

As in (8)

$$\bar{Y}_2 \bar{A}_4 = \begin{bmatrix} \bar{c}_2 \\ \bar{c}_2 \bar{A}_4 \\ \vdots \\ \bar{c}_2 \bar{A}_4^{n-1} \end{bmatrix} \bar{A}_4 = \begin{bmatrix} \bar{c}_2 \bar{A}_4 \\ \bar{c}_2 \bar{A}_4^2 \\ \vdots \\ \bar{c}_2 \bar{A}_4^n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & \cdots & 1 \\ -b_1 & -b_2 & -b_3 & \cdots & -b_n \end{bmatrix} \bar{Y}_2 \tag{16}$$

where elements b_i are given by

$$\det(\lambda I - A_4) = \lambda^n + b_n \lambda^{n-1} + \cdots + b_1.$$

Equations (16) and (14) give

$$\bar{A}_4 = \bar{Y}_2^{-1} \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & \cdots & 1 \\ -b_1 & \cdot & \cdot & \cdot & -b_n \end{bmatrix} \bar{Y}_2 = \begin{bmatrix} * & \frac{\bar{r}_{22}}{\bar{r}_{21}} & & & \\ * & * & \frac{\bar{r}_{23}}{\bar{r}_{22}} & & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ * & * & * & \cdots & \frac{\bar{r}_{2r}}{\bar{r}_{2,n-1}} \\ * & * & * & \cdots & * \end{bmatrix}. \tag{17}$$

We now consider the case where S_1 is neither reachable nor observable. In this case

$$\operatorname{rank} F_1 = m_1 < m.$$

Because of the special structure of F_1, we have

$$\operatorname{rank} \tilde{F}_1 = m_1$$

where

$$\tilde{F}_1 \equiv [b_1 \quad A_1 b_1 \cdots A_1^{m_1 - 1} b_1].$$

Applying QRD to $\tilde{F}_1$ yields

$$Q_{11} \tilde{F}_1 = \begin{bmatrix} \tilde{R}_1 \\ 0 \end{bmatrix} \tag{18}$$

with Q_{11} orthogonal and

$$\tilde{R}_1 = \begin{bmatrix} \tilde{r}_{11} & & * \\ & \ddots & \\ 0 & & \tilde{r}_{1m_1} \end{bmatrix}, \quad \tilde{r}_{1_i} \neq 0 \qquad (1 \leqslant i \leqslant m_1). \tag{19}$$

Let

$$\bar{A}_1 = Q_{11} A_1 Q_{11}^t = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{13} & \bar{A}_{14} \end{bmatrix}$$

$$\bar{b}_1 = Q_{11} b_1, \quad \text{and} \quad \bar{c}_1 = c_1 Q_{11}^t$$

where $\bar{A}_{11} \in R^{m_1 \times m_1}$ and $\bar{A}_{14} \in R^{(m-m_1)\times(m-m_1)}$. Now from (18) and (19)

$$[\bar{b}_1 \quad \bar{A}_1 \bar{b}_1 \cdots \bar{A}_1^{m_1-1} \bar{b}_1] = \begin{bmatrix} \tilde{r}_{11} & & * \\ & \ddots & \\ 0 & & \tilde{r}_{1m_1} \\ \hline & 0 & \end{bmatrix} \tag{20}$$

which gives

$$\bar{b}_1 = \begin{bmatrix} \tilde{r}_{11} \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{21}$$

In the present case, the column vector $\bar{A}_1^{m_1}\bar{b}_1$ is a linear combination of the vectors $\{\bar{A}_1^i\bar{b}_1,\ 0 \leqslant i \leqslant m_1\}$. That is, there are m_1 real numbers $\{\alpha_i,\ 1 \leqslant i \leqslant m_1\}$ such that

$$\bar{A}_1^{m_1}\bar{b}_1 = \alpha_1\bar{b}_1 + \cdots + \alpha_{m_1}\bar{A}_1^{m_1-1}\bar{b}_1.$$

In effect

$$\begin{aligned}\bar{A}_1&\left[\bar{b}_1 \quad \bar{A}_1\bar{b}_1 \cdots \bar{A}_1^{m_1-1}\bar{b}_1\right]\\ &= \left[\bar{b}_1 \quad \bar{A}_1\bar{b}_1 \cdots \bar{A}_1^{m_1-1}\bar{b}_1\right] \cdot \begin{bmatrix} 0 & 0 & \cdots & 0 & \alpha_1 \\ 1 & 0 & \cdots & 0 & \cdot \\ 0 & 1 & \cdots & 0 & \cdot \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & \alpha_{m_1} \end{bmatrix}_{m_1 \times m_1}\end{aligned} \tag{22}$$

and by virtue of (20), eq. (22) implies

$$\begin{bmatrix} \bar{A}_{11} \\ \bar{A}_{13} \end{bmatrix} = \begin{bmatrix} * & * & \cdots & * & * \\ \frac{\tilde{r}_{12}}{\tilde{r}_{11}} & * & \cdots & & \\ & \frac{\tilde{r}_{13}}{\tilde{r}_{12}} & & \vdots & \vdots \\ & & \ddots & & \\ 0 & & & \frac{\tilde{r}_{1m_1}}{\tilde{r}_{1,m_1-1}} & * \\ \hline & & 0 & & \end{bmatrix}_{m \times m_1} \tag{23}$$

In other words, because of the lack of reachability, the use of matrix Q_{11} cannot zero the entries in $\bar{A}_{14}$. However, one may use another orthogonal transformation Q_{12}, which is of the form

$$Q_{12} = \begin{bmatrix} I_{m_1} & 0 \\ 0 & T \end{bmatrix}$$

where T is an $(m-m_1)\times(m-m_1)$ orthogonal matrix to $\bar{A}_1$ such that

$$Q_{12}\bar{A}_1Q_{12}^t = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12}T \\ T\bar{A}_{13} & T\bar{A}_{14}T^t \end{bmatrix} = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12}T^t \\ 0 & T\bar{A}_{14}T^t \end{bmatrix}. \tag{24}$$

By Theorem A.3, we can choose an orthogonal matrix T such that $T\bar{A}_{14}T^t$ is the real Schur decomposition (RSD) of $\bar{A}_{14}$, namely

$$T\bar{A}_{14}T^t = \begin{bmatrix} H_1 & & * \\ & H_2 & \\ & & \ddots & \\ & 0 & & H_k \end{bmatrix} \tag{25}$$

where each H_i is either a scalar or a 2×2 matrix having complex conjugate eigenvalues. Matrix Q_{12} preserves the zeros in $\bar{b}_1$. Consequently, by means of the orthogonal similarity transformation $Q_1 \equiv Q_{12}Q_{11}$ we have

$$Q_1A_1Q_1^t = Q_{12}\bar{A}_1Q_{12}^t = \left[\begin{array}{ccccc|cccc} * & * & \cdots & * & * & & & & \\ \frac{\tilde{r}_{12}}{\tilde{r}_{11}} & * & \cdots & * & * & & & & \\ & \frac{\tilde{r}_{13}}{\tilde{r}_{12}} & & & \vdots & & * & & \\ & & \ddots & & & & & & \\ & 0 & & \frac{\tilde{r}_{1m_1}}{\tilde{r}_{1,m_1-1}} & * & & & & \\ \hline & & & & & H_1 & & * & \\ & & 0 & & & & H_2 & & \\ & & & & & & & \ddots & \\ & & & & & & 0 & & H_k \end{array}\right] \tag{26}$$

and

$$Q_1b_1 = Q_{12}\bar{b}_1 = \begin{bmatrix} \tilde{r}_{11} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{27}$$

in which the number of zero entries is at least $m(m-1)/2$. When subsystem S_2 is neither reachable nor observable, one can zero at least $n(n-1)/2$ entries in (A_2, b_2, c_2) in a similar manner. These results establish the following theorem.

Theorem 1: There exists an orthogonal similarity transformation $Q = Q_1 \oplus Q_2$ which forces at least $[m(m-1)+n(n-1)]/2$ zero entries in the realization (QAQ^t, Qb, cQ^t).

Theorem 1 implies that at least $[m(m-1)+n(n-1)]/2$ multiplications can be eliminated in the realization of (1) by the above technique.

III. Sensitivity Analysis

In this section, the similarity transformation

$$T = DQ \tag{28}$$

where

$$D = \operatorname{diag}(d_1 \cdots d_{m+n}), \quad d_i \neq 0 \qquad (1 \leqslant i \leqslant m+n)$$

and Q is orthogonal will be used to introduce $m+n$ free parameters which may be chosen to reduce the sensitivities of a digital filter with respect to the multipliers or to eliminate $m+n$ additional multiplications while keeping the digital filter free of overflow oscillations.

Since $\|A\| < 1$, where A is the system matrix in (1a), matrix W defined as

$$W = I - A^tA \tag{29}$$

is positive definite. With the following assignments

$$\tilde{A} = TAT^{-1}, \quad \tilde{b} = Tb, \quad \tilde{c} = cT^{-1} \tag{30}$$

where T is given in (28), $W > 0$ implies that

$$\begin{aligned} D^{-1}QWQ^tD^{-1} &= D^{-1}Q(I - A^tA)Q^tD^{-1} \\ &= D^{-2} - D^{-1}QA^tAQ^tD^{-1} \\ &= D^{-2} - \tilde{A}^tD^{-2}\tilde{A} > 0. \end{aligned} \tag{31}$$

Therefore, by Theorem 2 of [4] the realization $(\tilde{A}, \tilde{b}, \tilde{c})$ is free of overflow oscillations.

It should be noted that if the orthogonal matrix Q is chosen to force at least $[m(m-1)+n(n-1)]/2$ zero entries in realization (QAQ^t, Qb, cQ^t), then the similarity transformation given by (28) will preserve these zero entries in realization (TAT^{-1}, Tb, cT^{-1}). Now let

$$G(z_1, z_2) = c[\Gamma(z_1, z_2) - A]^{-1}b$$

with

$$\Gamma(z_1, z_2) = \begin{bmatrix} z_1 I_n & 0 \\ 0 & z_2 I_m \end{bmatrix}$$

be the transfer function of the filter characterized by (1). A common measure of sensitivity with respect to the coefficients in (A, b, c) is given by [9] as

$$S(z_1, z_2) = \operatorname{trace}\left\{\left[\frac{\partial G}{\partial A}\right]\left[\frac{\partial G}{\partial A}\right]^* + \left[\frac{\partial G}{\partial b}\right]\left[\frac{\partial G}{\partial b}\right]^* + \left[\frac{\partial G}{\partial c}\right]\left[\frac{\partial G}{\partial c}\right]^*\right\} \tag{32}$$

where

$$\frac{\partial G}{\partial A} = [\Gamma(z_1, z_2) - A]^{-t}c^tb^t[\Gamma(z_1, z_2) - A]^{-t}$$

$$\frac{\partial G}{\partial b} = [\Gamma(z_1, z_2) - A]^{-t}c^t$$

$$\frac{\partial G}{\partial c} = b^t[\Gamma(z_1, z_2) - A]^{-t}$$

and $[\Gamma(z_1, z_2) - A]^{-t}$ denotes the transpose of $[\Gamma(z_1, z_2) - A]^{-1}$, and $*$ is the complex-conjugate transpose. Likewise, if the similarity transformation T given by (28) is applied, the sensitivity with respect to the coefficients in $(\tilde{A}, \tilde{b}, \tilde{c})$ is given by

$$\tilde{S}(z_1, z_2) = \operatorname{trace}\left\{\left[\frac{\partial G}{\partial A}\right]Q^tD^2Q\left[\frac{\partial G}{\partial A}\right]^*Q^tD^{-2}Q + \left[\frac{\partial G}{\partial b}\right]\left[\frac{\partial G}{\partial b}\right]^*Q^tD^{-2}Q + \left[\frac{\partial G}{\partial c}\right]Q^tD^2Q\left[\frac{\partial G}{\partial c}\right]^*\right\}. \tag{33}$$

To demonstrate the possibility of reducing the sensitivity with respect to the coefficients in $(\tilde{A}, \tilde{b}, \tilde{c})$ in a given frequency range by adjusting parameters $\{d_i, 1 \leqslant i \leqslant m+n\}$, let us examine the simplest case where $D = dI_{m+n}$, and d is a nonzero scalar parameter. In this case (33) becomes

$$\tilde{S}(z_1, z_2) = \operatorname{trace}\left\{\left[\frac{\partial G}{\partial A}\right]\left[\frac{\partial G}{\partial A}\right]^* + d^{-2}\left[\frac{\partial G}{\partial b}\right]\left[\frac{\partial G}{\partial b}\right]^* + d^2\left[\frac{\partial G}{\partial c}\right]\left[\frac{\partial G}{\partial c}\right]^*\right\}$$

and if

$$\operatorname{trace}\left[\frac{\partial G}{\partial b}\right]\left[\frac{\partial G}{\partial b}\right]^* < \operatorname{trace}\left[\frac{\partial G}{\partial c}\right]\left[\frac{\partial G}{\partial c}\right]^*$$

in the given frequency range, then the use of a smaller d leads to lower sensitivity in that frequency range.

IV. Further Reduction in the Number of Multiplications

An alternative way to take advantage of the similarity transformation in (28) is to force $m+n$ multiplier constants in the digital filter represented by (1) to be unity. This technique leads to the elimination of further $m+n$ multiplications in addition to the possible $[m(m-1)+n(n-1)]/2$ multiplications that can be eliminated by the technique of Section II.

For the sake of simplicity, we assume that both pairs (A_1, b_1) and (A_4, b_2) are reachable so that by the analysis given in the previous section there exists an orthogonal similarity transformation $Q = Q_1 \oplus Q_2$ such that

$$
\bar{A} = QAQ^t = \left[\begin{array}{ccccc|ccccc}
* & * & \cdots & * & * & & & & & \\
\frac{r_{12}}{r_{11}} & * & \cdots & * & * & & & & & \\
 & \frac{r_{13}}{r_{12}} & & \vdots & \vdots & & & * & & \\
 & & \ddots & & & & & & & \\
 & 0 & & \frac{r_{1m}}{r_{1,m-1}} & * & & & & & \\
\hline
 & & & & & * & * & & * & * \\
 & & & & & \frac{r_{22}}{r_{21}} & * & & * & * \\
 & & * & & & & \frac{r_{23}}{r_{22}} & & \vdots & \vdots \\
 & & & & & & & \ddots & & \\
 & & & & & & 0 & & \frac{r_{2n}}{r_{2,n-1}} & *
\end{array}\right] \tag{34a}
$$

and

$$
\bar{b} = Qb = \left[\begin{array}{c}
r_{11} \\ 0 \\ \vdots \\ 0 \\ \hline r_{21} \\ 0 \\ \vdots \\ 0
\end{array}\right]. \tag{34b}
$$

Since r_{1i} and r_{2j} for $1 \leqslant i \leqslant m$, $1 \leqslant j \leqslant n$ are nonzero, the diagonal matrix D in (28) can be constructed as

$$
D = \begin{bmatrix}
r_{11}^{-1} & & & & & \\
 & \ddots & & & 0 & \\
 & & r_{1m}^{-1} & & & \\
 & & & r_{21}^{-1} & & \\
 & 0 & & & \ddots & \\
 & & & & & r_{2n}^{-1}
\end{bmatrix}. \tag{35}
$$

Hence, (34) assumes the form

$$
TAT^{-1} = D\bar{A}D^{-1} = \left[\begin{array}{ccccc|ccccc}
* & * & \cdots & * & * & & & & & \\
1 & * & \cdots & * & * & & & & & \\
 & 1 & & \vdots & \vdots & & & * & & \\
 & & \ddots & & & & & & & \\
 & 0 & & 1 & * & & & & & \\
\hline
 & & & & & * & * & \cdots & * & * \\
 & & & & & 1 & * & \cdots & * & * \\
 & & * & & & & 1 & & \vdots & \vdots \\
 & & & & & & & \ddots & & \\
 & & & & & & 0 & & 1 & *
\end{array}\right] \tag{36a}
$$

and

$$Tb = D\bar{b} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ \hline 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{36b}$$

Thus, the similarity transformation given in (28) leads to $n+m$ unit multipliers and preserves the zero entries appearing in (34). Since D is a nonsingular diagonal matrix, the resulting realization (TAT^{-1}, Tb, cT^{-1}) is also free of overflow oscillations. However, the sensitivity of the frequency response with respect to the multipliers may increase.

If (A_4, b_2) is a controllable pair, but pair (A_1, b_1) is neither reachable nor observable, the approach described in the latter part of Section II can be used to find an orthogonal matrix $Q = Q_1 \oplus Q_2$ such that (26) and (27) hold. The diagonal matrix D in (1) can then be constructed as

$$D = \left[\begin{array}{ccc|c|cccc} \tilde{r}_{11}^{-1} & & & & & & & \\ & \ddots & & 0 & & & 0 & \\ & & \tilde{r}_{1,m_1}^{-1} & & & & & \\ \hline & 0 & & D_1 & & & & \\ \hline & & & & r_{21}^{-1} & & & \\ & & 0 & & & r_{22}^{-1} & & \\ & & & & & & \ddots & \\ & & & & & & & r_{2n}^{-1} \end{array}\right] \tag{37}$$

where D_1 is a diagonal matrix of dimension $m - m_1$, which can easily be determined by noting the structure of matrices H_i $(1 \leqslant i \leqslant k)$ given in (26). For example, if $k=2$ and both H_1 and H_2 are 2×2 matrices, i.e.,

$$\begin{bmatrix} H_1 & * \\ 0 & H_2 \end{bmatrix} = \left[\begin{array}{cc|cc} * & * & & \\ r_1 & * & & * \\ \hline & & * & * \\ & 0 & r_2 & * \end{array}\right]$$

where $r_1 \neq 0$ and $r_2 \neq 0$, then matrix D_1 in (37) can be taken as

$$D_1 = \begin{bmatrix} 1 & & & 0 \\ & r_1^{-1} & & \\ & & 1 & \\ 0 & & & r_2^{-1} \end{bmatrix} \tag{38}$$

which gives

$$D_1 \begin{bmatrix} H_1 & * \\ 0 & H_2 \end{bmatrix} D_1^{-1} = \left[\begin{array}{cc|cc} * & * & & \\ 1 & * & & * \\ \hline & & * & * \\ & 0 & 1 & * \end{array}\right].$$

Therefore, using the similarity transformation given by (28) with D defined as in (37), we can eliminate at least $[m(m-1)+n(n-1)]/2$ multiplications. In addition, we can force the subdiagonal entries in A_1 and A_4 to be unity or zero. The resulting realization (TAT^{-1}, Tb, cT^{-1}) is free of overflow oscillations, as in the previous cases.

V. Example

Let us consider a first-quadrant Gaussian filter of the order (4,2) characterized by

$$A = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

$$c = [c_1 \quad c_2], \quad \text{and} \quad d \tag{39}$$

where

$$A_1 = \begin{bmatrix} 0.174340E+01 & 0.117383E+01 & 0.143891E+00 & 0.296357E-01 \\ -0.921900E+00 & -0.225628E+00 & 0.278089E-01 & 0.875035E-01 \\ 0.297146E-01 & -0.180827E-01 & -0.498595E-01 & 0.919117E+00 \\ -0.427139E-03 & -0.836201E-02 & 0.302893E-01 & -0.114475E+00 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} -0.462118E-01 & 0.538983E-01 \\ 0.444979E-01 & -0.567200E-01 \\ -0.456776E-02 & 0.347877E-02 \\ 0.155149E-01 & 0.355290E-02 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} 0.112382E+01 & -0.165127E+00 & 0.315924E-01 & -0.577690E-01 \\ 0.358407E-01 & 0.338645E-01 & -0.288409E-01 & 0.575798E-01 \end{bmatrix}$$

$$A_4 = \begin{bmatrix} 0.188585E+01 & -0.109236E+01 \\ 0.110738E+01 & -0.229426E+00 \end{bmatrix}$$

$$b_1 = \begin{bmatrix} 0.229943E+01 \\ -0.389516E+00 \\ -0.253897E-01 \\ -0.650878E-02 \end{bmatrix}$$

$$b_2 = \begin{bmatrix} 0.104029E+01 \\ -0.376250E-01 \end{bmatrix}$$

$$c_1 = [0.310808E-01 \quad 0.708642E-01 \quad 0.870614E+00 \quad 0.353070E-01]$$

$$c_2 = [0.124361E-01 \quad 0.171934E-02]$$

and

$$d = 0.943040E-02.$$

This filter was designed by Aly and Fahmy [10] and was used by Aboulnasr and Fahmy [8] to illustrate their approach for the elimination of some multiplications in the system matrix. The approach suggested in [8] can eliminate six multiplications in the system matrix in a total of 49 multipliers.

It is easy to verify that both subsystems (A_1, b_1) and (A_4, b_2) in (39) are reachable so that the desirable orthogonal similarity matrix can be formed as

$$Q = Q_1 \oplus Q_2 \tag{40}$$

where Q_1 and Q_2 are obtained from QRD of F_1 and F_2, respectively. Numerical computation [11], [12] gives

$$Q_1 = \begin{bmatrix} 0.985891 & -0.167006 & -0.010885 & -0.002790 \\ -0.165606 & -0.982951 & 0.079353 & 0.009146 \\ 0.022919 & 0.070613 & 0.865399 & 0.495552 \\ 0.008165 & 0.030474 & 0.494639 & -0.868525 \end{bmatrix} \tag{41}$$

and

$$Q_2 = \begin{bmatrix} 0.999346 & -0.036144 \\ 0.036144 & 0.999346 \end{bmatrix}. \tag{42}$$

By applying the orthogonal transformation Q given by (40)–(42) to (39), we have $\{\bar{A} = QAQ^T, \bar{b} = Qb, \bar{c} = cQ^T, \bar{d} = d\}$ where

$$\bar{A} = \begin{bmatrix} \bar{A}_1 & \bar{A}_2 \\ \bar{A}_3 & \bar{A}_4 \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} \bar{b}_1 \\ \bar{b}_2 \end{bmatrix}, \quad \bar{c} = [\bar{c}_1 \quad \bar{c}_2]$$

and

$$\bar{A}_1 = \begin{bmatrix} 1.644913 & -1.473914 & 0.248891 & 0.115540 \\ 0.607402 & -0.132606 & -0.040859 & -0.004748 \\ 0.0 & 0.013613 & 0.349038 & -0.659096 \\ 0.0 & 0.0 & 0.233746 & -0.507907 \end{bmatrix}$$

$$\bar{A}_2 = \begin{bmatrix} -0.055211 & 0.060606 \\ -0.037986 & 0.045792 \\ 0.005742 & 0.002210 \\ -0.014650 & -0.003185 \end{bmatrix}$$

$$\bar{A}_3 = \begin{bmatrix} 1.133539 & -0.020326 & 0.012556 & 0.072174 \\ 0.070849 & -0.041746 & 0.007246 & -0.060381 \end{bmatrix}$$

$$\bar{A}_4 = \begin{bmatrix} 1.882544 & -1.015974 \\ 1.183765 & -0.226120 \end{bmatrix}$$

$$\bar{b}_1 = \begin{bmatrix} 2.332335 \\ 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$

$$\bar{b}_2 = \begin{bmatrix} 1.040970 \\ 0.0 \end{bmatrix}$$

$$\bar{c}_1 = [0.009231 \quad -0.005393 \quad 0.776641 \quad 0.402388]$$

$$\bar{c}_2 = [0.012365 \quad 0.002167].$$

This realization entails the elimination of seven multiplications.

Further notice that QRD of F_1 and F_2 also gives

$$R_1 = \begin{bmatrix} 2.332335 & & & * \\ & 1.141666 & & \\ & & 0.019286 & \\ 0 & & & 0.004508 \end{bmatrix}$$

and

$$R_2 = \begin{bmatrix} 1.040970 & * \\ 0.0 & 1.232264 \end{bmatrix}$$

so that one can form matrix D in (28) as

$$D = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}$$

where

$$D_1 = \operatorname{diag}\{2.332335^{-1}, 1.416666^{-1}, 0.019286^{-1}, 0.004508^{-1}\}$$

and

$$D_2 = \operatorname{diag}\{1.040970^{-1}, 1.232264^{-1}\}.$$

By applying the transformation $T = DQ$, we have $\{\tilde{A} = TAT^{-1},\ \tilde{b} = Tb,\ \tilde{c} = cT^{-1},\ \tilde{d} = d\}$ where

$$\tilde{A} = \begin{bmatrix} \tilde{A}_1 & \tilde{A}_2 \\ \tilde{A}_3 & \tilde{A}_4 \end{bmatrix}, \quad \tilde{b} = \begin{bmatrix} \tilde{b}_1 \\ \tilde{b}_2 \end{bmatrix}, \quad \tilde{c} = [\tilde{c}_1 \quad \tilde{c}_2]$$

and

$$\tilde{A}_1 = \begin{bmatrix} 1.644913 & -0.895259 & 0.002058 & 0.000223 \\ 1.0 & -0.132606 & -0.000556 & -0.000015 \\ 0.0 & 1.0 & 0.349038 & 0.154060 \\ 0.0 & 0.0 & 1.0 & -0.507907 \end{bmatrix}$$

$$\tilde{A}_2 = \begin{bmatrix} -0.024642 & 0.032021 \\ -0.027912 & 0.039831 \\ 0.309927 & 0.141206 \\ -3.382922 & -0.870621 \end{bmatrix}$$

$$\tilde{A}_3 = \begin{bmatrix} 2.539739 & -0.027662 & 0.000233 & 0.000313 \\ 0.134098 & -0.047993 & 0.000113 & -0.000221 \end{bmatrix}$$

$$\tilde{A}_4 = \begin{bmatrix} 1.882544 & -1.202675 \\ 1.0 & -0.226120 \end{bmatrix}$$

$$\tilde{b}_1 = \begin{bmatrix} 1.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$

$$\tilde{b}_2 = \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

$$\tilde{c}_1 = [0.021530 \quad 0.007640 \quad 0.014978 \quad 0.001814]$$

$$\tilde{c}_2 = [0.012872 \quad 0.002670].$$

It is seen that in addition to the seven zero entries, six other entries have been forced to be unity in the resulting realization. This increases the number of multiplications eliminated to 13.

V. Conclusions

It has been shown that as many as $[m(m-1)+n(n-1)]/2$ entries can be forced to be zero in an LSS realization through the use of an appropriate transformation from the class of orthogonal similarity transformations. This desirable transformation can be obtained by the QR decomposition of the reachability or the observability matrices. Further, it has been demonstrated that a suitable use of a broader class of similarity transformations can result in either introducing $m+n$ free parameters in the sensitivity function which may be appropriately chosen to reduce the sensitivity with respect to the multipliers or to force additional $m+n$ multipliers to be unity while keeping the filter free of overflow oscillations. In effect, a total of $[m(m+1)+n(n+1)]/2$ multiplications can be eliminated in an LSS realization. This is a significant improvement relative to the result given in [8], particularly in the case where $|m-n|$ is large. A numerical example has been given which illustrates the techniques described.

Appendix

This appendix summarizes three theorems of linear algebra which have been used in the paper. The proofs of these theorems can be found in any standard text of numerical analysis, e.g., [11].

Theorem A.1 (Singular Value Decomposition (SVD)): If $A \in R^{M\times N}$ $(M \geq N)$, then there exist orthogonal matrices $U = [u_1 \cdots u_M] \in R^{M\times M}$ and $V = [v_1 \cdots v_N] \in R^{N\times N}$ such that

$$A = U \begin{bmatrix} \begin{matrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_N \end{matrix} \\ \hline 0 \end{bmatrix} V^t \tag{2.1}$$

where $\sigma_1 \geq \cdots \geq \sigma_N$ are the singular values of A (i.e., the nonnegative square roots of the eigenvalues of A^tA); the columns $\{v_i,\ 1 \leq i \leq N\}$ form a complete orthonormal basis of the eigenvectors of A^tA; the columns $\{u_i,\ 1 \leq i \leq M\}$ form a complete orthonormal basis of the eigenvectors of AA^t.

Theorem A.2 (QR Decomposition (QRD)): If $A \in R^{M\times N}$ $(M \geq N)$, then there exists an orthogonal matrix Q such that

$$A = Q^t \begin{bmatrix} R \\ 0 \end{bmatrix}$$

where R is an $N \times N$ upper triangular matrix. If A has rank N, then the first N columns of Q^t form an orthonormal basis for the space spanned by the columns of A.

Theorem A.3 (Real Schur Decomposition (RSD)): If $A \in R^{N\times N}$, then there exists an orthogonal $Q \in R^{N\times N}$ such that

$$Q^TAQ = \begin{bmatrix} H_1 & & & * \\ 0 & H_2 & & \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & H_k \end{bmatrix}$$

where each H_i is either a scalar or a 2×2 matrix having complex conjugate eigenvalues.

Quite a few numerically stable algorithms for obtaining SVD, QRD, and RSD for a given matrix are available. The interested reader is referred to [11] and [12].

References

[1] C. W. Barnes and A. T. Fam, "Minimum norm recursive digital filters that are free of overflow limit cycles," *IEEE Trans. Circuits Syst.*, vol. CAS-24, pp. 567–574, 1977.
[2] W. L. Mills, C. T. Mullis, and R. A. Roberts, "Digital filter realization without overflow oscillations," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 334–338, 1978.
[3] C. W. Barnes, "Roundoff noise and overflow in normal digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 154–159, 1979.
[4] N. G. El-Agizi and M. M. Fahmy, "2-D filters with no overflow oscillations," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 465–469, 1979.
[5] N. G. El-Agizi, "Minimization of finite word length effects in 2-D digital filters," Ph.D. dissertation, Dept. Elec. Eng., Queen's Univ., Kingston, Canada, 1979.
[6] J. H. Lodge and M. M. Fahmy, "Stability and overflow oscillations in 2-D state-space digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 1161–1171, 1981.
[7] R. P. Roesser, "A discrete state-space model for linear image processing," *IEEE Trans. Automat. Contr.*, vol. AC-20, pp. 1–10, 1975.
[8] T. T. Aboulnasr and M. M. Fahmy, "2-D state-space realizations with fewer multipliers and invariant norms," *IEEE Trans. Circuits Syst.*, vol. CAS-30, pp. 804–808, 1983.
[9] D. S. K. Chan, "Theory and implementation of multidimensional discrete systems for signal processing," Ph.D. dissertation, Dept. Electrical Eng. Computer Science, MIT, Cambridge, MA, 1978.
[10] S. A. H. Aly and M. M. Fahmy, "Spatial-domain design of two-dimensional recursive digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-27, pp. 892–901, Oct. 1980.
[11] G. H. Golub and C. F. Van Loan, *Matrix Computation*. Washington, D.C.: Johns Hopkins Univ. Press, 1983.
[12] J. J. Dongarra, C. B. Moler, J. R. Bunch, and G. W. Stewart, *LIN-PACK User's Guide*. Philadelphia, PA: SIAM, 1979.
[13] S. Y. Kung, B. C. Levy, M. Morf, and T. Kailath, "New results in 2-D systems theory, part II: 2-D state-space models—realization and the notions of controllability, observability, and minimality," *IEEE Proc.*, vol. 65, pp. 945–961, 1977.

Wu-Sheng Lu (S'81–M'86) received the B.S. and M.S. degrees in mathematics from Fudan University and East China Normal University, China, in 1964 and 1980, respectively. He received the M.S. degree in electrical engineering and the Ph.D. degree in control science from the University of Minnesota, in 1983 and 1984, respectively.

From October 1984 to December 1985, he was with the Department of Electrical Engineering, University of Victoria, Canada, as a Postdoctoral Fellow. Currently, he is a Visiting Assistant Professor of Electrical Engineering at the University of Minnesota. His research interests include systems theory, and analysis and synthesis of multidimensional digital filters.

Andreas Antoniou (M'69–SM'79–F'82) received the B.Sc.(Eng.) and Ph.D. degrees in electrical engineering from London University, London, UK, in 1963 and 1966, respectively.

From 1966 to 1969, he was Senior Scientific Officer at the Post Office Research Department, London, England, and from 1969 to 1970, he was a member of the Scientific Staff at the R&D Laboratories of Northern Electric Company Ltd., Ottawa, Ontario, Canada. From 1970 to 1983, he served in the Department of Electrical Engineering, Concordia University, Montreal, Quebec, Canada, as Professor from June 1973 and as Chairman from December 1977. On July 1, 1983, he was appointed Founding Chairman of the Department of Electrical Engineering, University of Victoria, Victoria, B.C., Canada.

His teaching and research interests are in the areas of electronics, network synthesis, digital system design, active and digital filters, and digital signal processing. He has published a number of papers on electronic circuits, active filters, and digital filters. He has authored *Digital Filters: Analysis and Design* (New York: McGraw-Hill). One of his papers on gyrator circuits was awarded the Ambrose Fleming Premium by the Institution of Electrical Engineers, UK.

Dr. Antoniou is a Member of the Order of Engineers of Quebec and a Fellow of the Institution of Electrical Engineers. He was Associate Editor for IEEE Transactions on Circuits and Systems during the period June 1983 to May 1985. He is now serving as Editor of the same Transactions.