Separate/Joint Optimization of Error Feedback and Coordinate Transformation for Roundoff Noise Minimization in Two-Dimensional State-Space Digital Filters

Takao Hinamoto, *Fellow, IEEE*, Keisuke Higashi, and Wu-Sheng Lu, *Fellow, IEEE*

***Abstract*—This paper is concerned with the minimization of roundoff noise subject to l_2-norm dynamic-range scaling constraints in two-dimensional (2-D) state-space digital filters. Two methods are proposed, with the first one using error feedback alone and the second one using joint error feedback and coordinate transformation optimization. In the first method, several techniques for the determination of optimal full-scale, block-diagonal, diagonal, and scalar error-feedback matrices for a given 2-D state-space digital filter are proposed. In the second method, an iterative approach for minimizing the roundoff noise under l_2-norm dynamic-range scaling constraints is developed by jointly optimizing a scalar error-feedback matrix and a coordinate transformation matrix, which may be regarded as an alternative approach to the conventional method for synthesizing the optimal 2-D filter structure with minimum roundoff noise. An analytical method for the joint optimization of a general error-feedback matrix and a coordinate transformation matrix under the scaling constraints is also proposed. A numerical example is presented to illustrate the utility of the proposed techniques.**

***Index Terms*—Optimal coordinate transformation, optimal error feedback, roundoff noise minimization, scaling constraints, 2-D state-space digital filters.**

I. Introduction

DUE to the existence of an infinite number of realizations for a given transfer function $H(z)$, there is a certain degree of freedom in choosing a particular realization of the filter. This freedom is often used to optimize some criterion associated with a particular algorithm or realization. If $H(z)$ is realized through hardware implementation using fixed-point arithmetic, then the internal noise caused by finite-word-length (FWL) registers may be the most serious concern with which to deal. One of the primary FWL register effects in fixed-point digital filters is the roundoff noise caused by the rounding of products/summations within the realization. Although hardware implementation of dynamic systems and digital signal processing modules with large data length becomes increasingly affordable in many applications, high implementation cost remains a concern, especially for multidimensional dynamic systems in which a large number multipliers are involved. In addition, the increased execution time needed for carrying out many multiplications of long-length numbers is obviously out of favor for real-time applications. The synthesis of state-space digital filter structures with minimum rounfoff noise under l_2-norm dynamic-range scaling constraints has been investigated in [1]–[4], and the investigation has been extended to 2-D state-space digital filters in [5]–[8]. Another technique for the reduction of roundoff noise at the filter output is to use error feedback (EF). The EF is achieved by extracting the quantization error after multiplication and addition and then feeding the error signal back to a certain point through a simple circuit. Many techniques for EF have been presented in the past for one-dimensional (1-D) digital filters [9]–[18], and more recently, for 2-D digital filters [19]–[23]. It has also been shown that the roundoff noise can be reduced by means of delta operator [24]–[26] and the digital filter in this case can be viewed as a special case of the filter with EF [24].

This paper proposes two new methods for the reduction of roundoff noise in 2-D state-space digital filters. Several closed-form formulas for evaluating the optimal full-scale, block-diagonal, diagonal, and scalar EF matrices for a given state-space digital filter are derived. Then, an iterative noise reduction technique for state-space digital filters is developed by jointly optimizing a scalar EF matrix and a coordinate transformation matrix subject to l_2-norm dynamic-range scaling constraints. An analytical method for the joint optimization of a general EF matrix and a coordinate transformation matrix under the scaling constraints is also proposed. Although the objective function involved in the joint optimization is not convex in general, and a rigorous mathematical proof of a global convergence property of the algorithm is not available at present, in every case of our fairly extensive computer simulations, the algorithm converges to an identical solution, regardless of the choice of an initial point. A numerical example is presented to illustrate the algorithms proposed and to demonstrate their performance.

Throughout the paper, $\boldsymbol{I}_n$ stands for the identity matrix of dimension $n \times n$, the transpose (conjugate transpose) of a matrix $\boldsymbol{A}$ is indicated by $\boldsymbol{A}^T(\boldsymbol{A}^*)$, and the trace and ith diagonal element of a square matrix $\boldsymbol{A}$ are denoted by tr$[\boldsymbol{A}]$ and $(\boldsymbol{A})_{ii}$, respectively.

Manuscript received April 15, 2002; revised February 11, 2003. The associate editor coordinating the review of this paper and approving it for publication was Prof. Derong Liu.

T. Hinamoto and K. Higashi are with Graduate School of Engineering, Hiroshima University, Higashi-Hiroshima, Japan (e-mail: hinamoto@hiroshima-u.ac.jp).

W.-S. Lu is with Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC V8W 3P6 Canada (e-mail: wslu@ece.uvic.ca).

Digital Object Identifier 10.1109/TSP.2003.815379

II. Two–Dimensional State-Space Digital Filters With Error Feedback

Consider the following single-input/single-output local state-space (LSS) model $(\boldsymbol{A},\boldsymbol{b},\boldsymbol{c},d)_{m,n}$ for 2-D digital filters which was originally proposed by Roesser [27]:

$$\begin{aligned}\boldsymbol{x}_{11}(i,j) &= \boldsymbol{A}\boldsymbol{x}(i,j)+\boldsymbol{b}u(i,j)\\ y(i,j) &= \boldsymbol{c}\boldsymbol{x}(i,j)+du(i,j)\end{aligned}\tag{1}$$

where

$$\boldsymbol{x}_{11}(i,j)=\begin{bmatrix}\boldsymbol{x}^h(i+1,j)\\ \boldsymbol{x}^v(i,j+1)\end{bmatrix},\quad \boldsymbol{x}(i,j)=\begin{bmatrix}\boldsymbol{x}^h(i,j)\\ \boldsymbol{x}^v(i,j)\end{bmatrix}$$
$$\boldsymbol{A}=\begin{bmatrix}\boldsymbol{A}_1 & \boldsymbol{A}_2\\ \boldsymbol{A}_3 & \boldsymbol{A}_4\end{bmatrix},\quad \boldsymbol{b}=\begin{bmatrix}\boldsymbol{b}_1\\ \boldsymbol{b}_2\end{bmatrix},\quad \boldsymbol{c}=[\boldsymbol{c}_1\quad \boldsymbol{c}_2].$$

Here, $\boldsymbol{x}^h(i,j)$ is an $m\times 1$ horizontal state vector, $\boldsymbol{x}^v(i,j)$ is an $n\times 1$ vertical state vector, $u(i,j)$ is a scalar input, $y(i,j)$ is a scalar output, and $\boldsymbol{A}_1,\boldsymbol{A}_2,\boldsymbol{A}_3,\boldsymbol{A}_4,\boldsymbol{b}_1,\boldsymbol{b}_2,\boldsymbol{c}_1,\boldsymbol{c}_2$, and d are real constant matrices of appropriate dimensions. The LSS model in (1) is assumed to be BIBO stable, separately locally controllable, and separately locally observable [28].

Because of finite register sizes, FWL constraints are imposed on the local state vector, input, output, and coefficients in the filter realization $(\boldsymbol{A},\boldsymbol{b},\boldsymbol{c},d)_{m,n}$. By considering the quantization carried out before matrix-vector multiplication, an FWL implementation of (1) can be expressed as

$$\begin{aligned}\tilde{\boldsymbol{x}}_{11}(i,j) &= \boldsymbol{A}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,j)]+\boldsymbol{b}u(i,j)\\ \tilde{y}(i,j) &= \boldsymbol{c}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,j)]+du(i,j)\end{aligned}\tag{2}$$

where each component of coefficient matrices $\boldsymbol{A},\boldsymbol{b},\boldsymbol{c}$, and d assumes an exact fractional B_c bit representation. The FWL local state vector $\tilde{\boldsymbol{x}}(i,j)$ and the output $\tilde{y}(i,j)$ all have a B bit fractional representation, whereas the input $u(i,j)$ is a $(B-B_c)$ bit fraction.

The quantizer $\boldsymbol{Q}[\,\cdot\,]$ in (2) rounds the B bit fraction $\tilde{\boldsymbol{x}}(i,j)$ to $(B-B_c)$ bits after the multiplications and additions, where the sign bit is not counted. In a fixed-point implementation, the quantization is usually performed by two's complement truncation that discards the lower bits of a double-precision accumulator. Thus, the quantization error

$$\boldsymbol{e}(i,j)=\tilde{\boldsymbol{x}}(i,j)-\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,j)]\tag{3}$$

coincides with the residue left in the lower part of $\tilde{\boldsymbol{x}}(i,j)$. The roundoff error $\boldsymbol{e}(i,j)$ is modeled as a zero-mean noise process of covariance $\sigma^2\boldsymbol{I}_{m+n}$ with

$$\sigma^2=\frac{1}{12}2^{-2(B-B_c)}.$$

In an effort to reduce the filter's roundoff noise, the quantization error $\boldsymbol{e}(i,j)$ is fed back to each input of delay operators through an $(m+n)\times(m+n)$ constant matrix $\boldsymbol{D}$ in the FWL filter (2). The 2-D filter with EF can be characterized by the LSS model

$$\begin{aligned}\tilde{\boldsymbol{x}}_{11}(i,j) &= \boldsymbol{A}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,j)]+\boldsymbol{b}u(i,j)+\boldsymbol{D}\boldsymbol{e}(i,j)\\ \tilde{y}(i,j) &= \boldsymbol{c}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,j)]+du(i,j)\end{aligned}\tag{4}$$

where $\boldsymbol{D}$ is referred to as an *EF matrix*.

Subtracting (4) from (1) yields

$$\begin{aligned}\Delta\boldsymbol{x}_{11}(i,j) &= \boldsymbol{A}\Delta\boldsymbol{x}(i,j)+(\boldsymbol{A}-\boldsymbol{D})\boldsymbol{e}(i,j)\\ \Delta y(i,j) &= \boldsymbol{c}\Delta\boldsymbol{x}(i,j)+\boldsymbol{c}\boldsymbol{e}(i,j)\end{aligned}\tag{5}$$

where

$$\begin{aligned}\Delta\boldsymbol{x}(i,j) &= \boldsymbol{x}(i,j)-\tilde{\boldsymbol{x}}(i,j)\\ \Delta\boldsymbol{x}_{11}(i,j) &= \boldsymbol{x}_{11}(i,j)-\tilde{\boldsymbol{x}}_{11}(i,j)\\ \Delta y(i,j) &= y(i,j)-\tilde{y}(i,j).\end{aligned}$$

By taking the 2-D z-transform on both sides of (5) and setting $\Delta\boldsymbol{x}^h(0,j)=\boldsymbol{0}$ for $j=0,1,\ldots$, and $\Delta\boldsymbol{x}^v(i,0)=\boldsymbol{0}$ for $i=0,1,\ldots$, we obtain

$$\begin{aligned}\Delta Y(z_1,z_2) &= \boldsymbol{G}_D(z_1,z_2)\boldsymbol{E}(z_1,z_2)\\ \boldsymbol{G}_D(z_1,z_2) &= \boldsymbol{c}(\boldsymbol{Z}-\boldsymbol{A})^{-1}(\boldsymbol{A}-\boldsymbol{D})+\boldsymbol{c}\end{aligned}\tag{6}$$

where $\boldsymbol{Z}=z_1\boldsymbol{I}_m\oplus z_2\boldsymbol{I}_n$. Here, $\Delta Y(z_1,z_2)$ and $\boldsymbol{E}(z_1,z_2)$ represent the 2-D z-transform of $\Delta y(i,j)$ and $\boldsymbol{e}(i,j)$, respectively, and $\boldsymbol{G}_D(z_1,z_2)$ is the 2-D transfer function from the quantization error $\boldsymbol{e}(i,j)$ to the filter output $\Delta y(i,j)$.

The noise gain is defined as $I(\boldsymbol{D})=\sigma_{\text{out}}^2/\sigma^2$, where σ_{out}^2 denotes noise variance at the filter output and can be evaluated as

$$\begin{aligned}I(\boldsymbol{D}) &= \frac{1}{(2\pi j)^2}\oint_{\Gamma_1}\oint_{\Gamma_2}\boldsymbol{G}_D(z_1,z_2)\boldsymbol{G}_D^*(z_1,z_2)\frac{dz_1dz_2}{z_1z_2}\\ &= \text{tr}[\boldsymbol{W}_D]\end{aligned}\tag{7}$$

where $\Gamma_i=\{z_i:|z_i|=1\}$ for $i=1,2$, and

$$\boldsymbol{W}_D=\frac{1}{(2\pi j)^2}\oint_{\Gamma_1}\oint_{\Gamma_2}\boldsymbol{G}_D^*(z_1,z_2)\boldsymbol{G}_D(z_1,z_2)\frac{dz_1dz_2}{z_1z_2}.$$

By applying the 2-D Cauchy integral theorem, the matrix $\boldsymbol{W}_D$ defined in (7) can be expressed in closed-form as

$$\boldsymbol{W}_D=(\boldsymbol{A}-\boldsymbol{D})^T\boldsymbol{W}_o(\boldsymbol{A}-\boldsymbol{D})+\boldsymbol{c}^T\boldsymbol{c}\tag{8}$$

where $\boldsymbol{W}_o$ is called the local observability Gramian of the 2-D filter and is defined by

$$\begin{aligned}\boldsymbol{W}_o &= \frac{1}{(2\pi j)^2}\oint_{\Gamma_1}\oint_{\Gamma_2}(\boldsymbol{Z}^*-\boldsymbol{A}^T)^{-1}\boldsymbol{c}^T\boldsymbol{c}(\boldsymbol{Z}-\boldsymbol{A})^{-1}\frac{dz_1dz_2}{z_1z_2}\\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{g}(i,j)^T\boldsymbol{g}(i,j)\end{aligned}\tag{9}$$

with

$$\begin{aligned}\boldsymbol{g}(i,j) &= \boldsymbol{c}\boldsymbol{A}^{(i-1,j)}\begin{bmatrix}\boldsymbol{I}_m & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}+\boldsymbol{c}\boldsymbol{A}^{(i,j-1)}\begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}_n\end{bmatrix}\\ \boldsymbol{A}^{(1,0)} &= \begin{bmatrix}\boldsymbol{I}_m & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{A},\quad \boldsymbol{A}^{(0,1)}=\begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}_n\end{bmatrix}\boldsymbol{A}\\ \boldsymbol{A}^{(0,0)} &= \boldsymbol{I}_{m+n},\quad \boldsymbol{A}^{(-i,j)}=\boldsymbol{0}\quad (i\geq 1)\\ \boldsymbol{A}^{(i,-j)} &= \boldsymbol{0}\quad (j\geq 1)\\ \boldsymbol{A}^{(i,j)} &= \boldsymbol{A}^{(1,0)}\boldsymbol{A}^{(i-1,j)}+\boldsymbol{A}^{(0,1)}\boldsymbol{A}^{(i,j-1)}\\ &= \boldsymbol{A}^{(i-1,j)}\boldsymbol{A}^{(1,0)}+\boldsymbol{A}^{(i,j-1)}\boldsymbol{A}^{(0,1)}\\ &\qquad (i,j)>(0,0)\end{aligned}\tag{10}$$

and the partial ordering for integer pairs (i,j) used in [27, p. 2].

We remark that matrix $\boldsymbol{W}_o$ is referred to as the *unit noise matrix* for the 2-D filter, and $\boldsymbol{W}_D$ can be viewed as the unit noise matrix for the 2-D filter with EF specified by matrix $\boldsymbol{D}$.

From (10), it follows that

$$\begin{aligned}\boldsymbol{g}(i,j)\boldsymbol{A} &= \boldsymbol{c}\boldsymbol{A}^{(i-1,j)}\begin{bmatrix}\boldsymbol{I}_m & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{A} + \boldsymbol{c}\boldsymbol{A}^{(i,j-1)}\begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}_n\end{bmatrix}\boldsymbol{A}\\ &= \boldsymbol{c}\boldsymbol{A}^{(i-1,j)}\boldsymbol{A}^{(1,0)} + \boldsymbol{c}\boldsymbol{A}^{(i,j-1)}\boldsymbol{A}^{(0,1)}\\ &= \begin{cases}\boldsymbol{c}\boldsymbol{A}^{(i,j)}, & (i,j) > (0,0)\\ \boldsymbol{0}, & (i,j) = (0,0)\end{cases}\end{aligned}\tag{11}$$

which leads to

$$\begin{aligned}&\boldsymbol{c}^T\boldsymbol{c} + \boldsymbol{A}^T\boldsymbol{W}_o\boldsymbol{A}\\ &\quad= \boldsymbol{c}^T\boldsymbol{c} + \boldsymbol{A}^T\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{g}(i,j)^T\boldsymbol{g}(i,j)\boldsymbol{A}\\ &\quad= \boldsymbol{c}^T\boldsymbol{c} + \sum_{i=1}^{\infty}\boldsymbol{A}^{(i,0)T}\boldsymbol{c}^T\boldsymbol{c}\boldsymbol{A}^{(i,0)}\\ &\qquad+ \sum_{j=1}^{\infty}\boldsymbol{A}^{(0,j)T}\boldsymbol{c}^T\boldsymbol{c}\boldsymbol{A}^{(0,j)} + \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\boldsymbol{A}^{(i,j)T}\boldsymbol{c}^T\boldsymbol{c}\boldsymbol{A}^{(i,j)}\\ &\quad= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{A}^{(i,j)T}\boldsymbol{c}^T\boldsymbol{c}\boldsymbol{A}^{(i,j)}.\end{aligned}\tag{12}$$

By comparing matrix $\boldsymbol{W}_o$ in (9) with (12), we obtain the relations [29]

$$\begin{aligned}\boldsymbol{W}_{o1} &= [\,\boldsymbol{I}_m \quad \boldsymbol{0}\,]\,[\boldsymbol{A}^T\boldsymbol{W}_o\boldsymbol{A} + \boldsymbol{c}^T\boldsymbol{c}]\begin{bmatrix}\boldsymbol{I}_m\\ \boldsymbol{0}\end{bmatrix}\\ \boldsymbol{W}_{o4} &= [\,\boldsymbol{0} \quad \boldsymbol{I}_n\,]\,[\boldsymbol{A}^T\boldsymbol{W}_o\boldsymbol{A} + \boldsymbol{c}^T\boldsymbol{c}]\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{I}_n\end{bmatrix}\end{aligned}\tag{13}$$

where

$$\boldsymbol{W}_o = \begin{bmatrix}\boldsymbol{W}_{o1} & \boldsymbol{W}_{o2}\\ \boldsymbol{W}_{o3} & \boldsymbol{W}_{o4}\end{bmatrix}.$$

Therefore, if there is no EF in the 2-D filter, then the noise gain $I(\boldsymbol{D})$ with $\boldsymbol{D} = \boldsymbol{0}$ becomes

$$\begin{aligned}I(\boldsymbol{0}) &= \mathrm{tr}[\boldsymbol{A}^T\boldsymbol{W}_o\boldsymbol{A} + \boldsymbol{c}^T\boldsymbol{c}]\\ &= \mathrm{tr}[\boldsymbol{W}_o].\end{aligned}\tag{14}$$

The l_2-norm dynamic-range scaling constraints on the local state vector involves the local controllability Gramian of the 2-D filter, which is defined by

$$\begin{aligned}\boldsymbol{K}_c &= \frac{1}{(2\pi j)^2}\oint_{\Gamma_1}\oint_{\Gamma_2}(\boldsymbol{Z}-\boldsymbol{A})^{-1}\boldsymbol{b}\boldsymbol{b}^T(\boldsymbol{Z}^*-\boldsymbol{A}^T)^{-1}\frac{dz_1dz_2}{z_1z_2}\\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{f}(i,j)\boldsymbol{f}(i,j)^T\end{aligned}\tag{15}$$

where

$$\boldsymbol{f}(i,j) = \boldsymbol{A}^{(i-1,j)}\begin{bmatrix}\boldsymbol{b}_1\\ \boldsymbol{0}\end{bmatrix} + \boldsymbol{A}^{(i,j-1)}\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{b}_2\end{bmatrix}.$$

A different yet equivalent state-space description of (1)—$(\bar{\boldsymbol{A}},\bar{\boldsymbol{b}},\bar{\boldsymbol{c}},d)_{m+n}$—can be obtained via a coordinate transformation $\bar{\boldsymbol{x}}(i,j) = \boldsymbol{T}^{-1}\boldsymbol{x}(i,j)$ with $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$, where

$$\bar{\boldsymbol{A}} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T},\quad \bar{\boldsymbol{b}} = \boldsymbol{T}^{-1}\boldsymbol{b},\quad \bar{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T}.\tag{16}$$

Accordingly, the local observability and local controllability Gramians for $(\bar{\boldsymbol{A}},\bar{\boldsymbol{b}},\bar{\boldsymbol{c}},d)_n$ become

$$\bar{\boldsymbol{W}}_o = \boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{T},\quad \bar{\boldsymbol{K}}_c = \boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T}\tag{17}$$

respectively. If the l_2-norm dynamic-range scaling constraints are imposed on the local state vector $\bar{\boldsymbol{x}}(i,j)$, i.e.,

$$(\bar{\boldsymbol{K}}_c)_{ii} = (\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T})_{ii} = 1,\quad i = 1,2,\ldots,m+n\tag{18}$$

then it can be shown that [6], [7]

$$\min_{\boldsymbol{T}}\mathrm{tr}[\bar{\boldsymbol{W}}_o] = \frac{1}{m}\left(\sum_{i=1}^{m}\sigma_{1i}\right)^2 + \frac{1}{n}\left(\sum_{i=1}^{n}\sigma_{4i}\right)^2\tag{19}$$

where σ_{1i}^2 for $i = 1,2,\ldots,m$ and σ_{4i}^2 for $i = 1,2,\ldots,n$ are the eigenvalues of matrices $\boldsymbol{K}_{c1}\boldsymbol{W}_{o1}$ and $\boldsymbol{K}_{c4}\boldsymbol{W}_{o4}$, respectively, and

$$\boldsymbol{K}_c = \begin{bmatrix}\boldsymbol{K}_{c1} & \boldsymbol{K}_{c2}\\ \boldsymbol{K}_{c3} & \boldsymbol{K}_{c4}\end{bmatrix}.$$

The state-space realization satisfying (18) and (19) is called the *optimal realization* (which is sometimes also referred to as the *optimal filter structure*). A method for constructing such a filter structure was proposed in [6] and [7].

If the coordinate transformation for the LSS model in (1) is taken into account, then the 2-D filter with EF can be characterized by

$$\begin{aligned}\tilde{\boldsymbol{x}}_{11}(i,j) &= \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,j)] + \boldsymbol{T}^{-1}\boldsymbol{b}u(i,j) + \boldsymbol{D}\boldsymbol{e}(i,j)\\ \tilde{y}(i,j) &= \boldsymbol{c}\boldsymbol{T}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,j)] + du(i,j)\end{aligned}\tag{20}$$

which corresponds to (4) in the original realization. In this case, the noise gain $I(\boldsymbol{D},\boldsymbol{T})$ becomes

$$\begin{aligned}I(\boldsymbol{D},\boldsymbol{T}) = \mathrm{tr}[(\boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}-\boldsymbol{D})^T\boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{T}(\boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}-\boldsymbol{D})]&\\ + \mathrm{tr}[\boldsymbol{T}^T\boldsymbol{c}^T\boldsymbol{c}\boldsymbol{T}].&\end{aligned}\tag{21}$$

Then, the problem is now formulated as follows. For given $\boldsymbol{A},\boldsymbol{b}$ and $\boldsymbol{c}$ (and therefore, $\boldsymbol{W}_o$ and $\boldsymbol{K}_c$), obtain matrices $\boldsymbol{D}$ and $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ that minimize (21) subject to the constraints in (18).

III. DETERMINATION OF OPTIMAL ERROR FEEDBACK MATRICES

In this section, suppose that the LSS model in (1) is expressed by the optimal realization, after choosing an appropriate coordinate transformation matrix $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ that satisfies (18) and (19) simultaneously. Then, closed-form formulas for determining the optimal full-scale, block-diagonal, diagonal, and scalar EF matrix $\boldsymbol{D}$ to minimize $I(\boldsymbol{D}) = \mathrm{tr}[\boldsymbol{W}_D]$ for a given 2-D state-space digital filter will be derived. It is noted that the optimal full-scale EF matrix is often too costly because it requires as many as $(m+n)^2$ explicit multiplications. The costs can be reduced, e.g., by constraining the EF matrix to be block-diagonal or diagonal, which reduces the number of distinct coefficients to $m^2 + n^2$ or $m + n$.

1) Case 1 $\boldsymbol{D}$ Is a General Matrix: Substituting (8) into (7), we obtain

$$\begin{aligned} I(\boldsymbol{D}) &= \mathrm{tr}[\boldsymbol{c}^T\boldsymbol{c} + (\boldsymbol{A}-\boldsymbol{D})^T\boldsymbol{W}_o(\boldsymbol{A}-\boldsymbol{D})] \\ &= \mathrm{tr}[\boldsymbol{c}^T\boldsymbol{c} + \boldsymbol{A}^T\boldsymbol{W}_o\boldsymbol{A}] + \mathrm{tr}[\boldsymbol{D}^T\boldsymbol{W}_o\boldsymbol{D}] - 2\mathrm{tr}[\boldsymbol{D}^T\boldsymbol{W}_o\boldsymbol{A}] \\ &= \mathrm{tr}[\boldsymbol{W}_o] + \mathrm{tr}[\boldsymbol{D}^T\boldsymbol{W}_o\boldsymbol{D}] - 2\,\mathrm{tr}[\boldsymbol{D}^T\boldsymbol{W}_o\boldsymbol{A}]. \end{aligned} \tag{22}$$

Differentiating (22) with respect to the EF matrix $\boldsymbol{D}$ yields

$$\frac{\partial I(\boldsymbol{D})}{\partial \boldsymbol{D}} = 2\boldsymbol{W}_o(\boldsymbol{D}-\boldsymbol{A}). \tag{23}$$

By choosing the EF matrix as $\boldsymbol{D} = \boldsymbol{A}$, the noise gain $I(\boldsymbol{D})$ in (22) achieves its minimum value

$$\begin{aligned} I_{\min}(\boldsymbol{D}) &= \mathrm{tr}[\boldsymbol{W}_o] - \mathrm{tr}[\boldsymbol{A}^T\boldsymbol{W}_o\boldsymbol{A}] \\ &= \mathrm{tr}[\boldsymbol{c}^T\boldsymbol{c}]. \end{aligned} \tag{24}$$

2) Case 2 $\boldsymbol{D}$ Is a Block-Diagonal Matrix: In this case, matrix $\boldsymbol{D}$ assumes the form

$$\boldsymbol{D} = \boldsymbol{D}_1 \oplus \boldsymbol{D}_4 \tag{25}$$

where $\boldsymbol{D}_1$ and $\boldsymbol{D}_4$ are $m \times m$ and $n \times n$ matrices, respectively, which leads (22) to

$$\begin{aligned} I(\boldsymbol{D}) &= \mathrm{tr}[\boldsymbol{W}_o] + \mathrm{tr}\left[\boldsymbol{D}_1^T\boldsymbol{W}_{o1}\boldsymbol{D}_1\right] + \mathrm{tr}\left[\boldsymbol{D}_4^T\boldsymbol{W}_{o4}\boldsymbol{D}_4\right] \\ &\quad - 2\mathrm{tr}\left[\boldsymbol{D}_1^T(\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3)\right] \\ &\quad - 2\mathrm{tr}\left[\boldsymbol{D}_4^T(\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4)\right]. \end{aligned} \tag{26}$$

Letting $\partial I(\boldsymbol{D})/\partial \boldsymbol{D}_1 = \boldsymbol{0}$ and $\partial I(\boldsymbol{D})/\partial \boldsymbol{D}_4 = \boldsymbol{0}$, it follows that

$$\begin{aligned} \boldsymbol{D}_1 &= \boldsymbol{A}_1 + \boldsymbol{W}_{o1}^{-1}\boldsymbol{W}_{o2}\boldsymbol{A}_3 \\ \boldsymbol{D}_4 &= \boldsymbol{A}_4 + \boldsymbol{W}_{o4}^{-1}\boldsymbol{W}_{o3}\boldsymbol{A}_2. \end{aligned} \tag{27}$$

By substituting (27) into (26), we obtain the minimum value of the noise gain $I(\boldsymbol{D})$ as

$$\begin{aligned} I_{\min}(\boldsymbol{D}) = \mathrm{tr}[\boldsymbol{W}_o] &- \mathrm{tr}\left[\boldsymbol{D}_1^T(\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3)\right] \\ &- \mathrm{tr}\left[\boldsymbol{D}_4^T(\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4)\right]. \end{aligned} \tag{28}$$

3) Case 3 $\boldsymbol{D}$ Is a Diagonal Matrix: In this case, matrix $\boldsymbol{D}$ assumes the form

$$\begin{aligned} \boldsymbol{D}_1 &= \mathrm{diag}\{\alpha_1, \alpha_2, \ldots, \alpha_m\} \\ \boldsymbol{D}_4 &= \mathrm{diag}\{\beta_1, \beta_2, \ldots, \beta_n\} \end{aligned} \tag{29}$$

which leads (26) to

$$\begin{aligned} I(\boldsymbol{D}) &- \mathrm{tr}[\boldsymbol{W}_o] \\ &= \sum_{i=1}^{m}(\boldsymbol{W}_{o1})_{ii}\alpha_i\left(\alpha_i - 2\frac{(\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3)_{ii}}{(\boldsymbol{W}_{o1})_{ii}}\right) \\ &\quad + \sum_{i=1}^{n}(\boldsymbol{W}_{o4})_{ii}\beta_i\left(\beta_i - 2\frac{(\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4)_{ii}}{(\boldsymbol{W}_{o4})_{ii}}\right). \end{aligned} \tag{30}$$

This implies that if α_i's and β_i's satisfy

$$\begin{aligned} \alpha_i\left(\alpha_i - 2\frac{(\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3)_{ii}}{(\boldsymbol{W}_{o1})_{ii}}\right) < 0, \quad i = 1, 2, \ldots, m \\ \beta_i\left(\beta_i - 2\frac{(\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4)_{ii}}{(\boldsymbol{W}_{o4})_{ii}}\right) < 0, \quad i = 1, 2, \ldots, n \end{aligned} \tag{31}$$

then the right-hand side of (30) becomes negative, that is, $I(\boldsymbol{D}) = \mathrm{tr}[\boldsymbol{W}_D] < \mathrm{tr}[\boldsymbol{W}_o]$ holds. Letting $\partial I(\boldsymbol{D})/\partial\alpha_i = 0$ and $\partial I(\boldsymbol{D})/\partial\beta_i = 0$, we obtain $\boldsymbol{D}_1 = \mathrm{diag}\{\alpha_1, \alpha_2, \ldots, \alpha_m\}$ and $\boldsymbol{D}_4 = \mathrm{diag}\{\beta_1, \beta_2, \ldots, \beta_n\}$ with

$$\begin{aligned} \alpha_i &= \frac{(\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3)_{ii}}{(\boldsymbol{W}_{o1})_{ii}}, \quad i = 1, 2, \ldots, m \\ \beta_i &= \frac{(\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4)_{ii}}{(\boldsymbol{W}_{o4})_{ii}}, \quad i = 1, 2, \ldots, n \end{aligned} \tag{32}$$

where $I(\boldsymbol{D})$ achieves its minimum as

$$\begin{aligned} I_{\min}(\boldsymbol{D}) = \mathrm{tr}[\boldsymbol{W}_o] &- \sum_{i=1}^{m}\frac{(\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3)_{ii}^2}{(\boldsymbol{W}_{o1})_{ii}} \\ &- \sum_{i=1}^{n}\frac{(\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4)_{ii}^2}{(\boldsymbol{W}_{o4})_{ii}}. \end{aligned} \tag{33}$$

4) Case 4 $\boldsymbol{D}_1$ and $\boldsymbol{D}_4$ Are Scalar Matrices $\alpha\boldsymbol{I}_m$ and $\beta\boldsymbol{I}_n$: If $\boldsymbol{D}_1 = \alpha\boldsymbol{I}_m$ and $\boldsymbol{D}_4 = \beta\boldsymbol{I}_n$ with scalars α and β, then (30) becomes

$$\begin{aligned} I(\boldsymbol{D}) &- \mathrm{tr}[\boldsymbol{W}_o] \\ &= \mathrm{tr}[\boldsymbol{W}_{o1}]\alpha\left(\alpha - 2\frac{\mathrm{tr}[\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3]}{\mathrm{tr}[\boldsymbol{W}_{o1}]}\right) \\ &\quad + \mathrm{tr}[\boldsymbol{W}_{o4}]\beta\left(\beta - 2\frac{\mathrm{tr}[\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4]}{\mathrm{tr}[\boldsymbol{W}_{o4}]}\right). \end{aligned} \tag{34}$$

Hence, if α and β satisfy

$$\begin{aligned} \alpha\left(\alpha - 2\frac{\mathrm{tr}[\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3]}{\mathrm{tr}[\boldsymbol{W}_{o1}]}\right) < 0 \\ \beta\left(\beta - 2\frac{\mathrm{tr}[\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4]}{\mathrm{tr}[\boldsymbol{W}_{o4}]}\right) < 0 \end{aligned} \tag{35}$$

then the right-hand side of (34) is negative, that is, $I(\boldsymbol{D}) = \mathrm{tr}[\boldsymbol{W}_D] < \mathrm{tr}[\boldsymbol{W}_o]$ holds. Moreover, from $\partial I(\boldsymbol{D})/\partial\alpha = 0$ and $\partial I(\boldsymbol{D})/\partial\beta = 0$, it follows that the values of α and β that minimize $I(\boldsymbol{D})$ are given by

$$\begin{aligned} \alpha &= \frac{\mathrm{tr}[\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3]}{\mathrm{tr}[\boldsymbol{W}_{o1}]} \\ \beta &= \frac{\mathrm{tr}[\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4]}{\mathrm{tr}[\boldsymbol{W}_{o4}]} \end{aligned} \tag{36}$$

which lead (34) to

$$\begin{aligned} I_{\min}(\boldsymbol{D}) = \mathrm{tr}[\boldsymbol{W}_o] &- \frac{(\mathrm{tr}[\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3])^2}{\mathrm{tr}[\boldsymbol{W}_{o1}]} \\ &- \frac{(\mathrm{tr}[\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4])^2}{\mathrm{tr}[\boldsymbol{W}_{o4}]}. \end{aligned} \tag{37}$$

IV. Noise Reduction by Joint Optimization of Error Feedback and Coordinate Transformation

First, the joint optimization of scalar EF matrices $\boldsymbol{D}_1 = \alpha \boldsymbol{I}_m$ and $\boldsymbol{D}_4 = \beta \boldsymbol{I}_n$ and coordinate transformation matrices $\boldsymbol{T}_1$ and $\boldsymbol{T}_4$ will be investigated for roundoff noise minimization under l_2-norm dynamic-range scaling constraints. Such constraints on the matrix $\boldsymbol{D}$ are introduced in order to guarantee $\boldsymbol{T}^{-1}\boldsymbol{D}\boldsymbol{T} = \boldsymbol{D}$ for every block-diagonal matrix $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$. Then, (21) is written as

$$\begin{aligned} I(\boldsymbol{D},\boldsymbol{T}) &= \mathrm{tr}[\boldsymbol{T}^T((\boldsymbol{A}-\boldsymbol{D})^T\boldsymbol{W}_o(\boldsymbol{A}-\boldsymbol{D}) + \boldsymbol{c}^T\boldsymbol{c})\boldsymbol{T}] \\ &= \mathrm{tr}[((\boldsymbol{A}-\boldsymbol{D})^T\boldsymbol{W}_o(\boldsymbol{A}-\boldsymbol{D}) + \boldsymbol{c}^T\boldsymbol{c})\boldsymbol{P}] \end{aligned} \tag{38}$$

where $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$, that is, $\boldsymbol{P}_i = \boldsymbol{T}_i\boldsymbol{T}_i^T$ for $i = 1, 4$. If the coordinate transformation for the LSS model in (1) is taken into account, then (36) is changed to

$$\begin{aligned} \alpha &= \frac{\mathrm{tr}[(\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3)\boldsymbol{P}_1]}{\mathrm{tr}[\boldsymbol{W}_{o1}\boldsymbol{P}_1]} \\ \beta &= \frac{\mathrm{tr}[(\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4)\boldsymbol{P}_4]}{\mathrm{tr}[\boldsymbol{W}_{o4}\boldsymbol{P}_4]}. \end{aligned} \tag{39}$$

Equations (38) and (39) imply that for fixed α and β, matrix $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ can be optimized to minimize $I(\boldsymbol{D},\boldsymbol{T})$ subject to the scaling constraints in (18) and vice versa. The proposed joint optimization will be performed in an iterative manner.

First, scalars α and β can be derived from (39) when the initial $\boldsymbol{P}$, say $\boldsymbol{P}_0$, is given. In what follows, let the unit noise matrix $\boldsymbol{W}_D$ in (8) with $\boldsymbol{D} = \alpha\boldsymbol{I}_m \oplus \beta\boldsymbol{I}_n$ be denoted by

$$\boldsymbol{W}_D = \begin{bmatrix} \boldsymbol{W}_{1\alpha} & \boldsymbol{W}_{\alpha\beta}^T \\ \boldsymbol{W}_{\alpha\beta} & \boldsymbol{W}_{4\beta} \end{bmatrix}. \tag{40}$$

Under the joint application of a scalar EF and a coordinate transformation, the noise gain $I(\boldsymbol{D},\boldsymbol{T})$ is given by $\mathrm{tr}[\boldsymbol{T}_1^T\boldsymbol{W}_{1\alpha}\boldsymbol{T}_1] + \mathrm{tr}[\boldsymbol{T}_4^T\boldsymbol{W}_{4\beta}\boldsymbol{T}_4]$. In order to minimize $I(\boldsymbol{D},\boldsymbol{T})$ (with α and β temporarily fixed) over an $m \times m$ nonsingular matrix $\boldsymbol{T}_1$ and an $n \times n$ nonsingular matrix $\boldsymbol{T}_4$ subject to the scaling constraints in (18), we define the Lagrange function

$$\begin{aligned} J(\alpha,\beta,\boldsymbol{P}) &= \mathrm{tr}[\boldsymbol{W}_{1\alpha}\boldsymbol{P}_1] + \lambda_1\left(\mathrm{tr}\left[\boldsymbol{K}_{c1}\boldsymbol{P}_1^{-1}\right] - m\right) \\ &\quad + \mathrm{tr}[\boldsymbol{W}_{4\beta}\boldsymbol{P}_4] + \lambda_4\left(\mathrm{tr}\left[\boldsymbol{K}_{c4}\boldsymbol{P}_4^{-1}\right] - n\right) \end{aligned} \tag{41}$$

where λ_1 and λ_4 are Lagrange multipliers. By using the formula for evaluating matrix gradient [30, p. 275]

$$\partial(\mathrm{tr}[\boldsymbol{M}\boldsymbol{X}^{-1}])/\partial\boldsymbol{X} = -[\boldsymbol{X}^{-1}\boldsymbol{M}\boldsymbol{X}^{-1}]^T$$

we compute

$$\begin{aligned} \frac{\partial J(\alpha,\beta,\boldsymbol{P})}{\partial\boldsymbol{P}_1} &= \boldsymbol{W}_{1\alpha} - \lambda_1\boldsymbol{P}_1^{-1}\boldsymbol{K}_{c1}\boldsymbol{P}_1^{-1} \\ \frac{\partial J(\alpha,\beta,\boldsymbol{P})}{\partial\boldsymbol{P}_4} &= \boldsymbol{W}_{4\beta} - \lambda_4\boldsymbol{P}_4^{-1}\boldsymbol{K}_{c4}\boldsymbol{P}_4^{-1} \\ \frac{\partial J(\alpha,\beta,\boldsymbol{P})}{\partial\lambda_1} &= \mathrm{tr}\left[\boldsymbol{K}_{c1}\boldsymbol{P}_1^{-1}\right] - m \\ \frac{\partial J(\alpha,\beta,\boldsymbol{P})}{\partial\lambda_4} &= \mathrm{tr}\left[\boldsymbol{K}_{c4}\boldsymbol{P}_4^{-1}\right] - n. \end{aligned} \tag{42}$$

Letting $\partial J(\alpha,\beta,\boldsymbol{P})/\partial\boldsymbol{P}_i = \boldsymbol{0}$ and $\partial J(\alpha,\beta,\boldsymbol{P})/\partial\lambda_i = 0$ for $i = 1, 4$, it is derived that

$$\begin{aligned} \boldsymbol{P}_1\boldsymbol{W}_{1\alpha}\boldsymbol{P}_1 &= \lambda_1\boldsymbol{K}_{c1}, \quad \mathrm{tr}\left[\boldsymbol{K}_{c1}\boldsymbol{P}_1^{-1}\right] = m \\ \boldsymbol{P}_4\boldsymbol{W}_{4\beta}\boldsymbol{P}_4 &= \lambda_4\boldsymbol{K}_{c4}, \quad \mathrm{tr}\left[\boldsymbol{K}_{c4}\boldsymbol{P}_4^{-1}\right] = n. \end{aligned} \tag{43}$$

Note that if matrices $\boldsymbol{W} > 0$ and $\boldsymbol{M} \geq 0$ are symmetric, then the matrix equation $\boldsymbol{P}\boldsymbol{W}\boldsymbol{P} = \boldsymbol{M}$ has the unique solution [31]

$$\boldsymbol{P} = \boldsymbol{W}^{-\frac{1}{2}}[\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{M}\boldsymbol{W}^{\frac{1}{2}}]^{\frac{1}{2}}\boldsymbol{W}^{-\frac{1}{2}}.$$

Then, it follows from (43) that

$$\begin{aligned} \boldsymbol{P}_1 &= \sqrt{\lambda_1}\boldsymbol{W}_{1\alpha}^{-\frac{1}{2}}\left[\boldsymbol{W}_{1\alpha}^{\frac{1}{2}}\boldsymbol{K}_{c1}\boldsymbol{W}_{1\alpha}^{\frac{1}{2}}\right]^{\frac{1}{2}}\boldsymbol{W}_{1\alpha}^{-\frac{1}{2}} \\ \boldsymbol{P}_4 &= \sqrt{\lambda_4}\boldsymbol{W}_{4\beta}^{-\frac{1}{2}}\left[\boldsymbol{W}_{4\beta}^{\frac{1}{2}}\boldsymbol{K}_{c4}\boldsymbol{W}_{4\beta}^{\frac{1}{2}}\right]^{\frac{1}{2}}\boldsymbol{W}_{4\beta}^{-\frac{1}{2}} \\ \frac{1}{\sqrt{\lambda_1}}\mathrm{tr}[\boldsymbol{K}_{c1}\boldsymbol{W}_{1\alpha}]^{\frac{1}{2}} &= \frac{1}{\sqrt{\lambda_1}}\left(\sum_{i=1}^{m}\mu_i\right) = m \\ \frac{1}{\sqrt{\lambda_4}}\mathrm{tr}[\boldsymbol{K}_{c4}\boldsymbol{W}_{4\beta}]^{\frac{1}{2}} &= \frac{1}{\sqrt{\lambda_4}}\left(\sum_{i=1}^{n}\nu_i\right) = n \end{aligned} \tag{44}$$

where μ_i^2 for $i = 1, 2, \ldots, m$ and ν_i^2 for $i = 1, 2, \ldots, n$ are the eigenvalues of $\boldsymbol{K}_{c1}\boldsymbol{W}_{1\alpha}$ and $\boldsymbol{K}_{c4}\boldsymbol{W}_{4\beta}$, respectively. Therefore, we obtain

$$\begin{aligned} \boldsymbol{P}_1 &= \frac{1}{m}\left(\sum_{i=1}^{m}\mu_i\right)\boldsymbol{W}_{1\alpha}^{-\frac{1}{2}}\left[\boldsymbol{W}_{1\alpha}^{\frac{1}{2}}\boldsymbol{K}_{c1}\boldsymbol{W}_{1\alpha}^{\frac{1}{2}}\right]^{\frac{1}{2}}\boldsymbol{W}_{1\alpha}^{-\frac{1}{2}} \\ \boldsymbol{P}_4 &= \frac{1}{n}\left(\sum_{i=1}^{n}\nu_i\right)\boldsymbol{W}_{4\beta}^{-\frac{1}{2}}\left[\boldsymbol{W}_{4\beta}^{\frac{1}{2}}\boldsymbol{K}_{c4}\boldsymbol{W}_{4\beta}^{\frac{1}{2}}\right]^{\frac{1}{2}}\boldsymbol{W}_{4\beta}^{-\frac{1}{2}}. \end{aligned} \tag{45}$$

Substituting (45) into (41) yields the minimum value of $J(\alpha,\beta,\boldsymbol{P})$ for fixed α and β as

$$\min_{\boldsymbol{P}} J(\alpha,\beta,\boldsymbol{P}) = \frac{1}{m}\left(\sum_{i=1}^{m}\mu_i\right)^2 + \frac{1}{n}\left(\sum_{i=1}^{n}\nu_i\right)^2. \tag{46}$$

Having obtained matrix $\boldsymbol{P} = \boldsymbol{P}_1 \oplus \boldsymbol{P}_4$, the improved values of scalars α and β can be obtained using (39). This iterative procedure for minimizing the roundoff noise under the scaling constraints in (18) with respect to scalar parameters α and β as well as an $(m+n) \times (m+n)$ symmetric positive-definite $\boldsymbol{P} = \boldsymbol{P}_1 \oplus \boldsymbol{P}_4$ can be summarized as follows.

1) Set $i = 1$, and

$$\boldsymbol{P}(0) = \mathrm{diag}\{(\boldsymbol{K}_c)_{11}, (\boldsymbol{K}_c)_{22}, \ldots, (\boldsymbol{K}_c)_{m+n,m+n}\}.$$

2) Compute scalars $\alpha(i)$ and $\beta(i)$ using

$$\begin{aligned} \alpha(i) &= \frac{\mathrm{tr}[(\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3)\boldsymbol{P}_1(i-1)]}{\mathrm{tr}[\boldsymbol{W}_{o1}\boldsymbol{P}_1(i-1)]} \\ \beta(i) &= \frac{\mathrm{tr}[(\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4)\boldsymbol{P}_4(i-1)]}{\mathrm{tr}[\boldsymbol{W}_{o4}\boldsymbol{P}_4(i-1)]}. \end{aligned}$$

3) Compute

$$\begin{aligned} I_{\min}(\alpha(i)\boldsymbol{I}_m \oplus \beta(i)\boldsymbol{I}_n) &= (1-\alpha(i)^2)\mathrm{tr}[\boldsymbol{W}_{o1}\boldsymbol{P}_1(i-1)] \\ &\quad + (1-\beta(i)^2)\mathrm{tr}[\boldsymbol{W}_{o4}\boldsymbol{P}_4(i-1)]. \end{aligned}$$

4) Replace $\boldsymbol{W}_{1\alpha}$ and $\boldsymbol{W}_{4\beta}$ by $\boldsymbol{W}_{1\alpha(i)}$ and $\boldsymbol{W}_{4\beta(i)}$ computed using

$$\begin{aligned}\boldsymbol{W}_{1\alpha(i)} &= (1+\alpha(i)^2)\boldsymbol{W}_{o1} - \alpha(i)[(\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{02}\boldsymbol{A}_3)^T \\ &\quad + \boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{02}\boldsymbol{A}_3] \\ \boldsymbol{W}_{4\beta(i)} &= (1+\beta(i)^2)\boldsymbol{W}_{o4} - \beta(i)[(\boldsymbol{W}_{o4}\boldsymbol{A}_4 + \boldsymbol{W}_{03}\boldsymbol{A}_2)^T \\ &\quad + \boldsymbol{W}_{o4}\boldsymbol{A}_4 + \boldsymbol{W}_{03}\boldsymbol{A}_2]\end{aligned}$$

respectively.

5) Derive $\boldsymbol{P} = \boldsymbol{P}_1 \oplus \boldsymbol{P}_4$ from (45), and take the resulting matrix as $\boldsymbol{P}(i) = \boldsymbol{P}_1(i) \oplus \boldsymbol{P}_4(i)$.
6) Compute tr$[\boldsymbol{W}_{1\alpha(i)}\boldsymbol{P}_1(i)]$ + tr$[\boldsymbol{W}_{4\alpha(i)}\boldsymbol{P}_4(i)]$.
7) Update $i := i+1$, and repeat from Step 2) until the change in either $I[\alpha(i)\boldsymbol{I}_m \oplus \beta(i)\boldsymbol{I}_n]$ or tr$[\boldsymbol{W}_{1\alpha(i)}\boldsymbol{P}_1(i)]+$ tr$[\boldsymbol{W}_{4\alpha(i)}\boldsymbol{P}_4(i)]$ becomes insignificant compared with a prescribed tolerance.

We remark that although the objective function involved in the joint optimization is not convex in general and a rigorous mathematical proof of the convergence property is not yet available at present, the above iterative algorithm was applied to quite a number of simulation examples, and fast convergence was observed in all the cases where all the final results were identical for any initial state-space realization. A sample of these examples will be illustrated in the next section.

Suppose the above algorithm converges after N iterations and the optimal coordinate transformation matrix $\boldsymbol{T}(N) = \boldsymbol{T}_1(N) \oplus \boldsymbol{T}_4(N)$ has been computed from the symmetric positive-definite matrix $\boldsymbol{P}(N) = \boldsymbol{P}_1(N) \oplus \boldsymbol{P}_4(N)$ (see the Appendix). Then, following (29)–(32), the diagonal EF matrix $\boldsymbol{D} = \boldsymbol{D}_1 \oplus \boldsymbol{D}_4$ with $\boldsymbol{D}_1 = \text{diag}\{\alpha_1, \alpha_2, \ldots, \alpha_m\}$ and $\boldsymbol{D}_4 = \text{diag}\{\beta_1, \beta_2, \ldots, \beta_n\}$ that minimizes

$$\begin{aligned}I(\boldsymbol{D}) &= \text{tr}[\boldsymbol{T}^T(N)\boldsymbol{W}_o\boldsymbol{T}(N)] \\ &\quad + \text{tr}\left[\boldsymbol{D}_1^2\boldsymbol{T}_1^T(N)\boldsymbol{W}_{o1}\boldsymbol{T}_1(N)\right] \\ &\quad - 2\,\text{tr}\left[\boldsymbol{D}_1\boldsymbol{T}_1^T(N)(\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3)\boldsymbol{T}_1(N)\right] \\ &\quad + \text{tr}\left[\boldsymbol{D}_4^2\boldsymbol{T}_4^T(N)\boldsymbol{W}_{o4}\boldsymbol{T}_4(N)\right] \\ &\quad - 2\,\text{tr}\left[\boldsymbol{D}_4\boldsymbol{T}_4^T(N)(\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4)\boldsymbol{T}_4(N)\right] \end{aligned} \tag{47}$$

is given by

$$\begin{aligned}\alpha_i &= \frac{\left(\boldsymbol{T}_1^T(N)(\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3)\boldsymbol{T}_1(N)\right)_{ii}}{\left(\boldsymbol{T}_1^T(N)\boldsymbol{W}_{o1}\boldsymbol{T}_1(N)\right)_{ii}} \\ &\qquad\qquad i = 1,2,\ldots,m \\ \beta_i &= \frac{\left(\boldsymbol{T}_4^T(N)(\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4)\boldsymbol{T}_4(N)\right)_{ii}}{\left(\boldsymbol{T}_4^T(N)\boldsymbol{W}_{o4}\boldsymbol{T}_4(N)\right)_{ii}} \\ &\qquad\qquad i = 1,2,\ldots,n. \end{aligned} \tag{48}$$

This diagonal EF matrix $\boldsymbol{D} = \boldsymbol{D}_1 \oplus \boldsymbol{D}_4$ leads to further reduction of the noise gain, i.e.,

$$I_{\min}(\boldsymbol{D}) < I_{\min}[\alpha(N)\boldsymbol{I}_m \oplus \beta(N)\boldsymbol{I}_n]. \tag{49}$$

Next, we discuss the joint optimization of a general EF matrix $\boldsymbol{D}$ and a coordinate transformation matrix $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ for roundoff noise minimization under the scaling constraints in (18). In this case, the problem can be reduced as follows: For given $\boldsymbol{A}, \boldsymbol{b}$, and $\boldsymbol{c}$ (and therefore $\boldsymbol{K}_c$ and $\boldsymbol{W}_o$), obtain matrix $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ that minimizes tr$[\boldsymbol{c}^T\boldsymbol{c}\boldsymbol{P}]$ subject to $(\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T})_{ii} = 1$ for $i = 1,2,\ldots,m+n$, where the optimal $\boldsymbol{D}$ can be obtained by $\boldsymbol{D} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}$. For tractability, we consider the minimization of tr$[((1-\mu)\boldsymbol{c}^T\boldsymbol{c} + \mu\boldsymbol{W}_o)\boldsymbol{P}]$ instead of tr$[\boldsymbol{c}^T\boldsymbol{c}\boldsymbol{P}]$, where $0 < \mu \le 1$. In other words, we define the Lagrange function

$$\begin{aligned}J_o(\boldsymbol{P}) &= \text{tr}[\hat{\boldsymbol{W}}_{o1}\boldsymbol{P}_1] + \lambda_1\left(\text{tr}\left[\boldsymbol{K}_{c1}\boldsymbol{P}_1^{-1}\right] - m\right) \\ &\quad + \text{tr}[\hat{\boldsymbol{W}}_{o4}\boldsymbol{P}_4] + \lambda_4\left(\text{tr}\left[\boldsymbol{K}_{c4}\boldsymbol{P}_4^{-1}\right] - n\right)\end{aligned} \tag{50}$$

where λ_1 and λ_4 are Lagrange multipliers, and

$$\begin{aligned}\hat{\boldsymbol{W}}_{o1} &= (1-\mu)\boldsymbol{c}_1^T\boldsymbol{c}_1 + \mu\boldsymbol{W}_{o1} \\ \hat{\boldsymbol{W}}_{o4} &= (1-\mu)\boldsymbol{c}_2^T\boldsymbol{c}_2 + \mu\boldsymbol{W}_{o4}.\end{aligned}$$

Employing steps similar to those used in deriving (45) from (41), we arrive at

$$\begin{aligned}\boldsymbol{P}_1 &= \frac{1}{m}\left(\sum_{i=1}^{m}\hat{\sigma}_{1i}\right)\hat{\boldsymbol{W}}_{o1}^{-\frac{1}{2}}\left[\hat{\boldsymbol{W}}_{o1}^{\frac{1}{2}}\boldsymbol{K}_{c1}\hat{\boldsymbol{W}}_{o1}^{\frac{1}{2}}\right]^{\frac{1}{2}}\hat{\boldsymbol{W}}_{o1}^{-\frac{1}{2}} \\ \boldsymbol{P}_4 &= \frac{1}{n}\left(\sum_{i=1}^{n}\hat{\sigma}_{4i}\right)\hat{\boldsymbol{W}}_{o4}^{-\frac{1}{2}}\left[\hat{\boldsymbol{W}}_{o4}^{\frac{1}{2}}\boldsymbol{K}_{c4}\hat{\boldsymbol{W}}_{o4}^{\frac{1}{2}}\right]^{\frac{1}{2}}\hat{\boldsymbol{W}}_{o4}^{-\frac{1}{2}}\end{aligned} \tag{51}$$

where $\hat{\sigma}_{1i}^2$ for $i = 1,2,\ldots,m$ and $\hat{\sigma}_{4i}^2$ for $i = 1,2,\ldots,n$ are the eigenvalues of $\boldsymbol{K}_{c1}\hat{\boldsymbol{W}}_{o1}$ and $\boldsymbol{K}_{c4}\hat{\boldsymbol{W}}_{o4}$, respectively. Note that matrices $\hat{\boldsymbol{W}}_{o1}$ and $\hat{\boldsymbol{W}}_{o4}$ are symmetric positive-definite, provided that $\mu > 0$. Once $\boldsymbol{P} = \boldsymbol{P}_1 \oplus \boldsymbol{P}_4$ are obtained, the coordinate transformation matrix $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ can be constructed from $\boldsymbol{P}_1 = \boldsymbol{T}_1\boldsymbol{T}_1^T$ and $\boldsymbol{P}_4 = \boldsymbol{T}_4\boldsymbol{T}_4^T$ to satisfy the scaling constraints in (18) (see the Appendix). The noise gain $I(\boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T})$ is then computed by tr$[\boldsymbol{T}^T\boldsymbol{c}^T\boldsymbol{c}\boldsymbol{T}]$.

V. Numerical Example

In this section, we present a numerical example to illustrate the algorithms proposed in Sections III and IV.

Consider a 2-D stable, separately locally controllable, and separately locally observable state-space digital filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{2,2}$ with $d = 0.0$ described by

$$\begin{aligned}&\begin{bmatrix}\boldsymbol{A}_1 & \boldsymbol{A}_2 \\ \boldsymbol{A}_3 & \boldsymbol{A}_4\end{bmatrix} \\ &\quad = \begin{bmatrix}1.888\,990 & -0.912\,190 & -1.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 \\ 0.027\,710 & -0.025\,800 & 1.888\,990 & 1.0 \\ -0.025\,800 & 0.024\,310 & -0.912\,190 & 0.0\end{bmatrix} \\ &[\boldsymbol{b}_1 \quad \boldsymbol{b}_2] \\ &\quad = [0.219\,089 \quad 0.0 \quad -0.028\,889 \quad 0.091\,219]^T \\ &[\boldsymbol{c}_1 \quad \boldsymbol{c}_2] \\ &\quad = [0.028\,889\,0 \quad -0.091\,219 \quad -0.219\,089 \quad 0.0].\end{aligned}$$

If a coordinate transformation matrix $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ is selected as

$$\begin{aligned}\boldsymbol{T}_1 &= \text{diag}\{9.336\,610, 9.336\,609\} \\ \boldsymbol{T}_4 &= \text{diag}\{1.065\,112, 0.986\,652\}\end{aligned}$$

then the above filter satisfies the scaling constraints in (18) and produces tr$[\boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{T}] = 367.508\,947$.

If a coordinate transformation matrix $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ is chosen as

$$\boldsymbol{T}_1 = \begin{bmatrix} 9.544\,965 & 1.373\,341 \\ 9.494\,676 & 3.318\,699 \end{bmatrix}$$
$$\boldsymbol{T}_4 = \begin{bmatrix} 0.329\,402 & -0.942\,406 \\ -0.136\,313 & 0.947\,397 \end{bmatrix}$$

then the above filter is transformed to the optimal realization $(\bar{\boldsymbol{A}}, \bar{\boldsymbol{b}}, \bar{\boldsymbol{c}}, d)_{2,2}$ with minimum roundoff noise subject to the scaling constraints in (18), where we have the first equation at the bottom of the page, whose local controllability and local observability Gramians are written as in the second equation at the bottom of the page, respectively, where the infinite sums in (9) and (15) are calculated by truncation $0 \le i \le 400$ and $0 \le j \le 400$, and the noise gain $I(\boldsymbol{0})$ is given by $\mathrm{tr}[\bar{\boldsymbol{W}}_o] = 13.688\,256$.

Let us now apply the EF described in Section III to the above optimal realization $(\bar{\boldsymbol{A}}, \bar{\boldsymbol{b}}, \bar{\boldsymbol{c}}, d)_{2,2}$. In the case when $\boldsymbol{D}$ is allowed to be a general EF matrix, then (23) suggests that we should choose $\boldsymbol{D} = \bar{\boldsymbol{A}}$, which yields $I_{\min}(\boldsymbol{D}) = 0.465\,549$. If $\boldsymbol{D}$ is constrained to be a block-diagonal EF matrix, then the optimal $\boldsymbol{D} = \boldsymbol{D}_1 \oplus \boldsymbol{D}_4$ is calculated using (27), which gives

$$\boldsymbol{D}_1 = \begin{bmatrix} 0.965\,580 & -0.178\,717 \\ 0.109\,304 & 0.933\,443 \end{bmatrix}$$
$$\boldsymbol{D}_4 = \begin{bmatrix} 0.935\,176 & 0.200\,611 \\ -0.156\,671 & 0.862\,050 \end{bmatrix}$$
$$I(\boldsymbol{D}) = 1.555\,329.$$

If $\boldsymbol{D}$ is constrained to be a diagonal EF matrix, then it can be calculated using (32) as

$$\boldsymbol{D} = \mathrm{diag}\{0.941\,314, 0.973\,118, 0.969\,957, 0.817\,514\}$$

which yields $I_{\min}(\boldsymbol{D}) = 1.908\,903$. If a scalar EF matrix is calculated using (36), then we obtain $\alpha = 0.957\,216$ and $\beta = 0.893\,736$, which yield $I_{\min}(\boldsymbol{D}) = 1.950\,396$.

Now, we apply the iterative optimization procedure described in Section IV to the original realization $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{2,2}$. The proposed algorithm converges after eight iterations to scalars $\alpha = 0.972\,437, \beta = 0.932\,447$, and a transformation matrix $\boldsymbol{T}(8) = \boldsymbol{T}_1(8) \oplus \boldsymbol{T}_4(8)$ with

$$\boldsymbol{T}_1(8) = \begin{bmatrix} 0.863\,617 & -0.306\,058 \\ 0.720\,296 & -0.508\,629 \end{bmatrix}$$
$$\boldsymbol{T}_4(8) = \begin{bmatrix} 0.606\,242 & -0.537\,680 \\ -0.425\,628 & 0.708\,075 \end{bmatrix}$$

which yield the noise gain $I(\alpha \boldsymbol{I}_2 \oplus \beta \boldsymbol{I}_2) = 1.614\,588$.

Next, a refined solution that offers further reduced noise gain is deduced by calculating an optimal diagonal EF matrix for the optimal realization $(\boldsymbol{T}(8)^{-1}\boldsymbol{A}\boldsymbol{T}(8), \boldsymbol{T}(8)^{-1}\boldsymbol{b}, \boldsymbol{c}\boldsymbol{T}(8), d)_{2,2}$. In this case, the optimal diagonal EF is obtained using (48) as

$$\boldsymbol{D} = \mathrm{diag}\{0.978\,520, 0.962\,184, 0.947\,598, 0.893\,592\}$$

which yields $I_{\min}(\boldsymbol{D}) = 1.610\,741$.

Finally, we apply the joint optimization of a general EF matrix $\boldsymbol{D}$ and a coordinate transformation matrix $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ described in Section IV to the original realization $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{2,2}$. In case $\mu = 0.001$, the transformation matrix $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ is computed as

$$\boldsymbol{T}_1 = \begin{bmatrix} 0.714\,538 & -0.598\,664 \\ 0.847\,981 & -0.415\,827 \end{bmatrix}$$
$$\boldsymbol{T}_4 = \begin{bmatrix} 0.526\,874 & -0.573\,867 \\ -0.313\,219 & 0.767\,659 \end{bmatrix}$$

which yield the noise gain $I(\boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}) = \mathrm{tr}[\boldsymbol{c}^T\boldsymbol{c}\boldsymbol{P}] = 0.347\,755$.

$$\bar{\boldsymbol{A}} = \begin{bmatrix} 0.965\,031 & -0.178\,310 & -0.058\,655 & 0.167\,811 \\ 0.115\,198 & 0.923\,959 & 0.167\,811 & -0.480\,100 \\ 0.021\,491 & -0.013\,210 & 0.965\,031 & 0.115\,198 \\ -0.013\,210 & 0.045\,857 & -0.178\,310 & 0.923\,959 \end{bmatrix}$$
$$\bar{\boldsymbol{b}} = [0.039\,012 \quad -0.111\,613 \quad 0.319\,129 \quad 0.142\,200]^T$$
$$\bar{\boldsymbol{c}} = [-0.590\,350 \quad -0.263\,054 \quad -0.072\,168 \quad 0.206\,471]$$

$$\bar{\boldsymbol{K}}_c = \begin{bmatrix} 1.0 & -0.221\,999 & 0.064\,066 & -0.184\,141 \\ -0.221\,999 & 1.0 & -0.036\,319 & 0.155\,751 \\ 0.063\,821 & -0.036\,079 & 1.0 & -0.221\,999 \\ -0.184\,141 & 0.155\,751 & -0.221\,999 & 1.0 \end{bmatrix}$$
$$\bar{\boldsymbol{W}}_o = \begin{bmatrix} 3.422\,064 & -0.759\,695 & 0.219\,239 & -0.124\,286 \\ -0.759\,695 & 3.422\,064 & -0.630\,143 & 0.532\,989 \\ 0.219\,239 & -0.630\,143 & 3.422\,064 & -0.759\,695 \\ -0.124\,286 & 0.532\,989 & -0.759\,695 & 3.422\,064 \end{bmatrix}$$

TABLE I
NOISE GAIN $I(\boldsymbol{D})$ FOR DIFFERENT EF SCHEMES

Error-Feedback Scheme	Accuracy of $\boldsymbol{D}$		
	Infinite Precision	3 Bit Quantization	Integer Quantization
$\boldsymbol{D}$=0	13.688256		
General $\boldsymbol{D}$	0.465549	0.555529	2.040208
Block-Diagonal $\boldsymbol{D}$	1.555329	1.612408	2.040208
Diagonal $\boldsymbol{D}$	1.908903	1.937559	2.040208
Scalar $\boldsymbol{D} = \alpha \boldsymbol{I}_m \oplus \beta \boldsymbol{I}_n$	1.950396	1.965326	2.040208
Jointly Optimized $\boldsymbol{T}$ and $\boldsymbol{D} = \alpha \boldsymbol{I}_m \oplus \beta \boldsymbol{I}_n$	1.614588	1.653024	1.660762
Optimal $\boldsymbol{T}$ and Diagonal $\boldsymbol{D}$	1.610741	1.635407	1.660762
Jointly Optimized $\boldsymbol{T}$ and General $\boldsymbol{D}$	0.347755	0.416500	1.773003

The simulations described above are summarized in Table I, where 3–bit quantization (integer quantization) implies that the elements of matrix $\boldsymbol{D}$ are rounded to power-of-two quantization with 3 bits after binary point (integer quantization). From this table, it is observed that the utilization of an optimal EF matrix leads to considerable reduction in roundoff noise, even when a scalar matrix $\boldsymbol{D} = \alpha \boldsymbol{I}_m \oplus \beta \boldsymbol{I}_n$ with quantized α and β. It is also observed that when the transformation matrix is jointly optimized, further noise reduction can be achieved compared with that in the conventional optimal realization.

VI. CONCLUSION

The minimization of roundoff noise in 2-D state-space digital filters by means of EF and joint EF/coordinate transformation optimization has been investigated. General, block-diagonal, diagonal, and scalar EF matrices for minimizing the noise gain in a given 2-D state-space digital filter have been derived. Then, an iterative procedure for minimizing the roundoff noise in a 2-D digital filter has also been developed by jointly optimizing a scalar EF matrix and a coordinate transformation subject to the usual l_2-norm dynamic-range scaling constraints. Furthermore, an analytical method for the joint optimization of a general EF matrix and a coordinate transformation matrix under the scaling constraints has been proposed. Simulation results have been presented to illustrate the validity of our proposed algorithms.

APPENDIX
DERIVATION OF MATRIX $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$

From (45), the optimal coordinate transformation matrices $\boldsymbol{T}_1$ and $\boldsymbol{T}_4$ that minimize (41) for fixed α and β can be obtained in closed form as

$$\begin{aligned} \boldsymbol{T}_1 &= \frac{1}{\sqrt{m}} \left(\sum_{i=1}^{m} \mu_i \right)^{\frac{1}{2}} \boldsymbol{W}_{1\alpha}^{-\frac{1}{2}} \left[\boldsymbol{W}_{1\alpha}^{\frac{1}{2}} \boldsymbol{K}_{c1} \boldsymbol{W}_{1\alpha}^{\frac{1}{2}} \right]^{\frac{1}{4}} \boldsymbol{U}_1 \\ \boldsymbol{T}_4 &= \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} \nu_i \right)^{\frac{1}{2}} \boldsymbol{W}_{4\beta}^{-\frac{1}{2}} \left[\boldsymbol{W}_{4\beta}^{\frac{1}{2}} \boldsymbol{K}_{c4} \boldsymbol{W}_{4\beta}^{\frac{1}{2}} \right]^{\frac{1}{4}} \boldsymbol{U}_4 \end{aligned} \quad \text{(A.1)}$$

where $\boldsymbol{U}_1$ and $\boldsymbol{U}_4$ are arbitrary $m \times m$ and $n \times n$ orthogonal matrices, respectively. From (A.1), it follows that

$$\begin{aligned} \boldsymbol{T}_1^{-1} \boldsymbol{K}_{c1} \boldsymbol{T}_1^{-T} &= m \left(\sum_{i=1}^{m} \mu_i \right)^{-1} \boldsymbol{U}_1 \left[\boldsymbol{W}_{1\alpha}^{\frac{1}{2}} \boldsymbol{K}_{c1} \boldsymbol{W}_{1\alpha}^{\frac{1}{2}} \right]^{\frac{1}{2}} \boldsymbol{U}_1 \\ \boldsymbol{T}_4^{-1} \boldsymbol{K}_{c4} \boldsymbol{T}_4^{-T} &= n \left(\sum_{i=1}^{n} \nu_i \right)^{-1} \boldsymbol{U}_4 \left[\boldsymbol{W}_{4\beta}^{\frac{1}{2}} \boldsymbol{K}_{c4} \boldsymbol{W}_{4\beta}^{\frac{1}{2}} \right]^{\frac{1}{2}} \boldsymbol{U}_4. \end{aligned} \quad \text{(A.2)}$$

Next, we choose the $m \times m$ and $n \times n$ orthogonal matrices $\boldsymbol{U}_1$ and $\boldsymbol{U}_4$ such that (A.2) satisfies the scaling constraints in (18). To this end, we carry out the eigenvalue-eigenvector decompositions

$$\begin{aligned} \left[\boldsymbol{W}_{1\alpha}^{\frac{1}{2}} \boldsymbol{K}_{c1} \boldsymbol{W}_{1\alpha}^{\frac{1}{2}} \right]^{\frac{1}{2}} &= \boldsymbol{R}_1 \boldsymbol{\Theta}_1 \boldsymbol{R}_1^T \\ \left[\boldsymbol{W}_{4\beta}^{\frac{1}{2}} \boldsymbol{K}_{c4} \boldsymbol{W}_{4\beta}^{\frac{1}{2}} \right]^{\frac{1}{2}} &= \boldsymbol{R}_4 \boldsymbol{\Theta}_4 \boldsymbol{R}_4^T \end{aligned} \quad \text{(A.3)}$$

where

$$\begin{aligned} \boldsymbol{\Theta}_1 &= \text{diag}\{\mu_1, \mu_2, \ldots, \mu_m\}, \quad \boldsymbol{R}_1 \boldsymbol{R}_1^T = \boldsymbol{I}_m \\ \boldsymbol{\Theta}_4 &= \text{diag}\{\nu_1, \nu_2, \ldots, \nu_n\}, \quad \boldsymbol{R}_4 \boldsymbol{R}_4^T = \boldsymbol{I}_n. \end{aligned}$$

As a result, it follows that

$$\begin{aligned} m \left(\sum_{i=1}^{m} \mu_i \right)^{-1} \left[\boldsymbol{W}_{1\alpha}^{\frac{1}{2}} \boldsymbol{K}_{c1} \boldsymbol{W}_{1\alpha}^{\frac{1}{2}} \right]^{\frac{1}{2}} &= \boldsymbol{R}_1 \boldsymbol{\Lambda}_1^{-2} \boldsymbol{R}_1^T \\ n \left(\sum_{i=1}^{n} \nu_i \right)^{-1} \left[\boldsymbol{W}_{4\beta}^{\frac{1}{2}} \boldsymbol{K}_{c4} \boldsymbol{W}_{4\beta}^{\frac{1}{2}} \right]^{\frac{1}{2}} &= \boldsymbol{R}_4 \boldsymbol{\Lambda}_4^{-2} \boldsymbol{R}_4^T \end{aligned} \quad \text{(A.4)}$$

where

$$\begin{aligned} \boldsymbol{\Lambda}_1^2 &= \frac{1}{m} \left(\sum_{i=1}^{m} \mu_i \right) \text{diag} \left\{ \mu_1^{-1}, \mu_2^{-1}, \ldots, \mu_m^{-1} \right\} \\ \boldsymbol{\Lambda}_4^2 &= \frac{1}{n} \left(\sum_{i=1}^{n} \nu_i \right) \text{diag} \left\{ \nu_1^{-1}, \nu_2^{-1}, \ldots, \nu_n^{-1} \right\}. \end{aligned}$$

Now, an $m \times m$ orthogonal matrix $\boldsymbol{S}_1$ and an $n \times n$ orthogonal matrix $\boldsymbol{S}_4$ such that

$$\begin{aligned} \boldsymbol{S}_1 \boldsymbol{\Lambda}_1^{-2} \boldsymbol{S}_1^T &= \begin{bmatrix} 1 & * & \cdots & * \\ * & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ * & \cdots & * & 1 \end{bmatrix} \\ \boldsymbol{S}_4 \boldsymbol{\Lambda}_4^{-2} \boldsymbol{S}_4^T &= \begin{bmatrix} 1 & * & \cdots & * \\ * & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ * & \cdots & * & 1 \end{bmatrix} \end{aligned} \quad \text{(A.5)}$$

can be obtained by numerical manipulations [3, p. 278]. By choosing $\boldsymbol{U}_1 = \boldsymbol{R}_1 \boldsymbol{S}_1^T$ and $\boldsymbol{U}_4 = \boldsymbol{R}_4 \boldsymbol{S}_4^T$ in (A.1), the optimal coordinate transformation matrix $T = T_1 \oplus T_4$ satisfying (18) and (46) simultaneously can now be constructed as

$$T_1 = \frac{1}{\sqrt{m}} \left(\sum_{i=1}^{m} \mu_i \right)^{\frac{1}{2}} W_{1\alpha}^{-\frac{1}{2}} \left[W_{1\alpha}^{\frac{1}{2}} K_{c1} W_{1\alpha}^{\frac{1}{2}} \right]^{\frac{1}{4}} R_1 S_1^T$$
$$T_4 = \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} \nu_i \right)^{\frac{1}{2}} W_{4\beta}^{-\frac{1}{2}} \left[W_{4\beta}^{\frac{1}{2}} K_{c4} W_{4\beta}^{\frac{1}{2}} \right]^{\frac{1}{4}} R_4 S_4^T. \quad \text{(A.6)}$$

REFERENCES

[1] S. Y. Hwang, "Roundoff noise in state-space digital filtering: A general analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 256–262, June 1976.

[2] C. T. Mullis and R. A. Roberts, "Synthesis of minimum roundoff noise fixed-point digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-23, pp. 551–562, Sept. 1976.

[3] 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.

[4] L. B. Jackson, A. G. Lindgren, and Y. Kim, "Optimal synthesis of second-order state-space structures for digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 149–153, Mar. 1979.

[5] M. Kawamata and T. Higuchi, "Synthesis of 2-D separable denominator digital filters with minimum roundoff noise and no overflow oscillations," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 365–372, Apr. 1986.

[6] ——, "A unified study on the roundoff noise in 2-D state-space digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 724–730, July 1986.

[7] W.-S. Lu and A. Antoniou, "Synthesis of 2-D state-space fixed-point digital filter structures with minimum roundoff noise," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 965–973, Oct. 1986.

[8] T. Hinamoto, T. Hamanaka, and S. Maekawa, "A generalized study on the synthesis of 2-D state-space digital filters with minimum roundoff noise," *IEEE Trans. Circuits Syst.*, vol. 35, pp. 1037–1042, Aug. 1988.

[9] H. A. Spang III and P. M. Shultheiss, "Reduction of quantizing noise by use of feedback," *IRE Trans. Commun. Syst.*, vol. CS-10, pp. 373–380, Dec. 1962.

[10] T. Thong and B. Liu, "Error spectrum shaping in narrowband recursive digital filters," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP–25, pp. 200–203, Apr. 1977.

[11] T. L. Chang and S. A. White, "An error cancellation digital filter structure and its distributed-arithmetic implementation," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 339–342, Apr. 1981.

[12] D. C. Munson and D. Liu, "Narrowband recursive filters with error spectrum shaping," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 160–163, Feb. 1981.

[13] W. E. Higgins and D. C. Munson, "Noise reduction strategies for digital filters: Error spectrum shaping versus the optimal linear state-space formulation," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-30, pp. 963–973, Dec. 1982.

[14] M. Renfors, "Roundoff noise in error-feedback state-space filters," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, Apr. 1983, pp. 619–622.

[15] W. E. Higgins and D. C. Munson, "Optimal and suboptimal error-spectrum shaping for cascade-form digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 429–437, May 1984.

[16] T. I. Laakso and I. O. Hartimo, "Noise reduction in recursive digital filters using high-order error feedback," *IEEE Trans. Signal Processing*, vol. 40, pp. 1096–1107, May 1992.

[17] P. P. Vaidyanathan, "On error-spectrum shaping in state-space digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 88–92, Jan. 1985.

[18] D. 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.

[19] T. Hinamoto, S. Karino, and N. Kuroda, "Error spectrum shaping in 2-D digital filters," *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 1, pp. 348–351, May 1995.

[20] P. Agathoklis and C. Xiao, "Low roundoff noise structures for 2-D filters," *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 2, pp. 352–355, May 1996.

[21] T. Hinamoto, S. Karino, and N. Kuroda, "2-D state-space digital filters with error spectrum shaping," *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 2, pp. 766–769, May 1996.

[22] T. Hinamoto, N. Kuroda, and T. Kuma, "Error feedback for noise reduction in 2-D digital filers with quadrantally symmetric or antisymmetric coefficients," *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 4, pp. 2461–2464, June 1997.

[23] T. Hinamoto, S. Karino, N. Kuroda, and T. Kuma, "Error spectrum shaping in two-dimensional recursive digital filters," *IEEE Trans. Circuits Syst.*, vol. 46, pp. 1203–1215, Oct. 1999.

[24] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Trans. Signal Processing*, vol. 41, pp. 629–637, Feb. 1993.

[25] D. Williamson, "Delay replacement in direct form structures," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 453–460, Apr. 1988.

[26] M. M. Ekanayake and K. Premaratne, "Two-dimensional delta-operator formulated discrete-time systems: Analysis and synthesis of minimum roundoff noise realizations," *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 2, pp. 213–216, May 1996.

[27] R. P. Roesser, "A discrete state-space model for linear image processing," *IEEE Trans. Automat. Contr.*, vol. AC-20, pp. 1–10, Feb. 1975.

[28] S. 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 notions of controllability, observability, and minimality," *Proc. IEEE*, vol. 65, pp. 945–961, June 1977.

[29] T. Hinamoto, T. Hamanaka, and S. Maekawa, "Relationship between two forms in the 2-D unit noise matrix," *IEEE Trans. Circuits Syst.*, vol. 35, pp. 609–610, May 1988.

[30] L. L. Scharf, *Statistical Signal Processing*. Reading, MA: Addison-Wesley, 1991.

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

Takao Hinamoto (M'77–SM'84–F'01) received the B.E. degree from Okayama University, Okayama, Japan, in 1969, the M.E. degree from Kobe University, Kobe, Japan, in 1971, and the Dr.Eng. degree from Osaka University, Osaka, Japan, in 1977, all in electrical engineering.

From 1972 to 1988, he was with the Faculty of Engineering, Kobe University. From 1979 to 1981, he was on leave from Kobe University as a Visiting Member of Staff with the Department of Electrical Engineering, Queen's University, Kingston, ON, Canada. From 1988 to 1991, he was a Professor of electronic circuits with the Faculty of Engineering, Tottori University, Tottori, Japan. Since January 1992, he has been a Professor of electronic control with the Department of Electrical Engineering, Hiroshima University, Hiroshima, Japan. His research interests include digital signal processing, system theory, and control engineering. He has published more than 280 papers in these areas, and is the co-editor and coauthor of *Two-Dimensional Signal and Image Processing* (Tokyo, Japan: SICE, 1996).

Dr. Hinamoto served as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II from 1993 to 1995 and presently serves as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I. He was the Guest Editor of the special section of Digital Signal Processing (DSP) for the August 1998 issue of the *IEICE Transactions on Fundamentals*. He served as Chair of the 12th DSP Symposium held in Hiroshima in November 1997, which is the 12th of a series of annual symposia sponsored by the DSP Technical Committee of IEICE. Since 1995, he has been a member of the steering committee of the IEEE Midwest Symposium on Circuits and Systems and, since 1998, a member of the DSP Technical Committee in the IEEE Circuits and Systems Society. He was a member of the Technical Program Committee for ISCAS'99. Since 1993, he has served as a senator or member of the Board of Directors in the Society of Instrument and Control Engineers (SICE), and from 1999 to 2001, he was Chair of the Chugoku Chapter of the SICE. He played a leading role in establishing the Hiroshima Section of IEEE and served as Interim Chair of the section. He is a recipient of the IEEE Third Millennium Medal.

Keisuke Higashi received the B.E. and M.E. degrees in electrical engineering from Hiroshima University, Hiroshima, Japan, in 2000 and 2002, respectively.

In April 2002, he joined the Mitsubishi Heavy Industries, Ltd., Nagasaki, Japan. His research interests are in digital signal processing and system theory.

Wu-Sheng Lu (S'81–M'85–SM'90–F'99) received the undergraduate degree in mathematics from Fudan University, Shanghai, China, in 1964 and the M.S. degree in electrical engineering and Ph.D. degree in control science from University of Minnesota, Minneapolis, in 1983, and 1984, respectively.

He was a Post-Doctoral Fellow with the University of Victoria, Victoria, BC, Canada, in 1985 and a Visiting Assistant Professor with the University of Minnesota in 1986. Since 1987, he has been with the University of Victoria, where he is currently a Professor. His teaching and research interests are in the areas of digital signal processing and application of optimization methods. He is the coauthor, with A. Antoniou, of *Two-Dimensional Digital Filters* (New York: Marcel Dekker, 1992). He was an Associate Editor of the *Canadian Journal of Electrical and Computer Engineering* in 1989 and was its Editor from 1990 to 1992. He is presently an Associate Editor for the *International Journal of Multidimensional Systems and Signal Processing*.

Dr. Lu served as an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II from 1993 to 1995 and for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I from 1999 to 2001. He is a Fellow of the Engineering Institute of Canada.