Roundoff Noise Minimization for 2-D State-Space Digital Filters Using Joint Optimization of Error Feedback and Realization

Takao Hinamoto, *Fellow, IEEE,* Hiroaki Ohnishi and Wu-Sheng Lu, *Fellow, IEEE*

***Abstract*— The joint optimization problem of error feedback and realization for two-dimensional (2-D) state-space digital filters to minimize the effects of roundoff noise at the filter output subject to L_2-norm dynamic-range scaling constraints is investigated. It is shown that the problem can be converted into an unconstrained optimization problem by using linear-algebraic techniques. The unconstrained optimization problem at hand is then solved iteratively by applying an efficient quasi-Newton algorithm with closed-form formulas for key gradient evaluation. Analytical details are given as to how the proposed technique can be applied to the cases where the error-feedback matrix is a general, block-diagonal, diagonal, or block-scalar matrix. A case study is presented to illustrate the utility of the proposed technique.**

***Index Terms*— 2-D digital filters, roundoff noise minimization, joint optimization, error feedback, state-space realization, L_2-scaling constraints.**

I. INTRODUCTION

When implementing recursive digital filters in fixed-point arithmetic, the problem of reducing the effects of roundoff noise at the filter output is of critical importance. Error feedback (EF) is a useful tool for the reduction of finite-word-length (FWL) effects in recursive digital filters. Many EF techniques have been reported in the past for one-dimensional (1-D) recursive digital filters [1]-[10], and more recently for 2-D recursive digital filters [11]-[15]. The roundoff noise can also be reduced by introducing a delta operator to recursive digital filters [16]-[18] or by applying a new structure based on the concept of polynomial operators for digital filter implementation [19]. Another useful approach is to construct the state-space filter structure for the roundoff noise gain to be minimized by applying a linear transformation to state-space coordinates subject to L_2-norm dynamic-range scaling constraints [20]-[23]. The problem of synthesizing such a state-space filter structure with minimum roundoff noise has been explored for 2-D state-space digital filters [24]-[27]. As a natural extension of the aforementioned methods, efforts have been made to develop new methods that combine EF and realization, for achieving better performance [28]-[30]. Separately-optimized analytical algorithms have been proposed for either 1-D [28] or 2-D [29] state-space digital filters. In [28] and [29], jointly-optimized iterative algorithms have also been considered for filters with a general or scalar EF matrix. In [30], a jointly-optimized iterative algorithm has been developed for 1-D state-space digital filters with a general, diagonal, or scalar EF matrix by applying a quasi-Newton method.

This paper investigates the problem of jointly optimizing EF and realization for 2-D state-space digital filters to minimize the roundoff noise subject to L_2-norm dynamic-range scaling constraints. To this end, an iterative technique which relies on an efficient quasi-Newton algorithm [31] is developed. It is shown that the constrained optimization problem can be converted into an unconstrained optimization problem by using linear-algebraic techniques. The proposed technique can be applied to the cases where the EF matrix is a general, block-diagonal, diagonal, or block-scalar matrix. A case study is presented to illustrate the algorithm proposed and to demonstrate its performance.

Throughout the paper, $\boldsymbol{I}_n$ stands for the identity matrix of dimension $n \times n$, $\oplus$ is used to denote the direct sum of matrices, 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.

II. 2-D STATE-SPACE DIGITAL FILTERS WITH ERROR FEEDBACK

Suppose that a local state-space (LSS) model $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{m,n}$ for 2-D recursive digital filters is described by [32]

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

where

$$\boldsymbol{x}_{11}(i,j) = \begin{bmatrix} \boldsymbol{x}^h(i+1,j) \\ \boldsymbol{x}^v(i,j+1) \end{bmatrix}, \qquad \boldsymbol{x}(i,j) = \begin{bmatrix} \boldsymbol{x}^h(i,j) \\ \boldsymbol{x}^v(i,j) \end{bmatrix},$$

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 \\ \boldsymbol{A}_3 & \boldsymbol{A}_4 \end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix}, \qquad \boldsymbol{c} = \begin{bmatrix} \boldsymbol{c}_1 & \boldsymbol{c}_2 \end{bmatrix},$$

with an $m \times 1$ horizontal state vector $\boldsymbol{x}^h(i,j)$, an $n \times 1$ vertical state vector $\boldsymbol{x}^v(i,j)$, a scalar input $u(i,j)$, a scalar output $y(i,j)$, and real constant matrices $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{A}_3$, $\boldsymbol{A}_4$, $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, $\boldsymbol{c}_1$, $\boldsymbol{c}_2$ and d of appropriate dimensions. The LSS model in (1) is assumed to be BIBO stable, separately locally controllable and separately locally observable [33].

Due to finite register sizes, we impose FWL constraints on the local state vector $\boldsymbol{x}(i,j)$, the input, the output, and on the

Manuscript received June 7, 2005; revised mm dd, 2005.

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

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

coefficients in the realization $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{m,n}$. Assuming that the quantization is performed before matrix-vector multiplication, the actual FWL filter of (1) is implemented 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 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, while 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 multiplications and additions, where the sign bit is not counted. In a fixed-point implementation, the quantization is usually carried out by two's complement truncation which 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 quantization error $\boldsymbol{e}(i,j)$ is modeled as a zero-mean white noise of covariance $\sigma^2 \boldsymbol{I}_{m+n}$ with

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

In order 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 $\boldsymbol{D}$. Under these circumstances, the filter model can be represented as

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

where $\boldsymbol{D}$ is referred to as the 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

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

From (5) it follows that the 2-D transfer function from the quantization error $\boldsymbol{e}(i,j)$ to the filter output $\Delta y(i,j)$ is given by

$$\boldsymbol{G}_D(z_1, z_2) = \boldsymbol{c}(\boldsymbol{Z} - \boldsymbol{A})^{-1}(\boldsymbol{A} - \boldsymbol{D}) + \boldsymbol{c}, \tag{6}$$

where $\boldsymbol{Z} = z_1 \boldsymbol{I}_m \oplus z_2 \boldsymbol{I}_n$.

For the 2-D filter in (4) with EF, the noise gain $I(\boldsymbol{D}) = \sigma_{out}^2/\sigma^2$ is evaluated by

$$I(\boldsymbol{D}) = \mathrm{tr}[\boldsymbol{W}_D], \tag{7}$$

where σ_{out}^2 denotes noise variance at the filter output and

$$\boldsymbol{W}_D = \frac{1}{(2\pi j)^2} \oint_{\Gamma_1} \oint_{\Gamma_2} \boldsymbol{G}_D^*(z_1, z_2) \boldsymbol{G}_D(z_1, z_2) \frac{dz_1 dz_2}{z_1 z_2},$$

with $\Gamma_i = \{z_i : |z_i| = 1\}$ for $i = 1, 2$. Utilizing the 2-D Cauchy integral theorem, we can express matrix $\boldsymbol{W}_D$ in (7) 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 matrix $\boldsymbol{W}_o$ is the local observability Gramian defined by

$$\begin{aligned} \boldsymbol{W}_o &= \frac{1}{(2\pi j)^2} \oint_{\Gamma_1} \oint_{\Gamma_2} (\boldsymbol{Z}^* - \boldsymbol{A}^T)^{-1} \boldsymbol{c}^T \boldsymbol{c} (\boldsymbol{Z} - \boldsymbol{A})^{-1} \frac{dz_1 dz_2}{z_1 z_2} \\ &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \boldsymbol{g}(i,j)^T \boldsymbol{g}(i,j), \end{aligned} \tag{9}$$

with

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

and the partial ordering for integer pairs (i,j) used in [32, p.2].

We remark that matrix $\boldsymbol{W}_o$ in (9) is referred to as the *unit noise matrix* for the 2-D filter (2), and matrix $\boldsymbol{W}_D$ in (8) is viewed as the *unit noise matrix* for the 2-D filter in (4) with EF specified by the matrix $\boldsymbol{D}$.

In the case where there is no EF in the 2-D filter, the noise gain $I(\boldsymbol{D})$ with $\boldsymbol{D} = \boldsymbol{0}$ can be expressed as

$$I(\boldsymbol{0}) = \mathrm{tr}[\boldsymbol{A}^T \boldsymbol{W}_o \boldsymbol{A} + \boldsymbol{c}^T \boldsymbol{c}] = \mathrm{tr}[\boldsymbol{W}_o]. \tag{11}$$

It is noted that the L_2-norm dynamic-range scaling constraints on the local state vector $\boldsymbol{x}(i,j)$ involves the local controllability Gramian defined by

$$\begin{aligned} \boldsymbol{K}_c &= \frac{1}{(2\pi j)^2} \oint_{\Gamma_1} \oint_{\Gamma_2} (\boldsymbol{Z} - \boldsymbol{A})^{-1} \boldsymbol{b} \boldsymbol{b}^T (\boldsymbol{Z}^* - \boldsymbol{A}^T)^{-1} \frac{dz_1 dz_2}{z_1 z_2} \\ &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \boldsymbol{f}(i,j) \boldsymbol{f}(i,j)^T, \end{aligned} \tag{12}$$

where

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

III. JOINT ERROR-FEEDBACK AND REALIZATION OPTIMIZATION

A. Probem Statement

The change of coordinates from local state vector $\boldsymbol{x}(i,j)$ to $\overline{\boldsymbol{x}}(i,j)$, defined by a linear transformation $\overline{\boldsymbol{x}}(i,j) = \boldsymbol{T}^{-1}\boldsymbol{x}(i,j)$ with $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$, transforms the LSS model $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{m,n}$ in (1) to a new realization $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_{m,n}$ with

$$\overline{\boldsymbol{A}} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}, \quad \overline{\boldsymbol{b}} = \boldsymbol{T}^{-1}\boldsymbol{b}, \quad \overline{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T}. \tag{13}$$

The local controllability Gramian $\overline{\boldsymbol{K}}_c$ and the local observability Gramian $\overline{\boldsymbol{W}}_o$ in the new realization then satisfy the relations

$$\overline{\boldsymbol{K}}_c = \boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T}, \qquad \overline{\boldsymbol{W}}_o = \boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{T}. \tag{14}$$

If the L_2-norm dynamic-range scaling constraints specified by

$$(\overline{\boldsymbol{K}}_c)_{ii} = (\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T})_{ii} = 1, \quad i = 1, 2, \cdots, m+n \tag{15}$$

are imposed on the new realization, then it is known that [25],[26]

$$\min_{\boldsymbol{T}} \ \mathrm{tr}[\overline{\boldsymbol{W}}_o] = \frac{1}{m}\left(\sum_{i=1}^{m}\sigma_{1i}\right)^2 + \frac{1}{n}\left(\sum_{i=1}^{n}\sigma_{4i}\right)^2 \tag{16}$$

where σ_{1i}^2 for $i = 1, 2, \cdots, m$ and σ_{4i}^2 for $i = 1, 2, \cdots, n$ are the eigenvalues of the $m \times m$ matrix $\boldsymbol{K}_{1c}\boldsymbol{W}_{1o}$ and the $n \times n$ matrix $\boldsymbol{K}_{4c}\boldsymbol{W}_{4o}$, respectively, and

$$\boldsymbol{K}_c = \begin{bmatrix} \boldsymbol{K}_{1c} & \boldsymbol{K}_{2c} \\ \boldsymbol{K}_{3c} & \boldsymbol{K}_{4c} \end{bmatrix}, \qquad \boldsymbol{W}_o = \begin{bmatrix} \boldsymbol{W}_{1o} & \boldsymbol{W}_{2o} \\ \boldsymbol{W}_{3o} & \boldsymbol{W}_{4o} \end{bmatrix}.$$

The LSS model $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_{m,n}$ satisfying (15) and (16) simultaneously is known as the *optimal realization* (which is sometimes also referred to as the *optimal filter structure*). A method for synthesizing such a filter structure was proposed in [25],[26].

If a coordinate transformation $\overline{\boldsymbol{x}}(i,j) = \boldsymbol{T}^{-1}\boldsymbol{x}(i,j)$ with $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ is applied to the LSS model in (1), then the 2-D filter in (4) with EF can be characterized by

$$\begin{aligned} \tilde{\boldsymbol{x}}_{11}(i,j) &= \overline{\boldsymbol{A}}\,\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,j)] + \overline{\boldsymbol{b}}\,u(i,j) + \boldsymbol{D}\boldsymbol{e}(i,j) \\ \tilde{y}(i,j) &= \overline{\boldsymbol{c}}\,\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,j)] + du(i,j). \end{aligned} \tag{17}$$

In this case, the noise gain $I(\boldsymbol{D}, \boldsymbol{T})$ can be expressed as a function of matrices $\boldsymbol{D}$ and $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ in the form

$$I(\boldsymbol{D}, \boldsymbol{T}) = \mathrm{tr}[\overline{\boldsymbol{W}}_D], \tag{18}$$

where

$$\overline{\boldsymbol{W}}_D = (\overline{\boldsymbol{A}} - \boldsymbol{D})^T\overline{\boldsymbol{W}}_o\,(\overline{\boldsymbol{A}} - \boldsymbol{D}) + \overline{\boldsymbol{c}}^T\overline{\boldsymbol{c}}.$$

The roundoff noise minimization problem can now be formulated as follows: *Given* $\boldsymbol{A}$, $\boldsymbol{b}$ *and* $\boldsymbol{c}$ (*and hence,* $\boldsymbol{W}_o$ *and* $\boldsymbol{K}_c$), *obtain matrices* $\boldsymbol{D}$ *and* $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ *which jointly minimize the noise gain in (18) subject to the scaling constraints in (15).*

B. Problem Relaxation and Conversion

In order to reduce solution sensitivity, the objective function in (18) is modified to

$$J(\boldsymbol{D}, \boldsymbol{T}) = \mathrm{tr}[(1-\mu)\overline{\boldsymbol{W}}_D + \mu\overline{\boldsymbol{W}}_o], \tag{19}$$

where $0 \le \mu \le 1$ is a scalar parameter that weights the importance of reducing $\mathrm{tr}[\overline{\boldsymbol{W}}_o]$ relative to reducing $\mathrm{tr}[\overline{\boldsymbol{W}}_D]$. Defining

$$\begin{aligned} \hat{\boldsymbol{T}} &= \hat{\boldsymbol{T}}_1 \oplus \hat{\boldsymbol{T}}_4 \\ &= (\boldsymbol{T}_1 \oplus \boldsymbol{T}_4)^T(\boldsymbol{K}_{1c} \oplus \boldsymbol{K}_{4c})^{-\frac{1}{2}}, \end{aligned} \tag{20}$$

it follows that

$$\overline{\boldsymbol{K}}_c = \hat{\boldsymbol{T}}^{-T}\begin{bmatrix} \boldsymbol{I}_m & \boldsymbol{K}_{1c}^{-\frac{1}{2}}\boldsymbol{K}_{2c}\boldsymbol{K}_{4c}^{-\frac{1}{2}} \\ \boldsymbol{K}_{4c}^{-\frac{1}{2}}\boldsymbol{K}_{3c}\boldsymbol{K}_{1c}^{-\frac{1}{2}} & \boldsymbol{I}_n \end{bmatrix}\hat{\boldsymbol{T}}^{-1}. \tag{21}$$

This enables one to reduce the scaling constraints in (15) to

$$\begin{aligned} (\hat{\boldsymbol{T}}_1^{-T}\hat{\boldsymbol{T}}_1^{-1})_{ii} &= 1, \qquad i = 1, 2, \cdots, m \\ (\hat{\boldsymbol{T}}_4^{-T}\hat{\boldsymbol{T}}_4^{-1})_{kk} &= 1, \qquad k = 1, 2, \cdots, n. \end{aligned} \tag{22}$$

The constraints in (22) simply state that each column in matrices $\hat{\boldsymbol{T}}_1^{-1}$ and $\hat{\boldsymbol{T}}_4^{-1}$ must be a unity vector. It can be verified that these constraints are satisfied if $\hat{\boldsymbol{T}}_1^{-1}$ and $\hat{\boldsymbol{T}}_4^{-1}$ assume the forms

$$\begin{aligned} \hat{\boldsymbol{T}}_1^{-1} &= \left[\frac{\boldsymbol{t}_{11}}{||\boldsymbol{t}_{11}||}, \frac{\boldsymbol{t}_{12}}{||\boldsymbol{t}_{12}||}, \cdots, \frac{\boldsymbol{t}_{1m}}{||\boldsymbol{t}_{1m}||}\right] \\ \hat{\boldsymbol{T}}_4^{-1} &= \left[\frac{\boldsymbol{t}_{41}}{||\boldsymbol{t}_{41}||}, \frac{\boldsymbol{t}_{42}}{||\boldsymbol{t}_{42}||}, \cdots, \frac{\boldsymbol{t}_{4n}}{||\boldsymbol{t}_{4n}||}\right] \end{aligned} \tag{23}$$

where $\boldsymbol{t}_{1i}$ for $i = 1, 2, \cdots, m$ and $\boldsymbol{t}_{4j}$ for $j = 1, 2, \cdots, n$ are $m \times 1$ and $n \times 1$ real vectors, respectively. In such a case, matrix $\overline{\boldsymbol{W}}_D$ in (18) can be written as

$$\overline{\boldsymbol{W}}_D = \hat{\boldsymbol{T}}\,[(\hat{\boldsymbol{A}} - \hat{\boldsymbol{T}}^T\boldsymbol{D}\hat{\boldsymbol{T}}^{-T})^T\hat{\boldsymbol{W}}_o(\hat{\boldsymbol{A}} - \hat{\boldsymbol{T}}^T\boldsymbol{D}\hat{\boldsymbol{T}}^{-T}) + \hat{\boldsymbol{C}}\,]\,\hat{\boldsymbol{T}}^T, \tag{24}$$

where $\hat{\boldsymbol{T}} = \hat{\boldsymbol{T}}_1 \oplus \hat{\boldsymbol{T}}_4$ and

$$\begin{aligned} \hat{\boldsymbol{A}} &= (\boldsymbol{K}_{1c} \oplus \boldsymbol{K}_{4c})^{-\frac{1}{2}}\,\boldsymbol{A}\,(\boldsymbol{K}_{1c} \oplus \boldsymbol{K}_{4c})^{\frac{1}{2}} \\ \hat{\boldsymbol{C}} &= (\boldsymbol{K}_{1c} \oplus \boldsymbol{K}_{4c})^{\frac{1}{2}}\,\boldsymbol{c}^T\boldsymbol{c}\,(\boldsymbol{K}_{1c} \oplus \boldsymbol{K}_{4c})^{\frac{1}{2}} \\ \hat{\boldsymbol{W}}_o &= (\boldsymbol{K}_{1c} \oplus \boldsymbol{K}_{4c})^{\frac{1}{2}}\,\boldsymbol{W}_o\,(\boldsymbol{K}_{1c} \oplus \boldsymbol{K}_{4c})^{\frac{1}{2}}. \end{aligned}$$

Under these circumstances, the objective function in (19) becomes

$$\begin{aligned} &J(\boldsymbol{D}, \hat{\boldsymbol{T}}) \\ &= (1-\mu)\,\mathrm{tr}[(\hat{\boldsymbol{T}}\hat{\boldsymbol{A}}^T - \boldsymbol{D}^T\hat{\boldsymbol{T}})\hat{\boldsymbol{W}}_o(\hat{\boldsymbol{A}}\hat{\boldsymbol{T}}^T - \hat{\boldsymbol{T}}^T\boldsymbol{D})] \\ &\quad + (1-\mu)\,\mathrm{tr}[\hat{\boldsymbol{T}}\,\hat{\boldsymbol{C}}\,\hat{\boldsymbol{T}}^T] + \mu\,\mathrm{tr}[\hat{\boldsymbol{T}}\,\hat{\boldsymbol{W}}_o\hat{\boldsymbol{T}}^T]. \end{aligned} \tag{25}$$

From the foregoing arguments, the problem of obtaining matrices $\boldsymbol{D}$ and $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ that minimize (19) subject to the scaling constraints in (15) is now converted into an unconstrained optimization problem of obtaining matrices $\boldsymbol{D}$ and $\hat{\boldsymbol{T}} = \hat{\boldsymbol{T}}_1 \oplus \hat{\boldsymbol{T}}_4$ that jointly minimize the noise gain in (25).

C. Optimization Method

Let $\boldsymbol{x}$ be the column vector that collects the variables in matrices $\boldsymbol{D}$, $[\boldsymbol{t}_{11}, \boldsymbol{t}_{12}, \cdots, \boldsymbol{t}_{1m}]$ and $[\boldsymbol{t}_{41}, \boldsymbol{t}_{42}, \cdots, \boldsymbol{t}_{4n}]$. Then, $J(\boldsymbol{D}, \hat{\boldsymbol{T}})$ is a function of $\boldsymbol{x}$, denoted by $J(\boldsymbol{x})$. The proposed algorithm starts with an initial point $\boldsymbol{x}_0$ obtained from an initial assignment $\boldsymbol{D} = \hat{\boldsymbol{T}} = \boldsymbol{I}_{m+n}$. In the kth iteration, a quasi-Newton algorithm updates the most recent point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as [31]

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{d}_k, \tag{26}$$

where

$$\begin{aligned} \boldsymbol{d}_k &= -\boldsymbol{S}_k\nabla J(\boldsymbol{x}_k) \\ \alpha_k &= arg\left[\min_{\alpha}\ J(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k)\right] \\ \boldsymbol{S}_{k+1} &= \boldsymbol{S}_k + \left(1 + \frac{\boldsymbol{\gamma}_k^T\boldsymbol{S}_k\boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_k^T\boldsymbol{\delta}_k}\right)\frac{\boldsymbol{\delta}_k\boldsymbol{\delta}_k^T}{\boldsymbol{\gamma}_k^T\boldsymbol{\delta}_k} - \frac{\boldsymbol{\delta}_k\boldsymbol{\gamma}_k^T\boldsymbol{S}_k + \boldsymbol{S}_k\boldsymbol{\gamma}_k\boldsymbol{\delta}_k^T}{\boldsymbol{\gamma}_k^T\boldsymbol{\delta}_k} \\ \boldsymbol{S}_0 &= \boldsymbol{I}, \quad \boldsymbol{\delta}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k, \quad \boldsymbol{\gamma}_k = \nabla J(\boldsymbol{x}_{k+1}) - \nabla J(\boldsymbol{x}_k). \end{aligned}$$

Here, $\nabla J(\boldsymbol{x})$ is the gradient of $J(\boldsymbol{x})$ with respect to $\boldsymbol{x}$, and $\boldsymbol{S}_k$ is a positive-definite approximation of the inverse Hessian matrix of $J(\boldsymbol{x})$. This iteration process continues until

$$|J(\boldsymbol{x}_{k+1}) - J(\boldsymbol{x}_k)| < \varepsilon, \tag{27}$$

where $\varepsilon > 0$ is a prescribed tolerance. If the iteration is terminated at step k, then $\boldsymbol{x}_k$ is deemed as a solution point.

The implementation of (26) requires the gradient of $J(\boldsymbol{x})$. Now we consider the cases where EF matrix is a general, block-diagonal, diagonal, or block-scalar matrix. It is noted that a general EF matrix is often too costly because it requires as many as $(m+n)^2$ explicit multiplications. The cost can be reduced, e.g., by constraining EF matrix to be a block-diagonal or diagonal (block-scalar), which reduces the number of distinct coefficients to m^2+n^2 or $m+n$.

A key quantity for the implementation of the quasi-Newton algorithm is the gradient $\nabla J(\boldsymbol{x})$. In what follows, we derive closed-form expressions of $\nabla J(\boldsymbol{x})$ for the cases where $\boldsymbol{D}$ assumes the form of a general, block-diagonal, diagonal, or block-scalar matrix.

Case 1: *$\boldsymbol{D}$ is a general matrix*

From (25), it is evident that the optimal choice of $\boldsymbol{D}$ is given by

$$\boldsymbol{D} = \hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{A}}\hat{\boldsymbol{T}}^T, \tag{28}$$

which leads to

$$J(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{A}}\hat{\boldsymbol{T}}^T, \hat{\boldsymbol{T}}) = \mathrm{tr}[\hat{\boldsymbol{T}}\{(1-\mu)\hat{\boldsymbol{C}} + \mu\hat{\boldsymbol{W}}_o\}\hat{\boldsymbol{T}}^T]. \tag{29}$$

In this case, the number of elements in vector $\boldsymbol{x}$ consisting of $\hat{\boldsymbol{T}} = \hat{\boldsymbol{T}}_1 \oplus \hat{\boldsymbol{T}}_4$ is equal to m^2+n^2 and the gradient of $J(\boldsymbol{x})$ is found to be

$$\begin{aligned}\frac{\partial J(\boldsymbol{x})}{\partial t_{ij}} &= \lim_{\Delta\to 0}\frac{J(\hat{\boldsymbol{T}}_{ij}) - J(\hat{\boldsymbol{T}})}{\Delta} \\ &= 2\boldsymbol{e}_j^T\hat{\boldsymbol{T}}\,[(1-\mu)\hat{\boldsymbol{C}} + \mu\hat{\boldsymbol{W}}_o]\hat{\boldsymbol{T}}^T\hat{\boldsymbol{T}}\boldsymbol{g}_{ij}\end{aligned} \tag{30}$$

for either $(1,1) \le (i,j) \le (m,m)$ or $(m+1,m+1) \le (i,j) \le (m+n,m+n)$ where $\hat{\boldsymbol{T}}_{ij}$ is the matrix obtained from $\hat{\boldsymbol{T}}$ with a perturbed (i,j)th component, which is given by [34, p.655]

$$\hat{\boldsymbol{T}}_{ij} = \hat{\boldsymbol{T}} + \frac{\Delta\hat{\boldsymbol{T}}\boldsymbol{g}_{ij}\boldsymbol{e}_j^T\hat{\boldsymbol{T}}}{1 - \Delta\boldsymbol{e}_j^T\hat{\boldsymbol{T}}\boldsymbol{g}_{ij}},$$

and $\boldsymbol{g}_{ij}$ is computed using

$$\boldsymbol{g}_{ij} = \partial\left\{\frac{\boldsymbol{t}_j}{||\boldsymbol{t}_j||}\right\}/\partial t_{ij} = \frac{1}{||\boldsymbol{t}_j||^3}(t_{ij}\boldsymbol{t}_j - ||\boldsymbol{t}_j||^2\boldsymbol{e}_i),$$

with

$$[\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_{m+n}] = [\boldsymbol{t}_{11}, \boldsymbol{t}_{12}, \cdots, \boldsymbol{t}_{1m}] \oplus [\boldsymbol{t}_{41}, \boldsymbol{t}_{42}, \cdots, \boldsymbol{t}_{4n}].$$

Case 2: *$\boldsymbol{D}$ is a block-diagonal matrix*

Matrix $\boldsymbol{D}$ in this case assumes the form

$$\boldsymbol{D} = \boldsymbol{D}_1 \oplus \boldsymbol{D}_4, \tag{31}$$

where $\boldsymbol{D}_1$ and $\boldsymbol{D}_4$ are $m\times m$ and $n\times n$ matrices, respectively. The gradient of $J(\boldsymbol{x})$ can be derived as follows:

$$\begin{aligned}\frac{\partial J(\boldsymbol{x})}{\partial t_{ij}} &= 2\beta_1 + 2(1-\mu)(\beta_2 - \beta_3) \\ \frac{\partial J(\boldsymbol{x})}{\partial d_{ij}} &= 2(1-\mu)\boldsymbol{e}_i^T\hat{\boldsymbol{T}}\hat{\boldsymbol{W}}_o(\hat{\boldsymbol{T}}^T\boldsymbol{D} - \hat{\boldsymbol{A}}\hat{\boldsymbol{T}}^T)\boldsymbol{e}_j,\end{aligned} \tag{32}$$

where

$$\begin{aligned}\beta_1 &= \boldsymbol{e}_j^T\hat{\boldsymbol{T}}\,[\,(1-\mu)(\hat{\boldsymbol{A}}^T\hat{\boldsymbol{W}}_o\hat{\boldsymbol{A}} + \hat{\boldsymbol{C}}) + \mu\hat{\boldsymbol{W}}_o]\hat{\boldsymbol{T}}^T\hat{\boldsymbol{T}}\boldsymbol{g}_{ij} \\ \beta_2 &= \boldsymbol{e}_j^T\hat{\boldsymbol{T}}\hat{\boldsymbol{W}}_o\hat{\boldsymbol{T}}^T\boldsymbol{D}\boldsymbol{D}^T\hat{\boldsymbol{T}}\boldsymbol{g}_{ij} \\ \beta_3 &= \boldsymbol{e}_j^T\hat{\boldsymbol{T}}\,(\hat{\boldsymbol{A}}^T\hat{\boldsymbol{W}}_o\hat{\boldsymbol{T}}^T\boldsymbol{D} + \hat{\boldsymbol{W}}_o\hat{\boldsymbol{A}}\hat{\boldsymbol{T}}^T\boldsymbol{D}^T)\hat{\boldsymbol{T}}\boldsymbol{g}_{ij},\end{aligned}$$

with $\boldsymbol{g}_{ij}$ defined in (30). In (32), $d_{ij} \in \boldsymbol{D}_1 \oplus \boldsymbol{D}_4$ is meant to be $d_{ij} \in \boldsymbol{D}_1$ for $(1,1) \le (i,j) \le (m,m)$ and $d_{ij} \in \boldsymbol{D}_4$ for $(m+1,m+1) \le (i,j) \le (m+n,m+n)$.

Case 3: *$\boldsymbol{D}$ is a diagonal matrix*

Here, matrix $\boldsymbol{D}$ assumes the form

$$\boldsymbol{D} = \mathrm{diag}\{d_{11}, d_{22}, \cdots, d_{m+n,m+n}\}. \tag{33}$$

In this case, $\partial J(\boldsymbol{x})/\partial d_{ij}$ can be obtained using (32) as

$$\frac{\partial J(\boldsymbol{x})}{\partial d_{ii}} = 2(1-\mu)\boldsymbol{e}_i^T\hat{\boldsymbol{T}}\hat{\boldsymbol{W}}_o(\hat{\boldsymbol{T}}^T\boldsymbol{D} - \hat{\boldsymbol{A}}\hat{\boldsymbol{T}}^T)\boldsymbol{e}_i, \tag{34}$$

where $1 \le i \le m+n$, and $\partial J(\boldsymbol{x})/\partial t_{ij}$ is also given by (32).

Case 4: *$\boldsymbol{D}$ is a block-scalar matrix*

It is assumed here that $\boldsymbol{D}_1 = \alpha\boldsymbol{I}_m$ and $\boldsymbol{D}_4 = \beta\boldsymbol{I}_n$ with scalars α and β. The gradient of $J(\boldsymbol{x})$ can then be calculated using

$$\begin{aligned}\frac{\partial J(\boldsymbol{x})}{\partial \alpha} &= 2(1-\mu)\sum_{i=1}^{m}\boldsymbol{e}_i^T\hat{\boldsymbol{T}}\hat{\boldsymbol{W}}_o(\hat{\boldsymbol{T}}^T\boldsymbol{D} - \hat{\boldsymbol{A}}\hat{\boldsymbol{T}}^T)\boldsymbol{e}_i \\ \frac{\partial J(\boldsymbol{x})}{\partial \beta} &= 2(1-\mu)\sum_{i=1}^{n}\boldsymbol{e}_{m+i}^T\hat{\boldsymbol{T}}\hat{\boldsymbol{W}}_o(\hat{\boldsymbol{T}}^T\boldsymbol{D} - \hat{\boldsymbol{A}}\hat{\boldsymbol{T}}^T)\boldsymbol{e}_{m+i}\end{aligned} \tag{35}$$

and $\partial J(\boldsymbol{x})/\partial t_{ij}$ is computed using (32).

IV. A CASE STUDY

In this section, we present a case study to illustrate the effectiveness of the proposed algorithm. Consider a 2-D BIBO stable, separately locally controllable, and separately locally observable state-space digital filter $(\boldsymbol{A}^o, \boldsymbol{b}^o, \boldsymbol{c}^o, d)_{2,2}$ of order (2,2) where

$$\begin{aligned}\boldsymbol{A}^o &= \begin{bmatrix} 1.88899 & -0.91219 & -1.00000 & 0.00000 \\ 1.00000 & 0.00000 & 0.00000 & 0.00000 \\ 0.02771 & -0.02580 & 1.88899 & 1.00000 \\ -0.02580 & 0.02431 & -0.91219 & 0.00000 \end{bmatrix} \\ \boldsymbol{b}^o &= \begin{bmatrix} 0.219089 & 0.000000 & -0.028889 & 0.091219 \end{bmatrix}^T \\ \boldsymbol{c}^o &= \begin{bmatrix} 0.028889 & -0.091219 & -0.219089 & 0.000000 \end{bmatrix} \\ d &= 0.08900.\end{aligned}$$

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

$$\boldsymbol{T}^o = \begin{bmatrix} -1.373341 & 9.544965 \\ -3.318699 & 9.494676 \end{bmatrix} \oplus \begin{bmatrix} 0.942406 & 0.329402 \\ -0.947397 & -0.136313 \end{bmatrix}$$

then the above filter is transformed to the *optimal realization* $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{2,2} = (\boldsymbol{T}^{o-1}\boldsymbol{A}^o\boldsymbol{T}^o, \boldsymbol{T}^{o-1}\boldsymbol{b}, \boldsymbol{c}\boldsymbol{T}^o, d)_{2,2}$ that satisfies (15) and (16) simultaneously [25],[26] where

$$\boldsymbol{A} = \begin{bmatrix} 0.923959 & -0.115198 & -0.480100 & -0.167811 \\ 0.178310 & 0.965031 & -0.167811 & -0.058655 \\ 0.045857 & 0.013210 & 0.923959 & 0.178310 \\ 0.013210 & 0.021491 & -0.115198 & 0.965031 \end{bmatrix}$$

$$\boldsymbol{b} = \begin{bmatrix} 0.111613 & 0.039012 & -0.142200 & 0.319129 \end{bmatrix}^T$$

$$\boldsymbol{c} = \begin{bmatrix} 0.263054 & -0.590350 & -0.206471 & -0.072168 \end{bmatrix}$$

$$d = 0.089000$$

and the local controllability and local observability Gramians were calculated by truncating the series in (12) and (9) to the range $(0, 0) \leq (i, j) \leq (200, 200)$ as

$$\boldsymbol{K}_c = \begin{bmatrix} 1.000000 & 0.221999 & 0.155751 & 0.036319 \\ 0.221999 & 1.000000 & 0.184141 & 0.064066 \\ 0.155751 & 0.184141 & 1.000000 & 0.221999 \\ 0.036319 & 0.064066 & 0.221999 & 1.000000 \end{bmatrix}$$

$$\boldsymbol{W}_o = \begin{bmatrix} 3.422064 & 0.759695 & 0.532989 & 0.630143 \\ 0.759695 & 3.422064 & 0.124286 & 0.219239 \\ 0.532989 & 0.124286 & 3.422064 & 0.759695 \\ 0.630143 & 0.219239 & 0.759695 & 3.422064 \end{bmatrix},$$

respectively. This gives the noise gain $I(\boldsymbol{0}) = \mathrm{tr}[\boldsymbol{W}_o] = 13.688256$. In what follows, EF and state-variable coordinate transformation are applied to the above *optimal realization* $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{2,2}$ in order to jointly minimize the roundoff noise, and the results obtained are then compared to their counterparts obtained in [29] where the minimization of the roundoff noise was carried out using EF and state-variable coordinate transformation, but in a *separate* manner.

Case 1: ***D*** *is a general matrix*

The quasi-Newton algorithm was applied to minimize (29) with $\mu = 0.01$ and tolerance $\varepsilon = 10^{-8}$. It took the algorithm 10 iterations to converge to the solution

$$\hat{\boldsymbol{T}} = \begin{bmatrix} 1.112303 & -0.262415 \\ 0.768079 & 0.846247 \end{bmatrix} \oplus \begin{bmatrix} 0.977230 & -0.434117 \\ 0.059862 & 1.067639 \end{bmatrix}$$

or equivalently,

$$\boldsymbol{T} = \begin{bmatrix} 1.076031 & 0.857797 \\ -0.136530 & 0.926745 \end{bmatrix} \oplus \begin{bmatrix} 0.922624 & 0.178741 \\ -0.322246 & 1.067644 \end{bmatrix}.$$

This leads to

$$\overline{\boldsymbol{A}} = \begin{bmatrix} 0.793657 & -0.235832 & -0.218781 & -0.149075 \\ 0.181787 & 1.095333 & -0.178900 & -0.121901 \\ 0.046747 & 0.047458 & 0.885610 & 0.190951 \\ 0.024675 & 0.043593 & -0.123522 & 1.003380 \end{bmatrix}$$

$$\overline{\boldsymbol{b}} = \begin{bmatrix} 0.062793 & 0.051347 & -0.200321 & 0.238447 \end{bmatrix}^T$$

$$\overline{\boldsymbol{c}} = \begin{bmatrix} 0.363655 & -0.321457 & -0.167239 & -0.113955 \end{bmatrix}$$

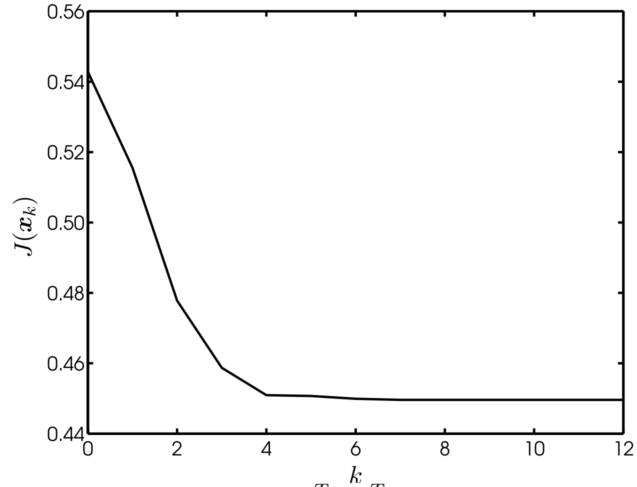

Fig. 1. Profile of $J(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{A}}\hat{\boldsymbol{T}}^T, \hat{\boldsymbol{T}})$ with $\mu = 0.01$ during the first 12 iterations.

$$\overline{\boldsymbol{K}}_c = \begin{bmatrix} 1.000000 & -0.484097 & -0.009234 & -0.020689 \\ -0.484097 & 1.000000 & 0.190252 & 0.119536 \\ -0.009234 & 0.190252 & 1.000000 & 0.354179 \\ -0.020689 & 0.119536 & 0.354179 & 1.000000 \end{bmatrix}$$

$$\overline{\boldsymbol{W}}_o = \begin{bmatrix} 3.802789 & 3.394235 & 0.304627 & 0.791440 \\ 3.394235 & 6.664921 & 0.288432 & 0.896328 \\ 0.304627 & 0.288432 & 2.816605 & 0.091564 \\ 0.791440 & 0.896328 & 0.091564 & 4.299965 \end{bmatrix}.$$

Using (28) and (29), the optimal EF matrix $\boldsymbol{D}$ and the noise gain in (18) were found to be

$$\boldsymbol{D} = \begin{bmatrix} 0.793657 & -0.235832 & -0.218781 & -0.149075 \\ 0.181787 & 1.095333 & -0.178900 & -0.121901 \\ 0.046747 & 0.047458 & 0.885610 & 0.190951 \\ 0.024675 & 0.043593 & -0.123522 & 1.003380 \end{bmatrix}$$

and $I(\boldsymbol{D}, \boldsymbol{T}) = 0.276534$, respectively. The profile of $J(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{A}}\hat{\boldsymbol{T}}^T, \hat{\boldsymbol{T}})$ with $\mu = 0.01$ in (29) during the first 12 iterations of the algorithm is depicted in Fig. 1.

Next, the above optimal EF matrix $\boldsymbol{D}$ was rounded to a power-of-two representation with 3 bits after the binary point, which resulted in

$$\boldsymbol{D}_{\text{3bit}} = \begin{bmatrix} 0.750 & -0.250 & -0.250 & -0.125 \\ 0.125 & 1.125 & -0.125 & -0.125 \\ 0.000 & 0.000 & 0.875 & 0.250 \\ 0.000 & 0.000 & -0.125 & 1.000 \end{bmatrix}.$$

The corresponding noise gain was found to be $I(\boldsymbol{D}_{\text{3bit}}, \boldsymbol{T}) = 0.379031$. Furthermore, when the optimal EF matrix $\boldsymbol{D}$ was rounded to the integer representation $\boldsymbol{D}_{\text{int}} = \mathrm{diag}\{1, 1, 1, 1\}$, the noise gain was found to be $I(\boldsymbol{D}_{\text{int}}, \boldsymbol{T}) = 1.786366$.

Case 2: ***D*** *is a block-diagonal matrix*

Again, the quasi-Newton algorithm was applied to minimize $J(\boldsymbol{D}, \hat{\boldsymbol{T}})$ in (25) with $\boldsymbol{D} = \boldsymbol{D}_1 \oplus \boldsymbol{D}_4$, $\mu = 0.01$, and $\varepsilon = 10^{-8}$. It took the algorithm 19 iterations to converge to the solution

$$\hat{\boldsymbol{T}} = \begin{bmatrix} 1.075413 & -0.290485 \\ 0.734598 & 0.837413 \end{bmatrix} \oplus \begin{bmatrix} 1.081669 & -1.093278 \\ -0.110922 & 1.533936 \end{bmatrix}$$

$$\boldsymbol{D} = \begin{bmatrix} 0.812641 & -0.217981 \\ 0.174373 & 1.086382 \end{bmatrix} \oplus \begin{bmatrix} 0.720185 & 0.234829 \\ -0.263724 & 1.077042 \end{bmatrix}.$$

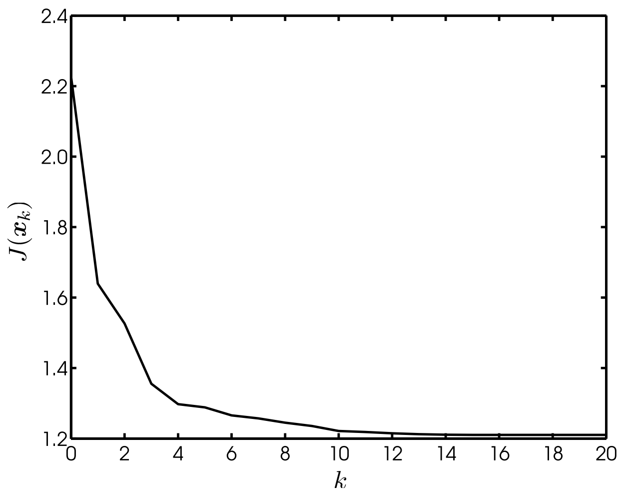

Fig. 2. Profile of $J(\boldsymbol{D}, \hat{\boldsymbol{T}})$ with $\mu = 0.01$ during the first 20 iterations.

This leads to

$$\boldsymbol{T} = \begin{bmatrix} 1.036236 & 0.823539 \\ -0.168545 & 0.914226 \end{bmatrix} \oplus \begin{bmatrix} 0.952782 & 0.061110 \\ -0.965616 & 1.511947 \end{bmatrix}$$

$$\overline{\boldsymbol{A}} = \begin{bmatrix} 0.805454 & -0.228456 & -0.170347 & -0.163237 \\ 0.172688 & 1.083536 & -0.144340 & -0.138315 \\ 0.045256 & 0.049009 & 0.756447 & 0.269578 \\ 0.035561 & 0.051491 & -0.205808 & 1.132543 \end{bmatrix}$$

$$\overline{\boldsymbol{b}} = \begin{bmatrix} 0.064366 & 0.054539 & -0.156380 & 0.111198 \end{bmatrix}^T$$

$$\overline{\boldsymbol{c}} = \begin{bmatrix} 0.372087 & -0.323078 & -0.127035 & -0.121732 \end{bmatrix}$$

$$\overline{\boldsymbol{K}}_c = \begin{bmatrix} 1.000000 & -0.440602 & -0.007858 & -0.016928 \\ -0.440602 & 1.000000 & 0.198776 & 0.171103 \\ -0.007858 & 0.198776 & 1.000000 & 0.759746 \\ -0.016928 & 0.171103 & 0.759746 & 1.000000 \end{bmatrix}$$

$$\overline{\boldsymbol{W}}_o = \begin{bmatrix} 3.506411 & 3.007275 & -0.088578 & 0.963868 \\ 3.007275 & 6.325040 & -0.168173 & 1.121432 \\ -0.088578 & -0.168173 & 4.899437 & -3.747273 \\ 0.963868 & 1.121432 & -3.747273 & 7.975946 \end{bmatrix}$$

and the minimized noise gain was found to be $I(\boldsymbol{D}, \boldsymbol{T}) = 0.993119$ from (18). The profile of $J(\boldsymbol{D}, \hat{\boldsymbol{T}})$ with $\mu = 0.01$ in (25) during the first 20 iterations of the algorithm is shown in Fig. 2.

Next, the optimal EF matrix $\boldsymbol{D} = \boldsymbol{D}_1 \oplus \boldsymbol{D}_4$ was rounded to a power-of-two representation with 3 bits after the binary point to yield

$$\boldsymbol{D}_{\text{3bit}} = \begin{bmatrix} 0.875 & -0.250 \\ 0.125 & 1.125 \end{bmatrix} \oplus \begin{bmatrix} 0.750 & 0.250 \\ -0.250 & 1.125 \end{bmatrix},$$

which leads to a noise gain $I(\boldsymbol{D}_{\text{3bit}}, \boldsymbol{T}) = 1.026055$. Furthermore, the optimal EF matrix $\boldsymbol{D} = \boldsymbol{D}_1 \oplus \boldsymbol{D}_4$ was rounded to the integer representation $\boldsymbol{D}_{\text{int}} = \text{diag}\{1, 1, 1, 1\}$ and the corresponding noise gain was found to be $I(\boldsymbol{D}_{\text{int}}, \boldsymbol{T}) = 1.779801$.

Case 3: ***D*** *is a diagonal matrix*

The quasi-Newton algorithm with $\mu = 0.0$ and $\varepsilon = 10^{-8}$ was applied to minimize (25) for a diagonal EF matrix $\boldsymbol{D}$. It took the algorithm 14 iterations to converge to the solution

$$\hat{\boldsymbol{T}} = \begin{bmatrix} 1.001398 & -0.305076 \\ 0.587614 & 0.866360 \end{bmatrix} \oplus \begin{bmatrix} 0.930738 & -0.766589 \\ 0.115699 & 1.200227 \end{bmatrix}$$

$$\boldsymbol{D} = \text{diag}\{0.959461,\ 0.979277,\ 0.896380,\ 0.950455\},$$

which leads to

$$\boldsymbol{T} = \begin{bmatrix} 0.961055 & 0.680708 \\ -0.191312 & 0.926574 \end{bmatrix} \oplus \begin{bmatrix} 0.839287 & 0.249038 \\ -0.657829 & 1.205640 \end{bmatrix}$$

$$\overline{\boldsymbol{A}} = \begin{bmatrix} 0.834922 & -0.203220 & -0.197375 & -0.217164 \\ 0.158082 & 1.054068 & -0.151112 & -0.166263 \\ 0.040783 & 0.038439 & 0.829877 & 0.216040 \\ 0.029372 & 0.044948 & -0.153937 & 1.059113 \end{bmatrix}$$

$$\overline{\boldsymbol{b}} = \begin{bmatrix} 0.075302 & 0.057652 & -0.213419 & 0.148249 \end{bmatrix}^T$$

$$\overline{\boldsymbol{c}} = \begin{bmatrix} 0.365751 & -0.367940 & -0.125814 & -0.138428 \end{bmatrix}$$

$$\overline{\boldsymbol{K}}_c = \begin{bmatrix} 1.000000 & -0.295774 & 0.021123 & 0.003433 \\ -0.295774 & 1.000000 & 0.193509 & 0.161263 \\ 0.021123 & 0.193509 & 1.000000 & 0.558757 \\ 0.003433 & 0.161263 & 0.558757 & 1.000000 \end{bmatrix}$$

$$\overline{\boldsymbol{W}}_o = \begin{bmatrix} 3.006599 & 2.209658 & 0.039163 & 0.801213 \\ 2.209658 & 5.481950 & -0.014649 & 0.881098 \\ 0.039163 & -0.014649 & 3.052508 & -1.354534 \\ 0.801213 & 0.881098 & -1.354534 & 5.642635 \end{bmatrix},$$

and the minimized noise gain was found to be $I(\boldsymbol{D}, \boldsymbol{T}) = 1.608812$ from (18). The profile of $J(\boldsymbol{D}, \hat{\boldsymbol{T}})$ with $\mu = 0.0$ in (25) during the first 16 iterations of the algorithm is shown in Fig. 3.

Next, the above optimal diagonal EF matrix $\boldsymbol{D}$ was rounded to a power-of-two representation with 3 bits after the binary point to yield $\boldsymbol{D}_{\text{3bit}} = \text{diag}\{1.000, 1.000, 0.875, 1.000\}$, which leads to a noise gain $I(\boldsymbol{D}_{\text{3bit}}, \boldsymbol{T}) = 1.631354$. Furthermore, when the optimized diagonal EF matrix $\boldsymbol{D}$ was rounded to the integer representation $\boldsymbol{D}_{\text{int}} = \text{diag}\{1, 1, 1, 1\}$, the noise gain was found to be $I(\boldsymbol{D}_{\text{int}}, \boldsymbol{T}) = 1.662735$.

Case 4: ***D*** *is a block-scalar matrix*

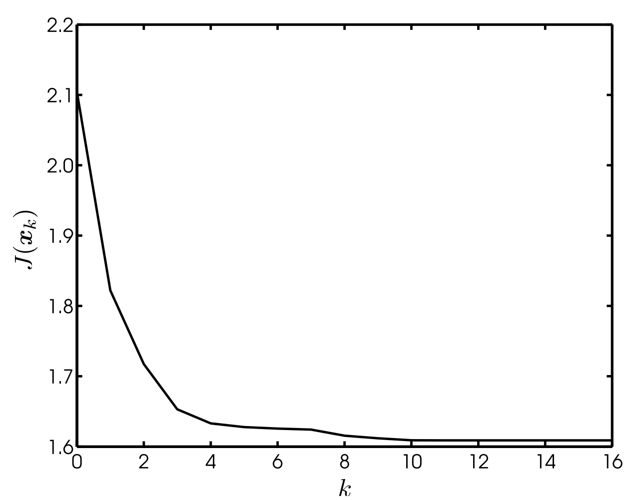

Fig. 3. Profile of $J(\boldsymbol{D}, \hat{\boldsymbol{T}})$ with $\mu = 0.0$ during the first 16 iterations.

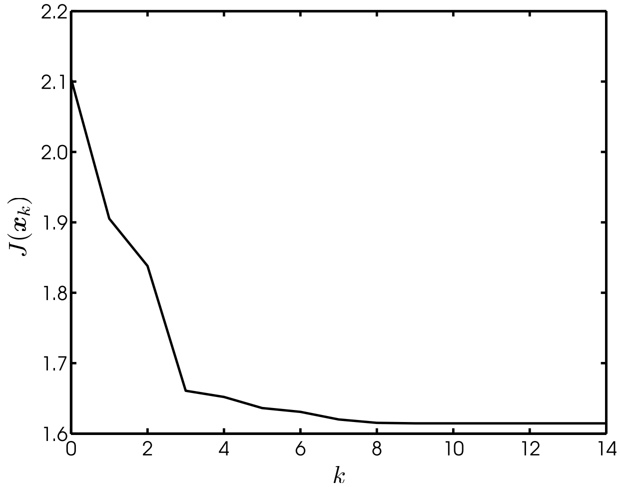

Fig. 4. Profile of $J(\boldsymbol{D}, \hat{\boldsymbol{T}})$ with $\mu = 0.0$ during the first 14 iterations.

In this case, the quasi-Newton algorithm with $\mu = 0.0$ and $\varepsilon = 10^{-8}$ was applied to minimize (25) for $\boldsymbol{D} = \alpha \boldsymbol{I}_2 \oplus \beta \boldsymbol{I}_2$ with scalars α and β. The algorithm converges after 12 iterations to the solution

$$\hat{\boldsymbol{T}} = \begin{bmatrix} 1.009533 & -0.279518 \\ 0.567440 & 0.880511 \end{bmatrix} \oplus \begin{bmatrix} 0.917919 & -0.788744 \\ 0.134695 & 1.202726 \end{bmatrix}$$

$$\alpha = 0.972437, \qquad \beta = 0.932446,$$

which leads to

$$\boldsymbol{T} = \begin{bmatrix} 0.971994 & 0.662241 \\ -0.165006 & 0.938383 \end{bmatrix} \oplus \begin{bmatrix} 0.824073 & 0.268195 \\ -0.681278 & 1.210245 \end{bmatrix}$$

$$\overline{\boldsymbol{A}} = \begin{bmatrix} 0.833441 & -0.200869 & -0.194700 & -0.229679 \\ 0.161558 & 1.055549 & -0.139020 & -0.163997 \\ 0.041366 & 0.037287 & 0.827306 & 0.217046 \\ 0.030965 & 0.044882 & -0.155969 & 1.061684 \end{bmatrix}$$

$$\overline{\boldsymbol{b}} = \begin{bmatrix} 0.077249 & 0.055157 & -0.218369 & 0.140763 \end{bmatrix}^T$$

$$\overline{\boldsymbol{c}} = \begin{bmatrix} 0.353098 & -0.379770 & -0.120980 & -0.142716 \end{bmatrix}$$

$$\overline{\boldsymbol{K}}_c = \begin{bmatrix} 1.000000 & -0.297762 & 0.026165 & 0.007977 \\ -0.297762 & 1.000000 & 0.190338 & 0.162372 \\ 0.026165 & 0.190338 & 1.000000 & 0.563261 \\ 0.007977 & 0.162372 & 0.563261 & 1.000000 \end{bmatrix}$$

$$\overline{\boldsymbol{W}}_o = \begin{bmatrix} 3.082557 & 2.282800 & 0.017388 & 0.830929 \\ 2.282800 & 5.458336 & -0.037480 & 0.879969 \\ 0.017388 & -0.037480 & 3.059210 & -1.446363 \\ 0.830929 & 0.879969 & -1.446363 & 5.751581 \end{bmatrix},$$

and the minimized noise gain was found to be $I(\boldsymbol{D}, \boldsymbol{T}) = 1.614538$ from (18). The profile of $J(\boldsymbol{D}, \hat{\boldsymbol{T}})$ with $\mu = 0.0$ in (25) during the first 14 iterations of the algorithm is drawn in Fig. 4.

Next, the optimal EF matrix $\boldsymbol{D} = \alpha \boldsymbol{I}_2 \oplus \beta \boldsymbol{I}_2$ was rounded to a power-of-two representation with 3 bits after the binary point as well as an integer representation. It was found that these representations were given by $\boldsymbol{D}_{\text{3bit}} = \text{diag}\{1.000, 1.000, 0.875, 0.875\}$ and $\boldsymbol{D}_{\text{int}} = \text{diag}\{1, 1, 1, 1\}$, respectively. The corresponding noise gains were obtained as $I(\boldsymbol{D}_{\text{3bit}}, \boldsymbol{T}) = 1.650103$ and $I(\boldsymbol{D}_{\text{int}}, \boldsymbol{T}) = 1.661235$, respectively. It is interesting to note that for this particular example the noise gain obtained from the integer approximation of the optimal matrix $\boldsymbol{D} = \alpha \boldsymbol{I}_m \oplus \beta \boldsymbol{I}_n$ is smaller than that obtained from the integer approximation of the optimal diagonal EF matrix $\boldsymbol{D}$, due to their different $\hat{\boldsymbol{T}}$ matrices.

The simulation results described above are summarized using the noise gain $I(\boldsymbol{D}, \boldsymbol{T})$ in (18) in Table I. For comparison purposes, their counterparts obtained using the method in [29] are also included in the Table. Specifically, the term "separate" means that the EF matrix was optimized by applying the existing method [29] to the optimal realization without EF, which satisfies (15) and (16) simultaneously [25],[26]. From the Table, it is observed that the proposed joint optimization offers greatly reduced roundoff noise gain for all cases of the matrix $\boldsymbol{D}$ when compared with that obtained by using *separate* optimization.

V. CONCLUSION

The joint optimization problem of EF and realization to minimize the effects of roundoff noise of 2-D state-space digital filters subject to L_2-norm dynamic-range scaling constraints has been investigated. It has been shown that the problem at hand can be converted into an unconstrained optimization problem by using linear algebraic techniques. Closed-form formulas for fast evaluation of the gradient of the objective function have been derived and an efficient quasi-Newton algorithm has been employed to solve the unconstrained optimization problem. The proposed technique has been applied to the cases where the EF matrix is a general, block-diagonal, diagonal, or block-scalar matrix, and its effectiveness compared with the existing method [29] has been demonstrated by a case study.

REFERENCES

[1] 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.

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

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

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

[5] 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. 30, pp. 963-973, Dec. 1982.

[6] M. Renfors, "Roundoff noise in error-feedback state-space filters," *Proc. Int. Conf. Acoustics, Speech, Signal Processing* (ICASSP'83), pp. 619-622, Apr. 1983.

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

[8] 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.

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

[10] D. Williamson, "Roundoff noise minimization and pole-zero sensitivity in fixed-point digital filters using residue feedback," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, pp. 1210-1220, Oct. 1986.

TABLE I
PERFORMANCE COMPARISON

Matrix $\boldsymbol{D}$	Optimization	Accuracy of $\boldsymbol{D}$		
		Infinite Precision	3-Bit Quantization	Integer Quantization
Null	Separate	13.688256		
General	Separate	0.465549	0.555529	2.040208
	Joint	0.276534	0.379031	1.786366
Block-Diagonal	Separate	1.555329	1.612408	2.040208
	Joint	0.993119	1.026055	1.779801
Diagonal	Separate	1.908903	1.937559	2.040208
	Joint	1.608812	1.631354	1.662735
Block-Scalar	Separate	1.950396	1.965326	2.040208
	Joint	1.614538	1.650103	1.661235

[11] T. Hinamoto, S. Karino and N. Kuroda, “Error spectrum shaping in 2-D digital filters,” *Proc. IEEE Int. Symp. Circuits Syst.* (ISCAS’95), vol. 1, pp. 348-351, May 1995.

[12] P. Agathoklis and C. Xiao, “Low roundoff noise structures for 2-D filters,” *Proc. IEEE Int. Symp. Circuits Syst.* (ISCAS’96), vol. 2, pp. 352-355, May 1996.

[13] T. Hinamoto, S. Karino and N. Kuroda, “2-D state-space digital filters with error spectrum shaping,” *Proc. IEEE Int. Symp. Circuits Syst.* (ISCAS’96), vol. 2, pp. 766-769, May 1996.

[14] 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.* (ISCAS’97), vol. 4, pp. 2461-2464, June 1997.

[15] 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.

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

[17] D. Williamson, “Delay replacement in direct form structures”, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 453-460, Apr. 1988.

[18] 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.* (ISCAS’96), vol. 2, pp. 213-216, May 1996.

[19] G. Li and Z. Zhao, “On the generalized DFIIt structure and its state-space realization in digital filter implementation,” *IEEE Trans. Circuits Syst. I*, vol. 51, pp. 769-778, Apr. 2004.

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

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

[22] S. Y. Hwang, “Minimum uncorrelated unit noise in state-space digital filtering,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 25, pp. 273-281, Aug. 1977.

[23] 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. 26, pp. 149-153, Mar. 1979.

[24] 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. 33, pp. 365-372, Apr. 1986.

[25] M. Kawamata and T. Higuchi, “A unified study on the roundoff noise in 2-D state-space digital filters,” *IEEE Trans. Circuits Syst.*, vol. 33, pp. 724-730, July 1986.

[26] 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. 33, pp. 965-973, Oct. 1986.

[27] 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.

[28] T. Hinamoto, H. Ohnishi and W.-S. Lu, “Roundoff noise minimization of state-space digital filters using separate and joint error feedback/coordinate transformation,” *IEEE Trans. Circuits Syst. I*, vol. 50, pp. 23-33, Jan. 2003.

[29] T. Hinamoto, K. Higashi and W.-S. Lu, “Separate/joint optimization of error feedback and coordinate transformation for roundoff noise minimization in two-dimensional state-space digital filters,” *IEEE Trans. Signal Processing*, vol. 51, pp. 2436-2445, Sept. 2003.

[30] W.-S. Lu and T. Hinamoto, “Jointly optimized error-feedback and realization for roundoff noise minimization in state-space digital filters,” *IEEE Trans. Signal Processing*, vol. 53, pp. 2135-2145, June 2005.

[31] R. Fletcher, Practical Methods of Optimization, 2nd ed. Wiley, New York, 1987.

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

[33] 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.

[34] T. Kailath, *Linear Systems*, Prentice Hall, 1980.