Separate/Joint Optimization of Error Feedback and Realization for Roundoff Noise Minimization in a Class of 2-D State-Space Digital Filters

TAKAO HINAMOTO hinamoto@hiroshima-u.ac.jp
Graduate School of Engineering, Hiroshima University, 1-4-1 Kagamiyama, Higashi-Hiroshima 739-8527, Japan

TORU OUMI oumi@hiroshima-u.ac.jp
Graduate School of Engineering, Hiroshima University, 1-4-1 Kagamiyama, Higashi-Hiroshima 739-8527, Japan

WU-SHENG LU wslu@ece.uvic.ca
Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada V8W 3P6

Received

Abstract. Techniques for the separate/joint optimization of error-feedback and realization are developed to minimize the roundoff noise subject to l_2-norm dynamic-range scaling constraints for a class of 2-D state-space digital filters. In the joint optimization, the problem at hand is converted into an unconstrained optimization problem by using linear-algebraic techniques. The unconstrained problem obtained is then solved by applying an efficient quasi-Newton algorithm. A numerical example is presented to illustrate the utility of the proposed techniques.

Key Words: 2-D IIR digital filters, Fornasini-Marchesini's second local state-space model, roundoff noise minimization, error feedback, coordinate transformation, separate/joint optimization, l_2-scaling constraints, no overflow oscillations

1. 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 is a useful tool for reducing finite-word-length (FWL) effects in recursive digital filters. Many error-feedback techniques have been proposed for 2-D recursive digital filters [1]-[5]. Another useful approach is to construct the 2-D state-space filter structure for the roundoff noise gain to be minimized by applying a linear transformation to the state-space coordinates subject to l_2-norm dynamic-range scaling constraints [6],[7]. As a natural extension of the fore-mentioned methods, efforts have been made to develop new methods that combine error feedback and coordinate transformation for better performance in the roundoff noise reduction. In [8],[9] separately/jointly-optimized iterative algorithms have been developed for 2-D filters described by the Roesser local state-space model with error-feedback matrix.

This paper investigates the problems of separately/jointly optimizing error feedback and realization subject to l_2-norm dynamic-range scaling constraints for 2-D state-space digital filters described by the Fornasini-Marchesini second local state-space model. The former is resolved by deriving analytical formulas in closed form while the latter is addressed by developing an iterative method based on an efficient quasi-Newton algorithm [10]. In this paper, error feedforward is also considered for the 2-D state-space digital filters. Computer simulation results demonstrate the validity of the proposed techniques.

2. Roundoff Noise Analysis and Scaling

Consider a 2-D recursive digital filter that is described by the Fornasini-Marchesini second local state-space (LSS) model $(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{c}, d)_n$ [11]:

$$\begin{aligned} \boldsymbol{x}(i,\,j) &= \boldsymbol{A}_1\boldsymbol{x}(i-1,\,j) + \boldsymbol{A}_2\boldsymbol{x}(i,\,j-1) + \boldsymbol{b}_1u(i-1,\,j) + \boldsymbol{b}_2u(i,\,j-1) \\ y(i,\,j) &= \boldsymbol{c}\boldsymbol{x}(i,\,j) + du(i,\,j) \end{aligned} \tag{1}$$

where $\boldsymbol{x}(i,j)$ is an $n \times 1$ local state vector, $u(i,j)$ is a scalar input, $y(i,j)$ is a scalar output, and $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, $\boldsymbol{c}$, and d are real matrices of appropriate dimensions. The LSS model in (1) is assumed to be stable, locally controllable and locally observable.

Due to finite register sizes, FWL constraints are imposed on the local state vector, input, output, and coefficients in the realization $(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{c}, d)_n$. Assuming that the quantization is carried out before matrix-vector multiplication, the actual 2-D FWL filter of (1) with EF and error feedforward can be implemented as

$$\begin{aligned}\tilde{\boldsymbol{x}}(i,\, j) &= \boldsymbol{A}_1\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i-1,\, j)] + \boldsymbol{A}_2\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,\, j-1)] \\ &\qquad + \boldsymbol{b}_1 u(i-1,\, j) + \boldsymbol{b}_2 u(i,\, j-1) + \boldsymbol{D}_1\boldsymbol{e}(i-1,\, j) + \boldsymbol{D}_2\boldsymbol{e}(i,\, j-1) \\ \tilde{y}(i,\, j) &= \boldsymbol{c}\,\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,\, j)] + du(i,\, j) + \boldsymbol{h}\boldsymbol{e}(i,\, j)\end{aligned} \tag{2}$$

where $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are $n \times n$ EF matrices, $\boldsymbol{h}$ is a $1 \times n$ error-feedforward vector,

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

and each component of matrices $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, $\boldsymbol{c}$ and d assumes an exact fractional B_c-bit representation. The FWL local state vector $\tilde{\boldsymbol{x}}(i,j)$ and 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 the multiplications and additions, where the sign bit is not counted. The quantization error $\boldsymbol{e}(i,j)$ is modeled as a zero-mean noise process with covariance $\sigma^2\boldsymbol{I}_n$ where

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

Subtracting (2) from (1) yields

$$\begin{aligned}\Delta\boldsymbol{x}(i,\, j) &= \boldsymbol{A}_1\Delta\boldsymbol{x}(i-1,\, j) + \boldsymbol{A}_2\Delta\boldsymbol{x}(i,\, j-1) + (\boldsymbol{A}_1 - \boldsymbol{D}_1)\boldsymbol{e}(i-1,\, j) + (\boldsymbol{A}_2 - \boldsymbol{D}_2)\boldsymbol{e}(i,\, j-1) \\ \Delta y(i,\, j) &= \boldsymbol{c}\Delta\boldsymbol{x}(i,\, j) + (\boldsymbol{c} - \boldsymbol{h})\boldsymbol{e}(i,\, j)\end{aligned} \tag{3}$$

where

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

By applying 2-D z transform to (3), the transfer function from the quantization error $\boldsymbol{e}(i,j)$ to the filter output $\Delta y(i,j)$ is obtained as

$$\boldsymbol{G}(z_1, z_2) = \boldsymbol{c}\left(\boldsymbol{I}_n - z_1^{-1}\boldsymbol{A}_1 - z_2^{-1}\boldsymbol{A}_2\right)^{-1}\left[z_1^{-1}(\boldsymbol{A}_1 - \boldsymbol{D}_1) + z_2^{-1}(\boldsymbol{A}_2 - \boldsymbol{D}_2)\right] + \boldsymbol{c} - \boldsymbol{h}. \tag{4}$$

For the 2-D filter in (2), the noise gain $I(\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{h}) = \sigma_{out}^2/\sigma^2$ can be evaluated by

$$I(\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{h}) = \mathrm{tr}\left[\frac{1}{(2\pi j)^2}\oint_{\Gamma_1}\oint_{\Gamma_2}\boldsymbol{G}^*(z_1, z_2)\boldsymbol{G}(z_1, z_2)\frac{dz_1 dz_2}{z_1 z_2}\right] \tag{5}$$

where σ_{out}^2 denotes noise variance at the output, and $\Gamma_i = \{z_i : |z_i| = 1\}$ for $i = 1, 2$.

Now if we express

$$\left(\boldsymbol{I}_n - z_1^{-1}\boldsymbol{A}_1 - z_2^{-1}\boldsymbol{A}_2\right)^{-1} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{A}^{(i,\, j)}z_1^{-i}z_2^{-j} \tag{6}$$

then the transition matrix $\boldsymbol{A}^{(i,\, j)}$ with $i, j \geq 0$ can be evaluated as follows:

$$\begin{aligned}\boldsymbol{A}^{(0,\, 0)} &= \boldsymbol{I}_n, \qquad \boldsymbol{A}^{(i,\, j)} = \boldsymbol{0} \quad \text{for } i < 0 \text{ or } j < 0 \\ \boldsymbol{A}^{(i,\, j)} &= \boldsymbol{A}_1\boldsymbol{A}^{(i-1,\, j)} + \boldsymbol{A}_2\boldsymbol{A}^{(i,\, j-1)} \\ &= \boldsymbol{A}^{(i-1,\, j)}\boldsymbol{A}_1 + \boldsymbol{A}^{(i,\, j-1)}\boldsymbol{A}_2 \quad \text{for } i, j > 0.\end{aligned} \tag{7}$$

Substituting (6) into (4) yields

$$\boldsymbol{G}(z_1, z_2) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\left[\boldsymbol{c}\boldsymbol{A}^{(i-1,\, j)}(\boldsymbol{A}_1 - \boldsymbol{D}_1) + \boldsymbol{c}\boldsymbol{A}^{(i,\, j-1)}(\boldsymbol{A}_2 - \boldsymbol{D}_2)\right]z_1^{-i}z_2^{-j} + \boldsymbol{c} - \boldsymbol{h}. \tag{8}$$

This leads the noise gain in (5) to

$$I(\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{h}) = J_1(\boldsymbol{D}_1, \boldsymbol{D}_2) + \mathrm{tr}[(\boldsymbol{c} - \boldsymbol{h})^T(\boldsymbol{c} - \boldsymbol{h})] \tag{9}$$

where

$$\begin{aligned} J_1(\boldsymbol{D}_1, \boldsymbol{D}_2) &= \mathrm{tr}\Big[[(\boldsymbol{A}_1 - \boldsymbol{D}_1)^T,\, (\boldsymbol{A}_2 - \boldsymbol{D}_2)^T]\boldsymbol{W}'\begin{bmatrix} \boldsymbol{A}_1 - \boldsymbol{D}_1 \\ \boldsymbol{A}_2 - \boldsymbol{D}_2 \end{bmatrix}\Big] \\ &= \mathrm{tr}\Big[(\boldsymbol{A}_1 - \boldsymbol{D}_1)^T\boldsymbol{W}_o(\boldsymbol{A}_1 - \boldsymbol{D}_1) + (\boldsymbol{A}_2 - \boldsymbol{D}_2)^T\boldsymbol{W}^T(\boldsymbol{A}_1 - \boldsymbol{D}_1) \\ &\qquad + (\boldsymbol{A}_1 - \boldsymbol{D}_1)^T\boldsymbol{W}(\boldsymbol{A}_2 - \boldsymbol{D}_2) + (\boldsymbol{A}_2 - \boldsymbol{D}_2)^T\boldsymbol{W}_o(\boldsymbol{A}_2 - \boldsymbol{D}_2)\Big]. \end{aligned}$$

Here, the $2n \times 2n$ matrix $\boldsymbol{W}'$ is defined by

$$\boldsymbol{W}' = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\begin{bmatrix} (\boldsymbol{c}\boldsymbol{A}^{(i-1,\,j)})^T \\ (\boldsymbol{c}\boldsymbol{A}^{(i,\,j-1)})^T \end{bmatrix}\Big[\boldsymbol{c}\boldsymbol{A}^{(i-1,\,j)},\ \boldsymbol{c}\boldsymbol{A}^{(i,\,j-1)}\Big] = \begin{bmatrix} \boldsymbol{W}_o & \boldsymbol{W} \\ \boldsymbol{W}^T & \boldsymbol{W}_o \end{bmatrix} \tag{10}$$

where matrix $\boldsymbol{W}_o$ is the local observability Gramian of the LSS model in (1). In the case when there is no EF but error feedforward exists, it follows from (9) that

$$\begin{aligned} I(\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{c}) &= \mathrm{tr}\Big[[\boldsymbol{A}_1^T,\, \boldsymbol{A}_2^T]\boldsymbol{W}'\begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{bmatrix}\Big] \\ &= \mathrm{tr}\Big[\boldsymbol{A}_1^T\boldsymbol{W}_o\boldsymbol{A}_1 + \boldsymbol{A}_2^T\boldsymbol{W}^T\boldsymbol{A}_1 + \boldsymbol{A}_1^T\boldsymbol{W}\boldsymbol{A}_2 + \boldsymbol{A}_2^T\boldsymbol{W}_o\boldsymbol{A}_2\Big]. \end{aligned} \tag{11}$$

The local controllability Gramian $\boldsymbol{K}_c$ is defined by

$$\boldsymbol{K}_c = \sum_{k=1}^{\infty}\sum_{i=0}^{k}\boldsymbol{f}(i,\, k-i)\boldsymbol{f}^T(i,\, k-i) \tag{12}$$

where

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

In order to prevent overflow in the 2-D state-space digital filter modeled by (1), signal scaling is applied based on the l_2 norm [7], which accordingly gives the following set of dynamic-range scaling constraints on the state variables:

$$(\boldsymbol{K}_c)_{ii} = 1 \quad \text{for} \quad i = 1, 2, \cdots, n \tag{13}$$

where $(\boldsymbol{K}_c)_{ii}$ denotes the ith entry of matrix $\boldsymbol{K}_c$.

With the above preliminaries, we are now in a position to present our solution methods for the roundoff noise minimization problems at hand.

3. Separate Optimization of Realization and Error Feedback

Applying a coordinate transformation defined by

$$\overline{\boldsymbol{x}}(i,\, j) = \boldsymbol{T}^{-1}\boldsymbol{x}(i,\, j) \tag{14}$$

to the LSS model $(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{c}, d)_n$ in (1), we obtain a new realization $(\overline{\boldsymbol{A}}_1, \overline{\boldsymbol{A}}_2, \overline{\boldsymbol{b}}_1, \overline{\boldsymbol{b}}_2, \overline{\boldsymbol{c}}, d)_n$ characterized by

$$\overline{\boldsymbol{A}}_1 = \boldsymbol{T}^{-1}\boldsymbol{A}_1\boldsymbol{T}, \qquad \overline{\boldsymbol{A}}_2 = \boldsymbol{T}^{-1}\boldsymbol{A}_2\boldsymbol{T}, \qquad \overline{\boldsymbol{b}}_1 = \boldsymbol{T}^{-1}\boldsymbol{b}_1, \qquad \overline{\boldsymbol{b}}_2 = \boldsymbol{T}^{-1}\boldsymbol{b}_2, \qquad \overline{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T} \tag{15}$$

where $\boldsymbol{T}$ is an $n \times n$ nonsingular matrix. The Gramians $\overline{\boldsymbol{K}}_c$, $\overline{\boldsymbol{W}}_o$ and $\overline{\boldsymbol{W}}$ in the new realization can then be written as

$$\overline{\boldsymbol{K}}_c = \boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T}, \qquad \overline{\boldsymbol{W}}_o = \boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{T}, \qquad \overline{\boldsymbol{W}} = \boldsymbol{T}^T\boldsymbol{W}\boldsymbol{T} \tag{16}$$

respectively. If the l_2-norm dynamic-range scaling constraints are imposed on the new realization, then we have

$$(\overline{\boldsymbol{K}}_c)_{ii} = (\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T})_{ii} = 1 \quad \text{for} \quad i = 1, 2, \cdots, n. \tag{17}$$

For the new realization with no EF and error feedforward, we consider the problem of minimizing a measure

$$M(\boldsymbol{P}, \lambda) = \mathrm{tr}[\boldsymbol{V}\boldsymbol{P}] + \lambda(\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] - n) \tag{18}$$

with respect to $n \times n$ nonsingular matrix $\boldsymbol{P}$ and scalar λ where $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$, λ is a Lagrange multiplier, and

$$\boldsymbol{V} = \boldsymbol{A}_1^T\boldsymbol{W}_o\boldsymbol{A}_1 + \boldsymbol{A}_2^T\boldsymbol{W}^T\boldsymbol{A}_1 + \boldsymbol{A}_1^T\boldsymbol{W}\boldsymbol{A}_2 + \boldsymbol{A}_2^T\boldsymbol{W}_o\boldsymbol{A}_2.$$

To solve the minimization problem, we compute

$$\begin{aligned} \frac{\partial J(\boldsymbol{P}, \lambda)}{\partial \boldsymbol{P}} &= \boldsymbol{V} - \lambda\boldsymbol{P}^{-1}\boldsymbol{K}_c\boldsymbol{P}^{-1} \\ \frac{\partial J(\boldsymbol{P}, \lambda)}{\partial \lambda} &= \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] - n \end{aligned} \tag{19}$$

and solve the equations $\partial J(\boldsymbol{P}, \lambda)/\partial \boldsymbol{P} = \boldsymbol{0}$ and $\partial J(\boldsymbol{P}, \lambda)/\partial \lambda = 0$. This leads to

$$\boldsymbol{P}\boldsymbol{V}\boldsymbol{P} = \lambda\boldsymbol{K}_c, \qquad \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] = n. \tag{20}$$

It follows from (20) that

$$\begin{aligned} &\boldsymbol{P} = \sqrt{\lambda}\boldsymbol{V}^{-\frac{1}{2}}[\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{V}^{\frac{1}{2}}]^{\frac{1}{2}}\boldsymbol{V}^{-\frac{1}{2}} \\ &\frac{1}{\sqrt{\lambda}}\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{V}]^{\frac{1}{2}} = \frac{1}{\sqrt{\lambda}}\left(\sum_{i=1}^{n}\theta_i\right) = n \end{aligned} \tag{21}$$

where θ_i^2 for $i = 1, 2, \cdots, n$ are the eigenvalues of $\boldsymbol{K}_c\boldsymbol{V}$. Therefore, we obtain

$$\boldsymbol{P} = \frac{1}{n}\left(\sum_{i=1}^{n}\theta_i\right)\boldsymbol{V}^{-\frac{1}{2}}[\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{V}^{\frac{1}{2}}]^{\frac{1}{2}}\boldsymbol{V}^{-\frac{1}{2}}. \tag{22}$$

Substituting (22) into (18) yields the minimum value of $M(\boldsymbol{P}, \lambda)$ as

$$\min_{\boldsymbol{P}, \lambda} M(\boldsymbol{P}, \lambda) = \frac{1}{n}\left(\sum_{i=1}^{n}\theta_i\right)^2. \tag{23}$$

From (22), the optimal coordinate transformation matrix $\boldsymbol{T}$ that minimizes (18) can be obtained in closed form as

$$\boldsymbol{T} = \left(\sum_{i=1}^{n}\frac{\theta_i}{n}\right)^{\frac{1}{2}}\boldsymbol{V}^{-\frac{1}{2}}[\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{V}^{\frac{1}{2}}]^{\frac{1}{4}}\boldsymbol{U} \tag{24}$$

where $\boldsymbol{U}$ is an arbitrary $n \times n$ orthogonal matrix. From (24) it follows that

$$\overline{\boldsymbol{K}}_c = \boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T} = \left(\sum_{i=1}^{n}\frac{\theta_i}{n}\right)^{-1}\boldsymbol{U}^T[\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{V}^{\frac{1}{2}}]^{\frac{1}{2}}\boldsymbol{U}. \tag{25}$$

Next, we choose the $n \times n$ orthogonal matrix $\boldsymbol{U}$ such that the matrix $\overline{\boldsymbol{K}}_c$ in (25) satisfies the l_2-norm dynamic-range scaling constraints in (17). To this end, we perform the eigenvalue-eigenvector decomposition

$$[\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{V}^{\frac{1}{2}}]^{\frac{1}{2}} = \boldsymbol{R}\boldsymbol{\Theta}\boldsymbol{R}^T \tag{26}$$

where $\boldsymbol{\Theta} = \mathrm{diag}\{\theta_1, \theta_2, \cdots, \theta_n\}$, $\boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}_n$. Consequently

$$\left(\sum_{i=1}^{n}\frac{\theta_i}{n}\right)^{-1}[\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{V}^{\frac{1}{2}}]^{\frac{1}{2}} = \boldsymbol{R}\boldsymbol{\Theta}'\boldsymbol{R}^T \tag{27}$$

where

$$\boldsymbol{\Theta}' = \mathrm{diag}\{\theta_1', \theta_2', \cdots, \theta_n'\}, \qquad \theta_i' = \frac{n\theta_i}{\theta_1 + \theta_2 + \cdots + \theta_n}, \quad i = 1, 2, \cdots, n.$$

Now an $n \times n$ orthogonal matrix $\boldsymbol{S}$ such that

$$\boldsymbol{S}\boldsymbol{\Theta}'\boldsymbol{S}^T = \begin{bmatrix} 1 & * & \cdots & * \\ * & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ * & \cdots & * & 1 \end{bmatrix} \tag{28}$$

can be obtained by numerical manipulations. By choosing $\boldsymbol{U} = \boldsymbol{R}\boldsymbol{S}^T$ in (24), the optimal coordinate transformation matrix $\boldsymbol{T}$ satisfying (17) and (23) simultaneously can now be constructed as

$$\boldsymbol{T} = \left(\sum_{i=1}^{n} \frac{\theta_i}{n} \right)^{\frac{1}{2}} \boldsymbol{V}^{-\frac{1}{2}} [\boldsymbol{V}^{\frac{1}{2}} \boldsymbol{K}_c \boldsymbol{V}^{\frac{1}{2}}]^{\frac{1}{4}} \boldsymbol{R}\boldsymbol{S}^T. \tag{29}$$

If the coordinate transformation in (14) is applied to the LSS model in (1), then (9) is changed to

$$\overline{I}(\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{h}, \boldsymbol{T}) = J_2(\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{T}) + \mathrm{tr}[(\overline{\boldsymbol{c}} - \boldsymbol{h})^T (\overline{\boldsymbol{c}} - \boldsymbol{h})] \tag{30}$$

where

$$\begin{aligned} J_2(\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{T}) = \mathrm{tr}\Big[& (\overline{\boldsymbol{A}}_1 - \boldsymbol{D}_1)^T \overline{\boldsymbol{W}}_o (\overline{\boldsymbol{A}}_1 - \boldsymbol{D}_1) + (\overline{\boldsymbol{A}}_2 - \boldsymbol{D}_2)^T \overline{\boldsymbol{W}}^T (\overline{\boldsymbol{A}}_1 - \boldsymbol{D}_1) \\ & + (\overline{\boldsymbol{A}}_1 - \boldsymbol{D}_1)^T \overline{\boldsymbol{W}} (\overline{\boldsymbol{A}}_2 - \boldsymbol{D}_2) + (\overline{\boldsymbol{A}}_2 - \boldsymbol{D}_2)^T \overline{\boldsymbol{W}}_o (\overline{\boldsymbol{A}}_2 - \boldsymbol{D}_2) \Big]. \end{aligned}$$

<u>Case 1</u>: $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are general matrices

In this case, we select the matrices $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ as

$$\boldsymbol{D}_1 = \overline{\boldsymbol{A}}_1, \qquad \boldsymbol{D}_2 = \overline{\boldsymbol{A}}_2. \tag{31}$$

<u>Case 2</u>: $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are diagonal matrices

We define

$$\begin{aligned} \boldsymbol{D}_1 &= \mathrm{diag}\{\alpha_1, \alpha_2, \cdots, \alpha_n\} \\ \boldsymbol{D}_2 &= \mathrm{diag}\{\beta_1, \beta_2, \cdots, \beta_n\}. \end{aligned} \tag{32}$$

By setting $\partial J_2(\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{T})/\partial \alpha_i = 0$ and $\partial J_2(\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{T})/\partial \beta_i = 0$ and solving these equations for $\{\alpha_i\}$ and $\{\beta_i\}$, we obtain

$$\begin{aligned} \alpha_i &= \frac{\overline{\boldsymbol{W}}_o(i,i)\boldsymbol{M}_1(i,i) - \overline{\boldsymbol{W}}(i,i)\boldsymbol{M}_2(i,i)}{\overline{\boldsymbol{W}}_o(i,i)^2 - \overline{\boldsymbol{W}}(i,i)^2} \\ \beta_i &= \frac{\overline{\boldsymbol{W}}_o(i,i)\boldsymbol{M}_2(i,i) - \overline{\boldsymbol{W}}(i,i)\boldsymbol{M}_1(i,i)}{\overline{\boldsymbol{W}}_o(i,i)^2 - \overline{\boldsymbol{W}}(i,i)^2} \end{aligned} \tag{33}$$

where $\boldsymbol{X}(i,j)$ denotes the (i,j)th element of matrix $\boldsymbol{X}$ and

$$\begin{aligned} \boldsymbol{M}_1 &= \overline{\boldsymbol{W}}_o \overline{\boldsymbol{A}}_1 + \overline{\boldsymbol{W}}\ \overline{\boldsymbol{A}}_2 \\ \boldsymbol{M}_2 &= \overline{\boldsymbol{W}}_o \overline{\boldsymbol{A}}_2 + \overline{\boldsymbol{W}}^T \overline{\boldsymbol{A}}_1. \end{aligned}$$

<u>Case 3</u>: $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are scalar matrices

In this case, matrices $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ assume the form

$$\boldsymbol{D}_1 = \alpha \boldsymbol{I}_n, \qquad \boldsymbol{D}_2 = \beta \boldsymbol{I}_n \tag{34}$$

where α and β are scalars. By setting $\partial J_2(\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{T})/\partial \alpha = 0$ and $\partial J_2(\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{T})/\partial \beta = 0$ and solving these equations for α and β, we obtain

$$\begin{aligned} \alpha &= \frac{\mathrm{tr}[\overline{\boldsymbol{W}}_o]\mathrm{tr}[\boldsymbol{M}_1] - \mathrm{tr}[\overline{\boldsymbol{W}}]\mathrm{tr}[\boldsymbol{M}_2]}{(\mathrm{tr}[\overline{\boldsymbol{W}}_o])^2 - (\mathrm{tr}[\overline{\boldsymbol{W}}])^2} \\ \beta &= \frac{\mathrm{tr}[\overline{\boldsymbol{W}}_o]\mathrm{tr}[\boldsymbol{M}_2] - \mathrm{tr}[\overline{\boldsymbol{W}}]\mathrm{tr}[\boldsymbol{M}_1]}{(\mathrm{tr}[\overline{\boldsymbol{W}}_o])^2 - (\mathrm{tr}[\overline{\boldsymbol{W}}])^2}. \end{aligned} \tag{35}$$

4. Joint Optimization of Error Feedback and Realization

Define

$$\hat{\boldsymbol{T}} = \boldsymbol{T}^T \boldsymbol{K}_c^{-\frac{1}{2}}. \tag{36}$$

Then (17) can be written as

$$(\hat{\boldsymbol{T}}^{-T} \hat{\boldsymbol{T}}^{-1})_{ii} = 1 \quad \text{for} \quad i = 1, 2, \cdots, n. \tag{37}$$

The constraints in (37) simply state that each column in matrix $\hat{\boldsymbol{T}}^{-1}$ must be a unity vector. These constraints are satisfied if $\hat{\boldsymbol{T}}^{-1}$ assumes the form

$$\hat{\boldsymbol{T}}^{-1} = \left[\frac{\boldsymbol{t}_1}{||\boldsymbol{t}_1||}, \frac{\boldsymbol{t}_2}{||\boldsymbol{t}_2||}, \cdots, \frac{\boldsymbol{t}_n}{||\boldsymbol{t}_n||} \right] \tag{38}$$

where $\boldsymbol{t}_i$'s for $i = 1, 2, \cdots, n$ are $n \times 1$ real vectors. In such a case, (30) can be expressed as

$$\hat{I}(\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{h}, \hat{\boldsymbol{T}}) = J_3(\boldsymbol{D}_1, \boldsymbol{D}_2, \hat{\boldsymbol{T}}) + \mathrm{tr}[(\hat{\boldsymbol{c}} - \boldsymbol{h})^T (\hat{\boldsymbol{c}} - \boldsymbol{h})] \tag{39}$$

where

$$\begin{aligned} J_3(\boldsymbol{D}_1, \boldsymbol{D}_2, \hat{\boldsymbol{T}}) = \mathrm{tr}\Big[& (\hat{\boldsymbol{A}}_1 - \boldsymbol{D}_1)^T \hat{\boldsymbol{W}}_o (\hat{\boldsymbol{A}}_1 - \boldsymbol{D}_1) + (\hat{\boldsymbol{A}}_2 - \boldsymbol{D}_2)^T \hat{\boldsymbol{W}}^T (\hat{\boldsymbol{A}}_1 - \boldsymbol{D}_1) \\ & + (\hat{\boldsymbol{A}}_1 - \boldsymbol{D}_1)^T \hat{\boldsymbol{W}} (\hat{\boldsymbol{A}}_2 - \boldsymbol{D}_2) + (\hat{\boldsymbol{A}}_2 - \boldsymbol{D}_2)^T \hat{\boldsymbol{W}}_o (\hat{\boldsymbol{A}}_2 - \boldsymbol{D}_2) \Big] \end{aligned}$$

with

$$\begin{aligned} &\hat{\boldsymbol{A}}_1 = \hat{\boldsymbol{T}}^{-T} (\boldsymbol{K}_c^{-\frac{1}{2}} \boldsymbol{A}_1 \boldsymbol{K}_c^{\frac{1}{2}}) \hat{\boldsymbol{T}}^T, \qquad \hat{\boldsymbol{A}}_2 = \hat{\boldsymbol{T}}^{-T} (\boldsymbol{K}_c^{-\frac{1}{2}} \boldsymbol{A}_2 \boldsymbol{K}_c^{\frac{1}{2}}) \hat{\boldsymbol{T}}^T, \qquad \hat{\boldsymbol{c}} = (\boldsymbol{c} \boldsymbol{K}_c^{\frac{1}{2}}) \hat{\boldsymbol{T}}^T \\ &\hat{\boldsymbol{W}}_o = \hat{\boldsymbol{T}} (\boldsymbol{K}_c^{\frac{1}{2}} \boldsymbol{W}_o \boldsymbol{K}_c^{\frac{1}{2}}) \hat{\boldsymbol{T}}^T, \qquad \hat{\boldsymbol{W}} = \hat{\boldsymbol{T}} (\boldsymbol{K}_c^{\frac{1}{2}} \boldsymbol{W} \boldsymbol{K}_c^{\frac{1}{2}}) \hat{\boldsymbol{T}}^T. \end{aligned}$$

When selecting vector $\boldsymbol{h}$ as $\boldsymbol{h} = \hat{\boldsymbol{c}}$, the problem of obtaining matrices $\boldsymbol{D}_1$, $\boldsymbol{D}_2$ and $\boldsymbol{T}$ that minimize $J_2(\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{T})$ in (30) subject to the scaling constraints in (17) can be converted into an unconstrained optimization problem of obtaining matrices $\boldsymbol{D}_1$, $\boldsymbol{D}_2$ and $\hat{\boldsymbol{T}}$ that minimize $J_3(\boldsymbol{D}_1, \boldsymbol{D}_2, \hat{\boldsymbol{T}})$ in (39).

Let $\boldsymbol{x}$ be the column vector that collects the variables in matrices $\boldsymbol{D}_1$, $\boldsymbol{D}_2$ and $[\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n]$. Then, $J_3(\boldsymbol{D}_1, \boldsymbol{D}_2, \hat{\boldsymbol{T}})$ is a function of $\boldsymbol{x}$, thus can be denoted by $J_3(\boldsymbol{x})$. We propose a quasi-Newton algorithm which starts with a trivial initial point $\boldsymbol{x}_0$ obtained from an initial assignment $\boldsymbol{D}_1 = \boldsymbol{D}_2 = \hat{\boldsymbol{T}} = \boldsymbol{I}_n$. In the kth iteration, the algorithm updates the most recent point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as [10]

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k \tag{40}$$

where

$$\begin{aligned} &\boldsymbol{d}_k = -\boldsymbol{S}_k \nabla J_3(\boldsymbol{x}_k), \qquad \alpha_k = arg \ \min_{\alpha} \ J_3(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k) \\ &\boldsymbol{S}_{k+1} = \boldsymbol{S}_k + \left(1 + \frac{\boldsymbol{\gamma}_k^T \boldsymbol{S}_k \boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_k^T \boldsymbol{\delta}_k} \right) \frac{\boldsymbol{\delta}_k \boldsymbol{\delta}_k^T}{\boldsymbol{\gamma}_k^T \boldsymbol{\delta}_k} - \frac{\boldsymbol{\delta}_k \boldsymbol{\gamma}_k^T \boldsymbol{S}_k + \boldsymbol{S}_k \boldsymbol{\gamma}_k \boldsymbol{\delta}_k^T}{\boldsymbol{\gamma}_k^T \boldsymbol{\delta}_k} \\ &\boldsymbol{S}_0 = \boldsymbol{I}, \quad \boldsymbol{\delta}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k, \quad \boldsymbol{\gamma}_k = \nabla J_3(\boldsymbol{x}_{k+1}) - \nabla J_3(\boldsymbol{x}_k). \end{aligned}$$

Here, $\nabla J_3(\boldsymbol{x})$ is the gradients of $J_3(\boldsymbol{x})$ with respect to $\boldsymbol{x}$, and $\boldsymbol{S}_k$ is a positive-definite approximation of the inverse Hessian matrix of $J_3(\boldsymbol{x})$. This iteration process continues until

$$|J_3(\boldsymbol{x}_{k+1}) - J_3(\boldsymbol{x}_k)| < \varepsilon \tag{41}$$

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

If $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are general matrices, vector $\boldsymbol{x}$ consists of matrix $[\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n]$ only. After obtaing matrix $\hat{\boldsymbol{T}}$, we select matrices $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ as

$$\boldsymbol{D}_1 = \hat{\boldsymbol{A}}_1, \qquad \boldsymbol{D}_2 = \hat{\boldsymbol{A}}_2. \tag{42}$$

If $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are scalar matrices, the gradient of $J(\boldsymbol{x}) = \mathrm{tr}[\hat{\boldsymbol{T}}\boldsymbol{M}\hat{\boldsymbol{T}}^T]$ with respect to the (i, j)th element t_{ij} of $[\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n]$ is found to be

$$\begin{aligned} \frac{\partial J(\boldsymbol{x})}{\partial t_{ij}} &= \lim_{\Delta \to 0} \frac{\mathrm{tr}[\hat{\boldsymbol{T}}_{ij}\boldsymbol{M}\hat{\boldsymbol{T}}_{ij}^T] - \mathrm{tr}[\hat{\boldsymbol{T}}\boldsymbol{M}\hat{\boldsymbol{T}}^T]}{\Delta} \\ &= 2\,\mathrm{tr}[(\hat{\boldsymbol{T}}\boldsymbol{g}_{ij})(\boldsymbol{e}_j^T\hat{\boldsymbol{T}})\boldsymbol{M}\hat{\boldsymbol{T}}^T] \\ &= 2\boldsymbol{e}_j^T(\hat{\boldsymbol{T}}\boldsymbol{M}\hat{\boldsymbol{T}}^T\hat{\boldsymbol{T}})\boldsymbol{g}_{ij} \end{aligned} \tag{43}$$

where matrix $\boldsymbol{M}$ is derived from (39) and $\hat{\boldsymbol{T}}_{ij}$ is the matrix obtained from $\hat{\boldsymbol{T}}$ with a perturbed (i, j)th component, and is given by [12]

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

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

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

Also, the gradients of $J(\boldsymbol{x})$ with respect to α and β are found to be

$$\begin{aligned} \frac{\partial J(\boldsymbol{x})}{\partial \alpha} &= 2\,\mathrm{tr}[\hat{\boldsymbol{W}}_o(\alpha\boldsymbol{I} - \hat{\boldsymbol{A}}_1) + \hat{\boldsymbol{W}}(\beta\boldsymbol{I} - \hat{\boldsymbol{A}}_2)] \\ \frac{\partial J(\boldsymbol{x})}{\partial \beta} &= 2\,\mathrm{tr}[\hat{\boldsymbol{W}}_o(\beta\boldsymbol{I} - \hat{\boldsymbol{A}}_2) + \hat{\boldsymbol{W}}^T(\alpha\boldsymbol{I} - \hat{\boldsymbol{A}}_1)], \end{aligned} \tag{44}$$

respectively.

If $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are diagonal matrices, the gradient of $J(\boldsymbol{x})$ with respect to the (i, j)th element t_{ij} of $[\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n]$ is derived from

$$\frac{\partial \mathrm{tr}[\boldsymbol{D}_1\hat{\boldsymbol{T}}\boldsymbol{M}\hat{\boldsymbol{T}}^T\boldsymbol{D}_2]}{\partial t_{ij}} = 2\boldsymbol{e}_j^T(\hat{\boldsymbol{T}}\boldsymbol{M}\hat{\boldsymbol{T}}^T\boldsymbol{D}_2\boldsymbol{D}_1\hat{\boldsymbol{T}})\boldsymbol{g}_{ij}. \tag{45}$$

In addition, the gradients of $J(\boldsymbol{x})$ with respect to $\{\alpha_i\}$ and $\{\beta_i\}$ become

$$\begin{aligned} \frac{\partial J(\boldsymbol{x})}{\partial \alpha_i} &= 2\boldsymbol{e}_i^T[\hat{\boldsymbol{W}}_o(\alpha_i\boldsymbol{I} - \hat{\boldsymbol{A}}_1) + \hat{\boldsymbol{W}}(\beta_i\boldsymbol{I} - \hat{\boldsymbol{A}}_2)]\boldsymbol{e}_i \\ \frac{\partial J(\boldsymbol{x})}{\partial \beta_i} &= 2\boldsymbol{e}_i^T[\hat{\boldsymbol{W}}_o(\beta_i\boldsymbol{I} - \hat{\boldsymbol{A}}_2) + \hat{\boldsymbol{W}}^T(\alpha_i\boldsymbol{I} - \hat{\boldsymbol{A}}_1)]\boldsymbol{e}_i. \end{aligned} \tag{46}$$

5. A Numerical Example

In this section, the proposed algorithms are illustrated by a numerical example. Consider a stable, locally controllable, and locally observable 2-D state-space digital filter with order $n = 4$ specified by

$$\boldsymbol{A}_1 = \begin{bmatrix} 0 & 0 & 0 & -0.00411 \\ 1 & 0 & 0 & 0.08007 \\ 0 & 1 & 0 & -0.42458 \\ 0 & 0 & 1 & 1.04460 \end{bmatrix}, \quad \boldsymbol{b}_1 = \begin{bmatrix} -0.01452 \\ 0.01234 \\ 0.02054 \\ 0.04762 \end{bmatrix}$$

$$\boldsymbol{A}_2 = \begin{bmatrix} -0.22608 & -0.40594 & -0.30955 & -0.14469 \\ 1.61428 & 1.61040 & 1.02336 & 0.43872 \\ 0.10054 & -0.60615 & -0.45322 & -0.31019 \\ -0.00723 & 0.24580 & 0.38668 & 0.56289 \end{bmatrix}$$

$$\boldsymbol{b}_2 = \begin{bmatrix} 0.01189 & 0.02351 & -0.00637 & 0.02094 \end{bmatrix}^T$$

$$\boldsymbol{c} = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}, \qquad d = 0.00943.$$

After carrying out the l_2-scaling for the above LSS model with a diagonal coordinate transformation matrix

$$\boldsymbol{T} = \text{diag}\{0.093632,\ 0.215308,\ 0.480320,\ 1.026019\},$$

the noise gain of the scaled LSS model with error feedforward was found to be $I(\mathbf{0},\mathbf{0},\mathbf{c}) = 76.641884$.

Next, matrix

$$\boldsymbol{P} = \begin{bmatrix} 0.017094 & -0.051123 & 0.047136 & -0.037341 \\ -0.051123 & 0.208643 & -0.311749 & 0.241456 \\ 0.047136 & -0.311749 & 0.851664 & -0.883201 \\ -0.037341 & 0.241456 & -0.883201 & 1.184618 \end{bmatrix}$$

was derived from (22) and substituted into (18) to yield $M(\boldsymbol{P},\lambda) = 3.230958$. From (29), the optimal coordinate transformation matrix $\boldsymbol{T}$ was computed as

$$\boldsymbol{T} = \begin{bmatrix} -0.038042 & 0.074039 & -0.050246 & -0.087410 \\ -0.086036 & -0.212354 & 0.108868 & 0.379861 \\ 0.369703 & 0.263782 & 0.391289 & -0.701638 \\ -0.035881 & -0.107954 & -0.748701 & 0.781743 \end{bmatrix}.$$

With the above preparation, the algorithms proposed in Sections 3 and 4 were applied and the results obtained are summarized in Tables I and II.

Concerning the results in Table I, for the general error-feedback with infinite-precision coefficients, the optimal error-feedback coefficient matrices were obtained as

$$\boldsymbol{D}_1 = \begin{bmatrix} 0.344657 & -0.179239 & 0.357772 & 0.285109 \\ 0.068602 & 0.242850 & 0.238424 & 0.258793 \\ -0.354861 & 0.039246 & 0.301203 & -0.041770 \\ 0.110406 & 0.256071 & -0.162096 & 0.155890 \end{bmatrix}$$

$$\boldsymbol{D}_2 = \begin{bmatrix} 0.260894 & 0.047842 & -0.175978 & 0.172022 \\ -0.373913 & 0.402753 & -0.183875 & 0.030024 \\ 0.150711 & 0.155278 & 0.513628 & -0.032792 \\ 0.235014 & 0.191819 & 0.147592 & 0.316715 \end{bmatrix}.$$

For the diagonal error-feedback with infinite-precision coefficients, the optimal error-feedback coefficients were computed as

$$\boldsymbol{D}_1 = \text{diag}\{0.470260,\ 0.397949,\ 0.415318,\ 0.360817\}$$
$$\boldsymbol{D}_2 = \text{diag}\{0.517457,\ 0.600668,\ 0.491349,\ 0.482393\}.$$

Furthermore, for the scalar error-feedback with infinite-precision coefficients, the optimal error-feedback coefficients were derived as

$$\boldsymbol{D}_1 = 0.413630\,\boldsymbol{I}_4 \qquad \boldsymbol{D}_2 = 0.510074\,\boldsymbol{I}_4.$$

Concerning the results in Table II, for the diagonal error-feedback with infinite-precision coefficients, the jointly optimal coordinate transformation matrix and error-feedback coefficients were obtained as

$$\boldsymbol{T} = \begin{bmatrix} 0.021174 & -0.086477 & 0.004948 & -0.045964 \\ 0.034036 & 0.231248 & -0.124387 & 0.142746 \\ -0.028179 & -0.061520 & 0.458501 & -0.584703 \\ 0.020559 & -0.023235 & -0.473504 & 1.203926 \end{bmatrix}$$

$$\boldsymbol{D}_1 = \text{diag}\{0.452058,\ 0.372026,\ 0.166759,\ 0.513014\}$$
$$\boldsymbol{D}_2 = \text{diag}\{0.577567,\ 0.668349,\ 0.349703,\ 0.444123\},$$

respectively. Moreover, for the scalar error-feedback with infinite-precision coefficients, the jointly optimal coordinate transformation matrix and error-feedback scalars were derived as

$$\boldsymbol{T} = \begin{bmatrix} 0.044415 & -0.081243 & -0.000255 & -0.005814 \\ -0.001838 & 0.231026 & -0.056584 & 0.030832 \\ -0.013301 & -0.188906 & 0.510909 & -0.294235 \\ -0.196598 & 0.263116 & -0.785674 & 0.807781 \end{bmatrix}$$

$$\boldsymbol{D}_1 = 0.500637\,\boldsymbol{I}_4 \qquad \boldsymbol{D}_2 = 0.480951\,\boldsymbol{I}_4,$$

respectively.

TABLE I
ROUNDOFF NOISE GAIN IN *SEPARATE OPTIMIZATION*

Error Feedback Matrices	General	Diagonal	Scalar
Infinite Precision	0	0.454620	0.469953
3-Bit Quantization	0.049936	0.461004	0.481157
Integer Quantization	3.100355	2.217879	1.586120

TABLE II
ROUNDOFF NOISE GAIN IN *JOINT OPTIMIZATION*

Error Feedback Matrices	General	Diagonal	Scalar
Infinite Precision	0	0.186190	0.233562
3-Bit Quantization	0.049936	0.216724	0.243361
Integer Quantization	3.100355	1.736732	1.808645

6. Conclusion

The separate/joint optimization of error feedback and realization to minimize the roundoff noise subject to l_2-norm dynamic-range scaling constraints for a class of 2-D state-space digital filters has been investigated. It has been shown that the joint optimization problem can be converted into an unconstrained optimization problem by using linear algebraic techniques. An efficient quasi-Newton algorithm has then been employed to solve the unconstrained optimization problem iteratively. Our computer simulation results have demonstrated the effectiveness of the proposed techniques.

References

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
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. 33, pp. 965-973, Oct. 1986.
8. 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.
9. T. Hinamoto, H. Ohnishi and W.-S. Lu, "Roundoff noise minimization for 2-D state-space digital filters using joint optimization of error feedback and realization," *IEEE Trans. Signal Processing*, vol. 54, pp. 4302-4310, Nov. 2006.
10. R. Fletcher, *Practical Methods of Optimization*, 2nd ed. Wiley, New York, 1987.
11. E. Fornasini and G. Marchesini, "Doubly-indexed dynamical systems: State-space models and structural properties," *Math. Syst. Theory*, vol. 12, pp. 59-72, 1978.
12. T. Kailath, *Linear Systems*, Prentice Hall, 1980.