Jointly Optimized Error-Feedback and Realization for Roundoff Noise Minimization in a Class of 2-D State-Space Digital Filters

Takao Hinamoto and Toru Oumi
Graduate School of Engineering
Hiroshima University
Higashi-Hiroshima 739-8527, Japan
Email: {hinamoto, oumi}@hiroshima-u.ac.jp

Wu-Sheng Lu
Dept. of Elec. and Comp. Engineering
University of Victoria
Victoria, BC, Canada V8W 3P6
Email: wslu@ece.uvic.ca

***Abstract*— The separate/joint optimization techniques of error-feedback and realization are explored so as to reduce or minimize the effects of roundoff noise subject to l_2-scaling constraints for a class of 2-D state-space digital filters. First, the joint optimization problem at hand is converted into an unconstrained optimization problem by using linear-algebraic techniques. Next, the resultant unconstrained problem is solved iteratively by applying an efficient quasi-Newton algorithm. Finally, a numerical example is presented to illustrate the utility of the proposed techniques.**

I. INTRODUCTION

In the case when an IIR digital filter is implemented in fixed-point arithmetic, the problem of reducing the effects of roundoff noise at the filter output is of practically significance. Error feedback is known as a useful tool for reducing finite-word-length (FWL) effects in IIR digital filters. Many error feedback techniques have been proposed for 2-D IIR digital filters [1]-[5]. Another useful approach is to realize the 2-D state-space filter structure so that the roundoff noise gain is minimized by optimally choosing a linear transformation to the state-space coordinates subject to l_2-scaling constraints [6]-[8]. As a natural extension of the fore-mentioned methods, efforts have been made to develop new methods that combine the error feedback and the coordinate transformation for better performance in the roundoff noise reduction. In this connection, iterative algorithms for obtaining separately/jointly optimized error feedback and realization have been developed to reduce and/or minimize the effects of roundoff noise in 2-D state-space digital filters [9],[10].

In this paper, the problems of separately/jointly optimizing error feedback and realization are investigated to reduce or minimize the effects of roundoff noise subject to l_2-scaling constraints for a class of 2-D state-space digital filters. The 2-D state-space digital filters treated here are described by a local state-space model reported in [11], that corresponds to a transposed structure of the Fornasini-Marchesini second model [12]. We propose a separately-optimized analytical technique as well as a jointly-optimized iterative technique, which relies on an efficient quasi-Newton algorithm [13]. Our computer simulation results demonstrate the validity and effectiveness of the proposed techniques.

II. ROUNDOFF NOISE ANALYSIS AND SCALING

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

$$\begin{bmatrix} \boldsymbol{x}(i+1,\, j+1) \\ y(i,\, j) \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 \\ \boldsymbol{c}_1 & \boldsymbol{c}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(i,\, j+1) \\ \boldsymbol{x}(i+1,\, j) \end{bmatrix} + \begin{bmatrix} \boldsymbol{b} \\ d \end{bmatrix} u(i,\, j) \tag{1}$$

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 $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{b}$, $\boldsymbol{c}_1$, $\boldsymbol{c}_2$, and d are real matrices of appropriate dimensions. The LSS model in (1) is assumed to be stable, locally controllable and locally observable. A block diagram of the LSS model in (1) is shown in Fig. 1.

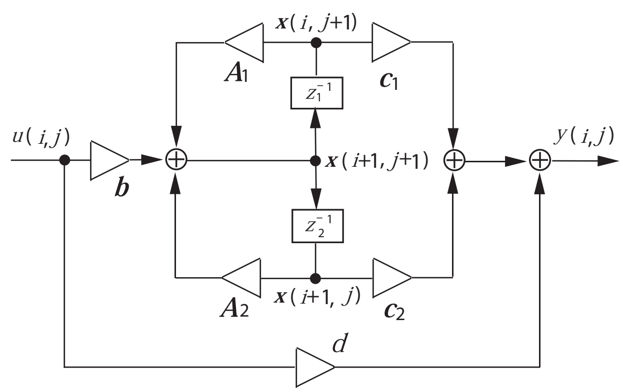

Fig. 1. A LSS model for 2-D filters.

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

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

where $\boldsymbol{e}(i,\, j)$ is an $n \times 1$ quantization-error vector defined by

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

$\boldsymbol{D}$ is $n \times n$ error feedback matrix, $\boldsymbol{h}$ is a $1 \times n$ error-feedforward vector, and each component of matrices $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{b}$, $\boldsymbol{c}_1$, $\boldsymbol{c}_2$ 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 vector $\boldsymbol{e}(i,j)$ is modeled as a zero-mean noise process of covariance $\sigma^2 \boldsymbol{I}_n$ with

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

Subtracting (2) from (1) yields

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

where

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

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

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

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

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

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

Let the transition matrix $\boldsymbol{A}^{(i,j)}$ be defined by

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

where $i, j \geq 0$. Then the following properties holds:

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

Substituting (6) into (4) yields

$$\begin{aligned} \boldsymbol{G}(z_1, z_2) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{g}(i,\, j)\,(z_1 z_2 \boldsymbol{I}_n - \boldsymbol{D})\, z_1^{-i} z_2^{-j} - \boldsymbol{h} \\ \boldsymbol{g}(i,\, j) &= \boldsymbol{c}_1 \boldsymbol{A}^{(i-1,j)} + \boldsymbol{c}_2 \boldsymbol{A}^{(i,j-1)} \end{aligned} \tag{8}$$

Substituting (8) into (5), it follows that

$$I(\boldsymbol{D}, \boldsymbol{h}) = J_1(\boldsymbol{D}) + \text{tr}\Big[\big\{\boldsymbol{g}(1,1) - \boldsymbol{h}\big\}^T \big\{\boldsymbol{g}(1,1) - \boldsymbol{h}\big\}\Big] \tag{9}$$

where

$$J_1(\boldsymbol{D}) = \text{tr}\Big[\boldsymbol{W}_o - 2\,\boldsymbol{W}\boldsymbol{D} + \boldsymbol{W}_o \boldsymbol{D}\boldsymbol{D}^T - \boldsymbol{g}^T(1,1)\,\boldsymbol{g}(1,1)\Big].$$

Here, the $n \times n$ matrices $\boldsymbol{W}_o$ and $\boldsymbol{W}$ are defined by

$$\begin{aligned} \boldsymbol{W}_o &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{g}^T(i,\, j)\,\boldsymbol{g}(i,\, j) \\ \boldsymbol{W} &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{g}^T(i+1,\, j+1)\,\boldsymbol{g}(i,\, j) \end{aligned} \tag{10}$$

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

$$I(\boldsymbol{0}, \boldsymbol{g}(1,1)) = \text{tr}\Big[\boldsymbol{W}_o - \boldsymbol{g}^T(1,1)\,\boldsymbol{g}(1,1)\Big]. \tag{11}$$

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

$$\boldsymbol{f}(i,\, j) = \boldsymbol{A}^{(i,j)}\boldsymbol{b}.$$

The LSS model in (1) is said to satisfy l_2-scaling constraints provided that

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

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

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

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

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

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

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

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

For the new realization with no error feedback and error feedforward, we consider the problem of minimizing

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

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

$$\boldsymbol{V} = \boldsymbol{W}_o - \boldsymbol{g}^T(1,1)\,\boldsymbol{g}(1,1).$$

We compute

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

By setting $\partial J(\boldsymbol{P},\lambda)/\partial \boldsymbol{P} = \boldsymbol{0}$ and $\partial J(\boldsymbol{P},\lambda)/\partial\lambda = 0$, we obtain

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

which lead to

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

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 $J(\boldsymbol{P},\lambda)$ as

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

It is noted that when matrix $\boldsymbol{T}$ assumes the form

$$\boldsymbol{T} = \boldsymbol{P}^{\frac{1}{2}}\boldsymbol{U} \tag{24}$$

where $\boldsymbol{P}^{1/2}$ is the square root of $\boldsymbol{P}$ obtained by (22), there exists an $n\times n$ orthogonal matrix $\boldsymbol{U}$ such that (17) is satisfied.

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

$$\overline{I}(\boldsymbol{D},\boldsymbol{h},\boldsymbol{T}) = J_2(\boldsymbol{D},\boldsymbol{T}) + ||\,\overline{\boldsymbol{g}}^T(1,1) - \boldsymbol{h}^T||^2 \tag{25}$$

where

$$\begin{aligned}J_2(\boldsymbol{D},\boldsymbol{T}) = \mathrm{tr}\Big[&\overline{\boldsymbol{W}}_o - 2\,\overline{\boldsymbol{W}}\boldsymbol{D} + \overline{\boldsymbol{W}}_o\boldsymbol{D}\boldsymbol{D}^T \\ &- \overline{\boldsymbol{g}}^T(1,1)\,\overline{\boldsymbol{g}}(1,1)\Big].\end{aligned}$$

The optimal selection of matrix $\boldsymbol{D}$ can be made as follows.

<u>Case 1</u>: *$\boldsymbol{D}$ is a general matrix*

In this case, we select the matrix $\boldsymbol{D}$ as

$$\boldsymbol{D} = \overline{\boldsymbol{W}}_o^{-1}\overline{\boldsymbol{W}}^T. \tag{26}$$

<u>Case 2</u>: *$\boldsymbol{D}$ is a diagonal matrix*

We define

$$\boldsymbol{D} = \mathrm{diag}\{\alpha_1, \alpha_2, \cdots, \alpha_n\}. \tag{27}$$

From $\partial J_2(\boldsymbol{D},\boldsymbol{T})/\partial\alpha_i = 0$, it is derived that for $i = 1, 2, \cdots, n$

$$\alpha_i = \frac{\overline{\boldsymbol{W}}(i,i)}{\overline{\boldsymbol{W}}_o(i,i)} \tag{28}$$

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

<u>Case 3</u>: *$\boldsymbol{D}$ is a scalar matrix*

With a scalar α, we define

$$\boldsymbol{D} = \alpha\boldsymbol{I}_n. \tag{29}$$

From $\partial J_2(\boldsymbol{D},\boldsymbol{T})/\partial\alpha = 0$, we obtain

$$\alpha = \frac{\mathrm{tr}[\overline{\boldsymbol{W}}]}{\mathrm{tr}[\overline{\boldsymbol{W}}_o]}. \tag{30}$$

IV. JOINT OPTIMIZATION OF ERROR FEEDBACK AND REALIZATION

Define

$$\hat{\boldsymbol{T}} = \boldsymbol{T}^T\boldsymbol{K}_c^{-\frac{1}{2}}, \tag{31}$$

then (17) can be written as

$$(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{T}}^{-1})_{ii} = 1 \;\text{ for }\; i = 1, 2, \cdots, n. \tag{32}$$

The constraints in (32) simply state that each column in matrix $\hat{\boldsymbol{T}}^{-1}$ must be a unity vector. These constraints are always satisfied if $\hat{\boldsymbol{T}}^{-1}$ assumes the form

$$\hat{\boldsymbol{T}}^{-1} = \left[\frac{\boldsymbol{t}_1}{||\boldsymbol{t}_1||}, \frac{\boldsymbol{t}_2}{||\boldsymbol{t}_2||}, \cdots, \frac{\boldsymbol{t}_n}{||\boldsymbol{t}_n||}\right] \tag{33}$$

where $\boldsymbol{t}_i$'s for $i = 1, 2, \cdots, n$ are $n\times 1$ real vectors. In such a case, (25) can be expressed as

$$\begin{aligned}\hat{I}(\boldsymbol{D},\boldsymbol{h},\hat{\boldsymbol{T}}) &= J_3(\boldsymbol{D},\hat{\boldsymbol{T}}) \\ &\quad + \mathrm{tr}\Big[\big\{\hat{\boldsymbol{g}}(1,1) - \boldsymbol{h}\big\}^T\big\{\hat{\boldsymbol{g}}(1,1) - \boldsymbol{h}\big\}\Big]\end{aligned} \tag{34}$$

where

$$\begin{aligned}J_3(\boldsymbol{D},\hat{\boldsymbol{T}}) = \mathrm{tr}\Big[&\hat{\boldsymbol{W}}_o - 2\,\hat{\boldsymbol{W}}\boldsymbol{D} + \hat{\boldsymbol{W}}_o\boldsymbol{D}\boldsymbol{D}^T \\ &- \hat{\boldsymbol{g}}^T(1,1)\,\hat{\boldsymbol{g}}(1,1)\Big]\end{aligned}$$

with

$$\begin{aligned}\hat{\boldsymbol{W}}_o &= \hat{\boldsymbol{T}}(\boldsymbol{K}_c^{\frac{1}{2}}\boldsymbol{W}_o\boldsymbol{K}_c^{\frac{1}{2}})\hat{\boldsymbol{T}}^T, \qquad \hat{\boldsymbol{g}}(i,j) = \boldsymbol{g}(i,j)\boldsymbol{K}_c^{\frac{1}{2}}\hat{\boldsymbol{T}}^T \\ \hat{\boldsymbol{W}} &= \hat{\boldsymbol{T}}(\boldsymbol{K}_c^{\frac{1}{2}}\boldsymbol{W}\boldsymbol{K}_c^{\frac{1}{2}})\hat{\boldsymbol{T}}^T.\end{aligned}$$

Selecting a vector $\boldsymbol{h}$ as $\boldsymbol{h} = \hat{\boldsymbol{g}}(1,1)$, the problem of obtaining matrices $\boldsymbol{D}$ and $\boldsymbol{T}$ that minimize $J_2(\boldsymbol{D},\boldsymbol{T})$ in (25) subject to the l_2-scaling constraints in (17) can be converted into an unconstrained optimization problem of obtaining matrices $\boldsymbol{D}$ and $\hat{\boldsymbol{T}}$ that minimize $J_3(\boldsymbol{D},\hat{\boldsymbol{T}})$ in (34).

Let $\boldsymbol{x}$ be the column vector that collects the variables in matrices $\boldsymbol{D}$ and $\hat{\boldsymbol{T}}$. Then, $J_3(\boldsymbol{D},\hat{\boldsymbol{T}})$ is a function of $\boldsymbol{x}$, denoted by $J_3(\boldsymbol{x})$. The algorithm starts with a trivial initial point $\boldsymbol{x}_0$ obtained from an initial assignment $\boldsymbol{D} = \hat{\boldsymbol{T}} = \boldsymbol{I}_n$. In the kth iteration, a quasi-Newton algorithm updates the most recent point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as [13]

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{d}_k \tag{35}$$

where

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

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

$$|J_3(\boldsymbol{x}_{k+1}) - J_3(\boldsymbol{x}_k)| < \varepsilon \tag{36}$$

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

The gradient of $J(\boldsymbol{x})$ with respect to the (i,j)th element of $\hat{\boldsymbol{T}}$ is found to be

$$\frac{\partial J(\boldsymbol{x})}{\partial t_{ij}} = \lim_{\Delta \to 0} \frac{J(\hat{\boldsymbol{T}}_{ij}) - J(\hat{\boldsymbol{T}})}{\Delta} \tag{37}$$

where $\hat{\boldsymbol{T}}_{ij}$ is the matrix obtained from $\hat{\boldsymbol{T}}$ with a perturbed (i,j)th component, and is given by [14]

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

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

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

V. A NUMERICAL EXAMPLE

Consider a stable, locally controllable, and locally observable 2-D state-space digital filter with order $n = 4$ specified by

$$\boldsymbol{A}_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -0.00411 & 0.08007 & -0.42458 & 1.04460 \end{bmatrix}$$

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

$$\boldsymbol{c}_1 = \begin{bmatrix} -0.01452 & 0.01234 & 0.02054 & 0.04762 \end{bmatrix}$$

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

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

After carrying out the l_2-scaling for the above LSS model with a diagonal coordinate transformation matrix, the noise gain of the scaled LSS model with error feedforward was computed as $I(\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{g}(1,1)) = 76.293336$. Next, the matrix $\boldsymbol{P}$ was derived from (22) and substituting it into (18) produced $J(\boldsymbol{P}, \lambda) = 3.950212$. The remaining results obtained by applying the proposed techniques are summarized in Tables I and II.

TABLE I
ROUNDOFF NOISE GAIN IN *SEPARATE OPTIMIZATION*

Error Feedback Matrix	General	Diagonal	Scalar
Infinite Precision	1.109415	1.482713	1.553030
3-Bit Quantization	1.138300	1.498537	1.561847
Integer Quantization	2.022930	1.919520	1.919520

TABLE II
ROUNDOFF NOISE GAIN IN *JOINT OPTIMIZATION*

Error Feedback Matrix	General	Diagonal	Scalar
Infinite Precision	1.103832	1.194349	1.281825
3-Bit Quantization	1.159342	1.217794	1.290750
Integer Quantization	4.357707	1.535906	1.529194

VI. CONCLUSION

In this paper, we have investigated the separate/joint optimization problems of error feedback and realization to reduce or minimize the effects of roundoff noise subject to l_2-scaling constraints for a class of 2-D state-space digital filters. It has been shown that the problem in the joint optimization can be converted into an unconstrained optimization problem by using linear algebraic techniques. The resultant unconstrained optimization problem has been solved by emplying an efficient quasi-Newton algorithm. Our computer simulation results have demonstrated the validity and 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, 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] 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.

[10] 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.

[11] T. Hinamoto, "A novel local state-space model for 2-D digital filers and its properties," *Proc. 2001 IEEE Int. Symp. Circuits Syst.*, vol. 2, pp. 545-548.

[12] E. Fornasini and G. Marchesini, "Doubly-indexed dynamical systems: State-space models and structural properties," *Math Syst. Theory*, vol.12, pp.59-72, 1978.

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

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