JOINTLY OPTIMAL HIGH-ORDER ERROR FEEDBACK AND REALIZATION FOR ROUNDOFF NOISE MINIMIZATION IN 2-D STATE-SPACE DIGITAL FILTERS

Takao Hinamoto, Akimitsu Doi
Hiroshima Institute of Technology
Hiroshima 731-5193, Japan
Emails: hinamoto@ieee.org, doi@cc.it-hiroshima.ac.jp

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

***Abstract*—The joint optimization problem of high-order error feedback and realization for minimizing roundoff noise at filter output subject to l_2-scaling constraints is investigated for two-dimensional (2-D) state-space digital filters. Linear algebraic techniques that convert the problem at hand into an unconstrained optimization problem are explored, and an efficient quasi-Newton algorithm is then applied to solve the unconstrained optimization problem iteratively. In this connection, closed-form formulas are derived for fast and accurate gradient evaluation. Finally a numerical example is presented to illustrate the validity and effectiveness of the proposed algorithm.**

1. INTRODUCTION

In the implementation of IIR digital filters with fixed-point arithmetic, it is of critical importance to reduce the effects of roundoff noise at the filter output. Error feedback (EF) is found effective for the reduction of finite-word-length (FWL) effects in IIR digital filters, and many EF methods have been proposed in the past [1]-[10]. Alternatively, the roundoff noise can also be reduced by introducing a delta operator to IIR digital filters [11]-[13], or by adopting a new structure based on the concept of polynomial operators for digital filter implementation [14]. Another useful approach is to synthesize 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-scaling constraints [15]-[18]. As a natural extension of the aforementioned methods, efforts have been made to develop new methods that combine EF and state-space realization, for achieving better performance [19]-[21]. Separately and jointly optimized scalar or general EF matrix for state-space filters have been explored [19]. A jointly-optimized iterative algorithm with a general, diagonal, or scalar EF matrix using a quasi-Newton method has been developed for state-space digital filters [20]. The similar technique is extended to 2-D digital filters with a general, block-diagonal, diagonal, or scalar EF matrix [21]. However, to date the use of high-order EF and its effect on noise-reduction performance for 2-D state-space digital filters have not been explored except for the case of the Fornasini-Marchesini second model [22].

This paper investigates the problem of jointly optimizing *high-order* EF and realization for minimizing the roundoff noise subject to l_2-scaling constraints for 2-D state-space digital filters described by the Roesser model. We convert the constrained optimization problem encountered into an unconstrained optimization problem by using linear-algebraic techniques. An efficient quasi-Newton algorithm is utilized to solve the unconstrained optimization problem, and is applied to the case where the *high-order* EF has diagonal matrices.

2. PROBLEM STATEMENT

Let us consider a 2-D stable, separately locally controllable and separately locally observable state-space digital filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{m+n}$ described by the Roessor model [23]

$$\begin{aligned} \boldsymbol{x}_1(i,j) &= \boldsymbol{A}\boldsymbol{x}_0(i,j) + \boldsymbol{b}u(i,j) \\ y(i,j) &= \boldsymbol{c}\,\boldsymbol{x}_0(i,j) + du(i,j) \end{aligned} \tag{1}$$

with

$$\boldsymbol{x}_k(i,j) = \begin{bmatrix} \boldsymbol{x}^h(i+k,j) \\ \boldsymbol{x}^v(i,j+k) \end{bmatrix}, \quad \boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 \\ \boldsymbol{A}_3 & \boldsymbol{A}_4 \end{bmatrix}$$

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

where $\boldsymbol{x}^h(i,j)$ is an $m \times 1$ horizontal state vector, $\boldsymbol{x}^v(i,j)$ is an $n \times 1$ vertical state vector, $u(i,j)$ is a scalar input, $y(i,j)$ is a scalar output, and $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{A}_3$, $\boldsymbol{A}_4$, $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, $\boldsymbol{c}_1$, $\boldsymbol{c}_2$ and d are real constant matrices of appropriate dimensions.

Assuming that the quantization is performed before matrix-vector multiplication, an actual FWL implementation of model (1) with error feedforward and *high-order* EF is obtained as

$$\begin{aligned} \tilde{\boldsymbol{x}}_1(i,j) &= \boldsymbol{A}\boldsymbol{Q}[\tilde{\boldsymbol{x}}_0(i,j)] + \boldsymbol{b}u(i,j) + \sum_{k=0}^{N-1} \boldsymbol{D}_k \boldsymbol{e}_{-k}(i,j) \\ \tilde{y}(i,j) &= \boldsymbol{c}\,\boldsymbol{Q}[\tilde{\boldsymbol{x}}_0(i,j)] + du(i,j) + \boldsymbol{h}\boldsymbol{e}_0(i,j) \end{aligned} \tag{2}$$

where $\boldsymbol{h}$ is a $1 \times (m+n)$ error-feedforward vector, $\boldsymbol{D}_k$'s are $(m+n) \times (m+n)$ diagonal *high-order* EF matrices and

$$\boldsymbol{e}_{-k}(i,j) = \begin{bmatrix} \boldsymbol{e}^h(i-k,j) \\ \boldsymbol{e}^v(i,j-k) \end{bmatrix}, \quad \boldsymbol{D}_k = \begin{bmatrix} \boldsymbol{D}_{1k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_{4k} \end{bmatrix}$$

$$\boldsymbol{e}_k(i,j) = \tilde{\boldsymbol{x}}_k(i,j) - \boldsymbol{Q}[\tilde{\boldsymbol{x}}_k(i,j)], \quad \boldsymbol{h} = \begin{bmatrix} \boldsymbol{h}_1 & \boldsymbol{h}_2 \end{bmatrix}.$$

The coefficient matrices $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{c}$ and d in (2) are assumed to have exact fractional B_c-bit representations. The FWL horizontal (vertical) state vector $\tilde{\boldsymbol{x}}^h(i,j)$ ($\tilde{\boldsymbol{x}}^v(i,j)$) and each output $\tilde{y}(i,j)$ has B-bit fractional representations, 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 $\tilde{\boldsymbol{x}}_k(i,j)$ to $(B-B_c)$-bit after the multiplications and additions, where the sign bit is not counted. It is assumed that the roundoff error $\boldsymbol{e}_0(i,j)$ can be modeled as a Gaussian random process with zero mean and covariance $\sigma^2\boldsymbol{I}_n$. By subtracting (2) from (1), it follows that

$$\begin{aligned} \Delta\boldsymbol{x}_1(i,j) &= \boldsymbol{A}\Delta\boldsymbol{x}_0(i,j) + \boldsymbol{A}\boldsymbol{e}_0(i,j) - \sum_{k=0}^{N-1} \boldsymbol{D}_k \boldsymbol{e}_{-k}(i,j) \\ \Delta y(i,j) &= \boldsymbol{c}\Delta\boldsymbol{x}_0(i,j) + (\boldsymbol{c}-\boldsymbol{h})\boldsymbol{e}_0(i,j) \end{aligned} \tag{3}$$

where $\Delta \boldsymbol{x}_k(i,j) = \boldsymbol{x}_k(i,j) - \tilde{\boldsymbol{x}}_k(i,j)$ for $k = 0, 1$ and $\Delta y(i,j) = y(i,j) - \tilde{y}(i,j)$. By taking (z_1, z_2)-transform on both sides of (3) and setting the boundary conditions to be null, we obtain

$$\begin{aligned}
&\Delta Y(z_1,z_2) = \boldsymbol{H}_e(z_1,z_2)\boldsymbol{E}_0(z_1,z_2)\\
&\boldsymbol{H}_e(z_1,z_2) = \boldsymbol{c}(\boldsymbol{Z}-\boldsymbol{A})^{-1}\Big(\boldsymbol{Z} - \sum_{k=0}^{N-1}\boldsymbol{D}_k\boldsymbol{Z}^{-k}\Big) - \boldsymbol{h}
\end{aligned} \tag{4}$$

where $\Delta Y(z_1,z_2)$ and $\boldsymbol{E}_0(z_1,z_2)$ are (z_1,z_2)-transforms of $\Delta y(i,j)$ and $\boldsymbol{e}_0(i,j)$, respectively, and $\boldsymbol{Z} = z_1\boldsymbol{I}_m \oplus z_2\boldsymbol{I}_n$. Define the transition matrix $\boldsymbol{A}^{(i,j)}$ by

$$(\boldsymbol{I}_{m+n} - \boldsymbol{Z}^{-1}\boldsymbol{A})^{-1} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{A}^{(i,j)}z_1^{-i}z_2^{-j}. \tag{5}$$

Since $\boldsymbol{c}(\boldsymbol{Z}-\boldsymbol{A})^{-1} = \boldsymbol{c}(\boldsymbol{I}_{m+n} - \boldsymbol{Z}^{-1}\boldsymbol{A})^{-1}\boldsymbol{Z}^{-1}$, the noise transfer function $\boldsymbol{H}_e(z_1,z_2)$ can be written as

$$\boldsymbol{H}_e(z_1,z_2) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{g}(i,j)z_1^{-i}z_2^{-j}\Big(\boldsymbol{Z} - \sum_{k=0}^{N-1}\boldsymbol{D}_k\boldsymbol{Z}^{-k}\Big) - \boldsymbol{h} \tag{6}$$

where

$$\begin{aligned}
\boldsymbol{A}^{(0,0)} &= \boldsymbol{I}_{m+n}, \quad \boldsymbol{A}^{(i,j)} = \boldsymbol{0}, \quad i<0 \text{ or } j<0\\
\boldsymbol{A}^{(1,0)} &= \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}, \quad \boldsymbol{A}^{(0,1)} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{A}_3 & \boldsymbol{A}_4 \end{bmatrix}\\
\boldsymbol{A}^{(i,j)} &= \boldsymbol{A}^{(1,0)}\boldsymbol{A}^{(i-1,j)} + \boldsymbol{A}^{(0,1)}\boldsymbol{A}^{(i,j-1)}\\
&= \boldsymbol{A}^{(i-1,j)}\boldsymbol{A}^{(1,0)} + \boldsymbol{A}^{(i,j-1)}\boldsymbol{A}^{(0,1)}, \quad (i,j) > (0,0)\\
\boldsymbol{g}(i,j) &= \boldsymbol{c}\Big(\boldsymbol{A}^{(i-1,j)}\begin{bmatrix} \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} + \boldsymbol{A}^{(i,j-1)}\begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_n \end{bmatrix}\Big).
\end{aligned}$$

We now define the normalized noise gain $J_{e1}(\boldsymbol{h},\boldsymbol{D}) = \sigma_{out}^2/\sigma^2$ with $\boldsymbol{D} = [\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}]$ as

$$\begin{aligned}
&J_{e1}(\boldsymbol{h},\boldsymbol{D})\\
&= \text{tr}\Big[\frac{1}{(2\pi j)^2}\oint_{|z_1|=1}\oint_{|z_2|=1}\boldsymbol{H}_e^*(z_1,z_2)\boldsymbol{H}_e(z_1,z_2)\frac{dz_1 dz_2}{z_1 z_2}\Big]
\end{aligned} \tag{7}$$

where σ_{out}^2 is noise variance at the filter output. Substituting (6) into (7), it follows that

$$\begin{aligned}
J_{e1}(\boldsymbol{h},\boldsymbol{D}) = \text{tr}\Big[&\boldsymbol{W}_{10} - \boldsymbol{c}_1^T\boldsymbol{c}_1 - 2\sum_{k=0}^{N-1}\boldsymbol{W}_{1,k+1}\boldsymbol{D}_{1k}\\
&+ \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\boldsymbol{W}_{1,|k-l|}\boldsymbol{D}_{1k}\boldsymbol{D}_{1l}\Big]\\
&+\text{tr}\Big[\boldsymbol{W}_{40} - \boldsymbol{c}_2^T\boldsymbol{c}_2 - 2\sum_{k=0}^{N-1}\boldsymbol{W}_{4,k+1}\boldsymbol{D}_{4k}\\
&+ \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\boldsymbol{W}_{4,|k-l|}\boldsymbol{D}_{4k}\boldsymbol{D}_{4l}\Big] + (\boldsymbol{c}-\boldsymbol{h})(\boldsymbol{c}-\boldsymbol{h})^T
\end{aligned} \tag{8}$$

where

$$\begin{aligned}
\boldsymbol{W}_{1r} &= [\ \boldsymbol{I}_m \quad \boldsymbol{0}\ ]\Big[\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{g}^T(i+r,j)\boldsymbol{g}(i,j)\Big]\begin{bmatrix} \boldsymbol{I}_m \\ \boldsymbol{0} \end{bmatrix}\\
\boldsymbol{W}_{4r} &= [\ \boldsymbol{0} \quad \boldsymbol{I}_n\ ]\Big[\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{g}^T(i,j+r)\boldsymbol{g}(i,j)\Big]\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I}_n \end{bmatrix}
\end{aligned}$$

are an extension of the local observability Gramian [23].

It should be noted that l_2-scaling constraints on the local state vector $\boldsymbol{x}_0(i,j)$ involve the local controllability Gramian $\boldsymbol{K}_c$ of the filter in (1), which can be computed by

$$\begin{aligned}
\boldsymbol{K}_c &= \frac{1}{(2\pi j)^2}\oint_{|z_1|=1}\oint_{|z_2|=1}\boldsymbol{F}(z_1,z_2)\boldsymbol{F}^*(z_1,z_2)\frac{dz_1 dz_2}{z_1 z_2}\\
&= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{f}(i,j)\boldsymbol{f}^T(i,j) = \begin{bmatrix} \boldsymbol{K}_{c1} & \boldsymbol{K}_{c2} \\ \boldsymbol{K}_{c3} & \boldsymbol{K}_{c4} \end{bmatrix}
\end{aligned} \tag{9}$$

where

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

A different yet equivalent local state-space description of (1), $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_n$, can be obtained via a coordinate transformation $\overline{\boldsymbol{x}}_0(i,j) = \boldsymbol{T}^{-1}\boldsymbol{x}_0(i,j)$ with $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ where

$$\overline{\boldsymbol{A}} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}, \quad \overline{\boldsymbol{b}} = \boldsymbol{T}^{-1}\boldsymbol{b}, \quad \overline{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T}. \tag{10}$$

The l_2-scaling constraints are imposed on the local state vector $\overline{\boldsymbol{x}}_0(i,j)$ so that

$$(\overline{\boldsymbol{K}}_c)_{ii} = (\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T})_{ii} = 1, \quad i = 1, 2, \cdots, m+n. \tag{11}$$

The problem being considered here is to design an optimal coordinate transformation matrix $\boldsymbol{T}$ as well as diagonal *high-order* EF matrices $\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}$ that jointly minimize the noise gain

$$\begin{aligned}
J_{e2}(\boldsymbol{D},\boldsymbol{T}) = \text{tr}\Big[&\overline{\boldsymbol{W}}_{10} - \overline{\boldsymbol{c}}_1^T\overline{\boldsymbol{c}}_1 - 2\sum_{k=0}^{N-1}\overline{\boldsymbol{W}}_{1,k+1}\boldsymbol{D}_{1k}\\
&+ \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\overline{\boldsymbol{W}}_{1,|k-l|}\boldsymbol{D}_{1k}\boldsymbol{D}_{1l}\Big]\\
&+\text{tr}\Big[\overline{\boldsymbol{W}}_{40} - \overline{\boldsymbol{c}}_2^T\overline{\boldsymbol{c}}_2 - 2\sum_{k=0}^{N-1}\overline{\boldsymbol{W}}_{4,k+1}\boldsymbol{D}_{4k}\\
&+ \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\overline{\boldsymbol{W}}_{4,|k-l|}\boldsymbol{D}_{4k}\boldsymbol{D}_{4l}\Big]
\end{aligned} \tag{12}$$

subject to l_2-scaling constraints in (11) where

$$\overline{\boldsymbol{W}}_{1r} = \boldsymbol{T}_1^T\boldsymbol{W}_{1r}\boldsymbol{T}_1, \quad \overline{\boldsymbol{W}}_{4r} = \boldsymbol{T}_4^T\boldsymbol{W}_{4r}\boldsymbol{T}_4$$

and the error feedforward vector $\boldsymbol{h}$ is chosen as $\boldsymbol{h} = \overline{\boldsymbol{c}}$.

3. JOINT OPTIMIZATION OF HIGH-ORDER ERROR FEEDBACK AND REALIZATION

To deal with the l_2-scaling constraints in (11), we define

$$\hat{\boldsymbol{T}} = \hat{\boldsymbol{T}}_1 \oplus \hat{\boldsymbol{T}}_4 = (\boldsymbol{T}_1 \oplus \boldsymbol{T}_4)^T(\boldsymbol{K}_{c1} \oplus \boldsymbol{K}_{c4})^{-\frac{1}{2}} \tag{13}$$

which leads (11) to

$$(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{T}}^{-1})_{ii} = 1, \quad i = 1, 2, \cdots, m+n. \tag{14}$$

It is noted that these constraints are always satisfied if $\hat{\boldsymbol{T}}_1^{-1}$ and $\hat{\boldsymbol{T}}_4^{-1}$ assumes the form

$$\begin{aligned}
\hat{\boldsymbol{T}}_1^{-1} &= \Big[\frac{\boldsymbol{t}_{11}}{||\boldsymbol{t}_{11}||}, \frac{\boldsymbol{t}_{12}}{||\boldsymbol{t}_{12}||}, \cdots, \frac{\boldsymbol{t}_{1m}}{||\boldsymbol{t}_{1m}||}\Big]\\
\hat{\boldsymbol{T}}_4^{-1} &= \Big[\frac{\boldsymbol{t}_{41}}{||\boldsymbol{t}_{41}||}, \frac{\boldsymbol{t}_{42}}{||\boldsymbol{t}_{42}||}, \cdots, \frac{\boldsymbol{t}_{4n}}{||\boldsymbol{t}_{4n}||}\Big].
\end{aligned} \tag{15}$$

Substituting (13) into (12), we obtain

$$\begin{aligned}
J_{e3}(\boldsymbol{D},\hat{\boldsymbol{T}}) = \operatorname{tr}\Big[&\hat{\boldsymbol{W}}_{10} - \hat{\boldsymbol{c}}_1^T\hat{\boldsymbol{c}}_1 - 2\sum_{k=0}^{N-1}\hat{\boldsymbol{W}}_{1,k+1}\boldsymbol{D}_{1k} \\
&+ \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\hat{\boldsymbol{W}}_{1,|k-l|}\boldsymbol{D}_{1k}\boldsymbol{D}_{1l}\Big] \\
&+\operatorname{tr}\Big[\hat{\boldsymbol{W}}_{40} - \hat{\boldsymbol{c}}_2^T\hat{\boldsymbol{c}}_2 - 2\sum_{k=0}^{N-1}\hat{\boldsymbol{W}}_{4,k+1}\boldsymbol{D}_{4k} \\
&+ \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\hat{\boldsymbol{W}}_{4,|k-l|}\boldsymbol{D}_{4k}\boldsymbol{D}_{4l}\Big]
\end{aligned} \tag{16}$$

where

$$\begin{aligned}
\hat{\boldsymbol{W}}_{1r} &= \hat{\boldsymbol{T}}_1\boldsymbol{K}_{c1}^{\frac{1}{2}}\boldsymbol{W}_{1r}\boldsymbol{K}_{c1}^{\frac{1}{2}}\hat{\boldsymbol{T}}_1^T, \qquad \hat{\boldsymbol{c}}_1 = \boldsymbol{c}_1\boldsymbol{K}_{c1}^{\frac{1}{2}}\hat{\boldsymbol{T}}_1^T \\
\hat{\boldsymbol{W}}_{4r} &= \hat{\boldsymbol{T}}_4\boldsymbol{K}_{c4}^{\frac{1}{2}}\boldsymbol{W}_{4r}\boldsymbol{K}_{c4}^{\frac{1}{2}}\hat{\boldsymbol{T}}_4^T, \qquad \hat{\boldsymbol{c}}_2 = \boldsymbol{c}_2\boldsymbol{K}_{c4}^{\frac{1}{2}}\hat{\boldsymbol{T}}_4^T.
\end{aligned}$$

From the foregoing arguments, the problem of obtaining matrices $\boldsymbol{T}$ and $\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}$ that jointly minimize (12) subject to the l_2-scaling constraints in (11) is now converted into an unconstrained optimization problem of obtaining $\hat{\boldsymbol{T}}$ and $\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}$ that jointly minimize (16).

Let $\boldsymbol{x}$ be the column vector that collects the variables in $\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}$, $\boldsymbol{t}_{11}, \boldsymbol{t}_{12}, \cdots, \boldsymbol{t}_{1m}$ and $\boldsymbol{t}_{41}, \boldsymbol{t}_{42}, \cdots, \boldsymbol{t}_{4n}$. Then $J_{e3}(\boldsymbol{D},\hat{\boldsymbol{T}})$ in (16) is a function of $\boldsymbol{x}$, denoted by $J(\boldsymbol{x})$. The proposed algorithm starts with the point $\boldsymbol{x}_0$ obtained from the assignment $\boldsymbol{T} = \boldsymbol{D}_0 = \boldsymbol{D}_1 = \cdots = \boldsymbol{D}_{N-1} = \boldsymbol{I}_{m+n}$. In the kth iteration, a quasi-Newton algorithm updates the most recent point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as [24]

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{d}_k, \tag{17}$$

where

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

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 $J(\boldsymbol{x}_k)$. This iteration process continues until

$$|J(\boldsymbol{x}_{k+1}) - J(\boldsymbol{x}_k)| < \varepsilon \tag{18}$$

is satisfied where $\varepsilon > 0$ is a prescribed tolerance.

We now consider the case where the high-order EF matrices are diagonal, i.e.,

$$\begin{aligned}
\boldsymbol{D}_{1k} &= \operatorname{diag}\{\alpha_{k1}, \alpha_{k2}, \cdots, \alpha_{km}\} \\
\boldsymbol{D}_{4k} &= \operatorname{diag}\{\beta_{k1}, \beta_{k2}, \cdots, \beta_{kn}\}
\end{aligned} \tag{19}$$

for $k = 0, 1, \ldots, N-1$. Then the gradients of $J(\boldsymbol{x})$ with respect to $\boldsymbol{t}_{11}, \boldsymbol{t}_{12}, \cdots, \boldsymbol{t}_{1m}, \boldsymbol{t}_{41}, \boldsymbol{t}_{42}, \cdots, \boldsymbol{t}_{4n}$ are found to be

$$\begin{aligned}
\frac{\partial J(\boldsymbol{x})}{\partial t_{1ij}} = 2\boldsymbol{e}_{1j}^T\Big[&\hat{\boldsymbol{W}}_{10} - \hat{\boldsymbol{c}}_1^T\hat{\boldsymbol{c}}_1 - \sum_{k=0}^{N-1}(\hat{\boldsymbol{W}}_{1,k+1} + \hat{\boldsymbol{W}}_{1,k+1}^T)\boldsymbol{D}_{1k} \\
&+ \frac{1}{2}\sum_{k=0}^{N-1}\sum_{l=0}^{N-1}(\hat{\boldsymbol{W}}_{1,|k-l|} + \hat{\boldsymbol{W}}_{1,|k-l|}^T)\boldsymbol{D}_{1k}\boldsymbol{D}_{1l}\Big]\hat{\boldsymbol{T}}_1\boldsymbol{g}_{1ij}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial J(\boldsymbol{x})}{\partial t_{4ij}} = 2\boldsymbol{e}_{4j}^T\Big[&\hat{\boldsymbol{W}}_{40} - \hat{\boldsymbol{c}}_2^T\hat{\boldsymbol{c}}_2 - \sum_{k=0}^{N-1}(\hat{\boldsymbol{W}}_{4,k+1} + \hat{\boldsymbol{W}}_{4,k+1}^T)\boldsymbol{D}_{4k} \\
&+ \frac{1}{2}\sum_{k=0}^{N-1}\sum_{l=0}^{N-1}(\hat{\boldsymbol{W}}_{4,|k-l|} + \hat{\boldsymbol{W}}_{4,|k-l|}^T)\boldsymbol{D}_{4k}\boldsymbol{D}_{4l}\Big]\hat{\boldsymbol{T}}_4\boldsymbol{g}_{4ij}
\end{aligned} \tag{20}$$

where t_{1ij} (t_{4ij}) is the ith element of vector $\boldsymbol{t}_{1j}$ ($\boldsymbol{t}_{4j}$) and

$$\begin{aligned}
\boldsymbol{g}_{1ij} &= \partial\left\{\frac{\boldsymbol{t}_{1j}}{||\boldsymbol{t}_{1j}||}\right\}/\partial t_{1ij} = \frac{1}{||\boldsymbol{t}_{1j}||^3}(t_{1ij}\boldsymbol{t}_{1j} - ||\boldsymbol{t}_{1j}||^2\boldsymbol{e}_{1i}) \\
\boldsymbol{g}_{4ij} &= \partial\left\{\frac{\boldsymbol{t}_{4j}}{||\boldsymbol{t}_{4j}||}\right\}/\partial t_{4ij} = \frac{1}{||\boldsymbol{t}_{4j}||^3}(t_{4ij}\boldsymbol{t}_{4j} - ||\boldsymbol{t}_{4j}||^2\boldsymbol{e}_{4i})
\end{aligned}$$

with $1 \le i,j \le m$ ($1 \le i,j \le n$) for t_{1ij}, $\boldsymbol{g}_{1ij}$ and $\boldsymbol{e}_{1i}$ (t_{4ij}, $\boldsymbol{g}_{4ij}$ and $\boldsymbol{e}_{4i}$). The gradients of $J(\boldsymbol{x})$ with respect to the diagonal $\boldsymbol{D}_{1k}$ and $\boldsymbol{D}_{4k}$ are given by

$$\begin{aligned}
\frac{\partial J(\boldsymbol{x})}{\partial \alpha_{kp}} &= -2(\hat{\boldsymbol{W}}_{1,k+1})_{pp} + 2\sum_{l=0}^{N-1}\beta_{lp}(\hat{\boldsymbol{W}}_{1,|k-l|})_{pp} \\
\frac{\partial J(\boldsymbol{x})}{\partial \beta_{kq}} &= -2(\hat{\boldsymbol{W}}_{4,k+1})_{qq} + 2\sum_{l=0}^{N-1}\alpha_{lq}(\hat{\boldsymbol{W}}_{4,|k-l|})_{qq}
\end{aligned} \tag{21}$$

for $k = 0, 1, \cdots, N-1$, $p = 1, 2, \cdots, m$ and $q = 1, 2, \cdots, n$.

4. AN ILLUSTRATIVE EXAMPLE

As a numerical example, consider a 2-D state-space digital filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{2+2}$ described by

$$\boldsymbol{A} = \begin{bmatrix} 1.88899 & -0.91219 & -1.00000 & 0 \\ 1 & 0 & 0 & 0 \\ 0.02771 & -0.02580 & 1.88899 & 1 \\ -0.02580 & 0.02431 & -0.91219 & 0 \end{bmatrix}$$

$$\boldsymbol{b} = \begin{bmatrix} 0.219089 & 0 & -0.028889 & 0.091219 \end{bmatrix}^T$$

$$\boldsymbol{c} = \begin{bmatrix} 0.028889 & -0.091219 & -0.219089 & 0 \end{bmatrix}$$

and $d = 0.87021$. The noise gain of the filter with no error feedforward and no EF was computed from (8) as

$$J_{e1}(\boldsymbol{0},\boldsymbol{0}) = \operatorname{tr}[\boldsymbol{W}_{10}] \oplus \operatorname{tr}[\boldsymbol{W}_{40}] = 176.452876.$$

Under the l_2-scaling constraints: $(\boldsymbol{T}_o^{-1}\boldsymbol{K}_c\boldsymbol{T}_o^{-T})_{ii} = 1$ for $i = 1, 2, 3, 4$ the above noise gain was changed to

$$\overline{J}_{e1}(\boldsymbol{0},\boldsymbol{0}) = \operatorname{tr}[\boldsymbol{T}_o^T(\boldsymbol{W}_{10} \oplus \boldsymbol{W}_{40})\boldsymbol{T}_o] = 367.51003$$

where $\boldsymbol{T}_o = \operatorname{diag}\{9.336620\ \ 9.336620\ \ 1.065113\ \ 0.986652\}$.

With the EF order set to $N = 2$, the quasi-Newton algorithm was applied to minimize (16) with tolerance $\varepsilon = 10^{-8}$ in (18). It took the algorithm 20 iterations to converge to the solution

$$\begin{aligned}
\hat{\boldsymbol{T}} &= \begin{bmatrix} 0.949460 & -0.381269 \\ 0.573481 & 0.847325 \end{bmatrix} \oplus \begin{bmatrix} 1.187712 & 0.543967 \\ 0.347831 & 1.259196 \end{bmatrix} \\
\boldsymbol{D}_0 &= \operatorname{diag}\{\, 1.904500 \quad 1.055826 \quad 1.348885 \quad 1.056108 \,\} \\
\boldsymbol{D}_1 &= -\operatorname{diag}\{\, 0.928480 \quad 0.141394 \quad 0.408198 \quad 0.167253 \,\}
\end{aligned}$$

whose noise gain was found to be $J_{e3}(\boldsymbol{D},\hat{\boldsymbol{T}}) = 1.055408$. The profile of $J_{e3}(\boldsymbol{D},\hat{\boldsymbol{T}})$ during the first 20 iterations of the algorithm is dipicted in Fig. 1.

To see how the use of *high-order* EF effects on the performance in noise reduction, simulations have also been

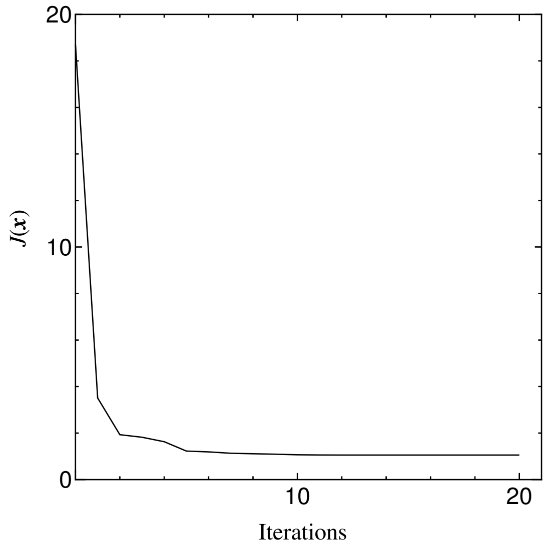

Fig. 1. Profile of $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}})$ during the first 20 iterations.

carried out with $N = 1$, i.e. $\boldsymbol{D} = \boldsymbol{D}_0$. The minimal noise gain in this case was found to be $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}}) = 1.250326$ which is considerably greater than what the system can achieve with $N = 2$. These and additional simulation results regarding the noise gain $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}}) + (\overline{\boldsymbol{c}} - \boldsymbol{h})(\overline{\boldsymbol{c}} - \boldsymbol{h})^T$ are summarized in Table I where the term "Infinite Precision" refers to the value of $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}})$ derived from the optimal $\boldsymbol{D}$ and $\hat{\boldsymbol{T}}$, the term "3-Bit Quantization" means that of $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}}) + (\overline{\boldsymbol{c}} - \boldsymbol{h})(\overline{\boldsymbol{c}} - \boldsymbol{h})^T$ where only each entry of the optimal $\boldsymbol{D}$ and $\boldsymbol{h}$ was rounded to a power-of-two representation with 3 bits after the binary point, and the term "Integer Quantization" means that of $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}}) + (\overline{\boldsymbol{c}} - \boldsymbol{h})(\overline{\boldsymbol{c}} - \boldsymbol{h})^T$ where only each entry of the optimal $\boldsymbol{D}$ and $\boldsymbol{h}$ was rounded to an integer.

TABLE I
PERFORMANCE COMPARISON

N	$\boldsymbol{D}_{1k} \oplus \boldsymbol{D}_{4k}$	Infinite Precision	3-Bit Quantization	Integer Quantization
1	Diagonal	1.250326	1.275315	1.687610
2	Diagonal	1.055408	1.087588	1.434863

5. CONCLUSION

For 2-D state-space digital filters described by the Roesser model [23], the joint optimization problem of *high-order* EF and realization to minimize the effects of roundoff noise subject to l_2-scaling constraints has been investigated. Linear algebraic techniques have been employed to convert the constrained optimization problem at hand into an unconstrained optimization problem, which is solved iteratively by applying an efficient quasi-Newton algorithm. Closed-form formulas for fast evaluation of the gradient of the objective function have also been derived. A numerical example has demonstrated the validity and effectiveness of the proposed technique.

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.

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

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] T. Hinamoto, A. Doi and W.-S. Lu, "Joint optimization of high-order error-feedback and realization for roundoff noise minimization in the Fornasini-Marchesini second model," *Proc. IEEE Int. Midwest Symp. Circuits Syst.* (MWSCAS 2012), pp.1120-1123, Aug. 2012.

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

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