Roundoff Noise Minimization for 2-D Separable-Denominator Digital Filters Using Jointly Optimal High-Order Error Feedback and Realization

Takao Hinamoto
Hiroshima University
Higashi-Hiroshima 739-8527, Japan
Email: hinamoto@ieee.org

Akimitsu Doi
Hiroshima Institute of Technology
Hiroshima 731-5193, Japan
Email: 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 the filter output subject to l_2-scaling constraints for two-dimensional (2-D) separable-denominator digital filters is investigated. 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. Closed-form formulas for fast and accurate gradient evaluation are derived. A numerical example is presented to demonstrate the validity and effectiveness of the proposed technique.**

I. Introduction

When IIR digital filters are implemented with fixed-point arithmetic, an important issue is reducing the roundoff noise at the filter's output. Some techniques for synthesizing the optimal state-space model that minimizes the roundoff noise subject to l_2-scaling constraints have been proposed [1]-[3]. In this regards, error feedback (EF) is known as an effective tool to reduce finite-word-length (FWL) effects in IIR digital filters, and many EF methods have been explored in the past [4]-[6]. As a natural extension of the aforementioned methods, new methods that combine EF and state-space realization have been developed to achieve better performance [7]-[10]. That is, separately and jointly optimized scalar or general EF matrices for state-space digital filters have been explored [7]. A jointly optimized iterative technique with a general, diagonal, or scalar EF matrix using a quasi-Newton algorithm has been developed for state-space digital filters [8]. The similar technique is extended to 2-D digital filters with a general, block-diagonal, diagonal, or scalar EF matrix [9]. More recently, the noise reduction performance for 2-D state-space digital filters has been efficiently improved by applying high-order EF [10].

In this paper, the problem of jointly optimizing *high-order* EF and realization for minimizing the roundoff noise subject to l_2-scaling constraints for 2-D separable-denominator digital filters is investigated. We convert the constrained optimization problem into an unconstrained optimization problem using linear-algebraic techniques. A quasi-Newton algorithm is then utilized to solve the unconstrained optimization problem, and is applied to the case where the *high-order* EF matrices are diagonal. A numerical example is presented to demonstrate the validity and effectiveness of the proposed technique.

II. Problem Statement

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

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

with

$$\boldsymbol{x}_k(i,j) = \begin{bmatrix} \boldsymbol{x}^h(i+k,j) \\ \boldsymbol{x}^v(i,j+k) \end{bmatrix}, \quad \boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 \\ \boldsymbol{0} & \boldsymbol{A}_4 \end{bmatrix}$$

$$\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix}, \qquad \boldsymbol{c} = [\ \boldsymbol{c}_1 \quad \boldsymbol{c}_2\ ]$$

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 $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{A}_4$, $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, $\boldsymbol{c}_1$, $\boldsymbol{c}_2$ 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

$$\begin{aligned} \tilde{\boldsymbol{x}}_1(i,j) &= \boldsymbol{A}\boldsymbol{Q}[\tilde{\boldsymbol{x}}_0(i,j)] + \boldsymbol{b}u(i,j) \\ &\quad + \sum_{k=0}^{M-1}[\boldsymbol{D}_{1k} \oplus \boldsymbol{0}]\boldsymbol{e}_{-k}(i,j) + \sum_{l=0}^{N-1}[\boldsymbol{0} \oplus \boldsymbol{D}_{4l}]\boldsymbol{e}_{-l}(i,j) \\ \tilde{y}(i,j) &= \boldsymbol{c}\,\boldsymbol{Q}[\tilde{\boldsymbol{x}}_0(i,j)] + du(i,j) + \boldsymbol{h}\boldsymbol{e}_0(i,j) \end{aligned} \tag{2}$$

where $\boldsymbol{h}$ is a $1 \times (m+n)$ error-feedforward vector, $\boldsymbol{D}_{1k}$'s and $\boldsymbol{D}_{4l}$'s are $m \times m$ and $n \times n$ *high-order* EF matrices, respectively, and

$$\boldsymbol{e}_{-k}(i,j) = \begin{bmatrix} \boldsymbol{e}^h(i-k,j) \\ \boldsymbol{e}^v(i,j-k) \end{bmatrix}, \quad \boldsymbol{h} = [\ \boldsymbol{h}_1 \quad \boldsymbol{h}_2\ ]$$

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

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 $\tilde{\boldsymbol{x}}_0(i,j)$ to $(B-B_c)$-bit after the multiplications and additions, where the sign bit is not counted. Assume 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

$$\begin{aligned}\Delta \boldsymbol{x}_1(i,j) &= \boldsymbol{A}\Delta\boldsymbol{x}_0(i,j) + \boldsymbol{A}\boldsymbol{e}_0(i,j)\\ &\quad - \sum_{k=0}^{M-1}[\boldsymbol{D}_{1k}\oplus \boldsymbol{0}]\boldsymbol{e}_{-k}(i,j) - \sum_{l=0}^{N-1}[\boldsymbol{0}\oplus\boldsymbol{D}_{4l}]\boldsymbol{e}_{-l}(i,j)\\ \Delta y(i,j) &= \boldsymbol{c}\Delta\boldsymbol{x}_0(i,j) + (\boldsymbol{c}-\boldsymbol{h})\boldsymbol{e}_0(i,j)\end{aligned} \tag{3}$$

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 zero, we obtain

$$\Delta Y(z_1,z_2) = [\boldsymbol{H}_e^h(z_1), \boldsymbol{H}_e^v(z_1,z_2)]\boldsymbol{E}_0(z_1,z_2) \tag{4a}$$

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

$$\begin{aligned}\boldsymbol{H}_e^h(z_1) &= \boldsymbol{c}_1\sum_{i=0}^{\infty}\left(\boldsymbol{A}_1^i - \sum_{k=0}^{M-1}\boldsymbol{A}_1^{i-k-1}\boldsymbol{D}_{1k}\right)z_1^{-i} - \boldsymbol{h}_1\\ \boldsymbol{H}_e^v(z_1,z_2) &= [\boldsymbol{c}_2 + \boldsymbol{c}_1(z_1\boldsymbol{I}_m - \boldsymbol{A}_1)^{-1}\boldsymbol{A}_2]\\ &\quad\cdot\sum_{i=0}^{\infty}\left(\boldsymbol{A}_4^i - \sum_{k=0}^{N-1}\boldsymbol{A}_4^{i-k-1}\boldsymbol{D}_{4k}\right)z_2^{-i} - \boldsymbol{h}_2\end{aligned} \tag{4b}$$

with $\boldsymbol{A}_1^i = \boldsymbol{0}$ and $\boldsymbol{A}_4^i = \boldsymbol{0}$ for $i < 0$.

Based on above analysis, we can now define the normalized noise gain of the 2-D filter as

$$J_o(\boldsymbol{h},\boldsymbol{D}_1,\boldsymbol{D}_4) = J_o^h(\boldsymbol{h}_1,\boldsymbol{D}_1) + J_o^v(\boldsymbol{h}_2,\boldsymbol{D}_4) = \sigma_{out}^2/\sigma^2 \tag{5a}$$

where

$$\begin{aligned}J_o^h(\boldsymbol{h}_1,\boldsymbol{D}_1) &= \mathrm{tr}\Big[\frac{1}{2\pi j}\oint_{|z_1|=1}[\boldsymbol{H}_e^h(z_1)]^*\boldsymbol{H}_e^h(z_1)\frac{dz_1}{z_1}\Big]\\ J_o^v(\boldsymbol{h}_2,\boldsymbol{D}_4)&\\ = \mathrm{tr}\Big[\frac{1}{(2\pi j)^2}&\oint_{|z_1|=1}\oint_{|z_2|=1}[\boldsymbol{H}_e^v(z_1,z_2)]^*\boldsymbol{H}_e^v(z_1,z_2)\frac{dz_1dz_2}{z_1z_2}\Big]\\ \boldsymbol{D}_1 &= [\boldsymbol{D}_{10},\boldsymbol{D}_{11},\cdots,\boldsymbol{D}_{1,M-1}]\\ \boldsymbol{D}_4 &= [\boldsymbol{D}_{40},\boldsymbol{D}_{41},\cdots,\boldsymbol{D}_{4,N-1}]\end{aligned} \tag{5b}$$

and σ_{out}^2 is noise variance at the filter output. By substituting (4b) into (5b), we have

$$\begin{aligned}&J_o^h(\boldsymbol{h}_1,\boldsymbol{D}_1)\\ &= \mathrm{tr}\Bigg[\boldsymbol{W}^h - \sum_{k=0}^{M-1}\Big\{\boldsymbol{D}_{1k}^T\boldsymbol{W}^h\boldsymbol{A}_1^{k+1} + (\boldsymbol{A}_1^T)^{k+1}\boldsymbol{W}^h\boldsymbol{D}_{1k}\Big\}\\ &\quad + \sum_{k=0}^{M-1}\sum_{l=0}^{M-1}\boldsymbol{D}_{1k}^T\Big\{\boldsymbol{W}^h\boldsymbol{A}_1^{k-l} + (\boldsymbol{A}_1^T)^{l-k}\boldsymbol{W}^h\Big\}\boldsymbol{D}_{1l}\\ &\quad - \sum_{k=0}^{M-1}\boldsymbol{D}_{1k}^T\boldsymbol{W}^h\boldsymbol{D}_{1k} - 2\boldsymbol{h}_1^T\boldsymbol{c}_1 + \boldsymbol{h}_1^T\boldsymbol{h}_1\Bigg]\end{aligned} \tag{6}$$

where $\boldsymbol{W}^h$ is the horizontal observability Grammian of the filter in (1) that can be obtained by solving the Lyapunov equation

$$\boldsymbol{W}^h = \boldsymbol{A}_1^T\boldsymbol{W}^h\boldsymbol{A}_1 + \boldsymbol{c}_1^T\boldsymbol{c}_1$$

and

$$\begin{aligned}&J_o^v(\boldsymbol{h}_2,\boldsymbol{D}_4)\\ &= \mathrm{tr}\Bigg[\boldsymbol{W}^v - \sum_{k=0}^{N-1}\Big\{\boldsymbol{D}_{4k}^T\boldsymbol{W}^v\boldsymbol{A}_4^{k+1} + (\boldsymbol{A}_4^T)^{k+1}\boldsymbol{W}^v\boldsymbol{D}_{4k}\Big\}\\ &\quad + \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\boldsymbol{D}_{4k}^T\Big\{\boldsymbol{W}^v\boldsymbol{A}_4^{k-l} + (\boldsymbol{A}_4^T)^{l-k}\boldsymbol{W}^v\Big\}\boldsymbol{D}_{4l}\\ &\quad - \sum_{k=0}^{N-1}\boldsymbol{D}_{4k}^T\boldsymbol{W}^v\boldsymbol{D}_{4k} - 2\boldsymbol{h}_2^T\boldsymbol{c}_2 + \boldsymbol{h}_2^T\boldsymbol{h}_2\Bigg]\end{aligned} \tag{7}$$

where $\boldsymbol{W}^v$ is the vertical observability Grammian of the filter in (1) that can be obtained by solving the Lyapunov equation

$$\boldsymbol{W}^v = \boldsymbol{A}_4^T\boldsymbol{W}^v\boldsymbol{A}_4 + \boldsymbol{A}_2^T\boldsymbol{W}^h\boldsymbol{A}_2 + \boldsymbol{c}_2^T\boldsymbol{c}_2.$$

If the high-order EF matrices $\boldsymbol{D}_{1k}$ for $k = 0,1,\cdots,M-1$ and $\boldsymbol{D}_{4l}$ for $l = 0,1,\cdots,N-1$ are diagonal then (6) and (7) can be considerably simplified to

$$\begin{aligned}J^h(\boldsymbol{h}_1,\boldsymbol{D}_1) &= \mathrm{tr}\Bigg[\boldsymbol{W}^h - \boldsymbol{c}_1^T\boldsymbol{c}_1 - 2\sum_{k=0}^{M-1}\boldsymbol{W}^h\boldsymbol{A}_1^{k+1}\boldsymbol{D}_{1k}\\ &\quad + \sum_{k=0}^{M-1}\sum_{l=0}^{M-1}\boldsymbol{W}^h\boldsymbol{A}_1^{|k-l|}\boldsymbol{D}_{1k}\boldsymbol{D}_{1l}\Bigg] + (\boldsymbol{c}_1-\boldsymbol{h}_1)(\boldsymbol{c}_1-\boldsymbol{h}_1)^T\end{aligned} \tag{8}$$

and

$$\begin{aligned}J^v(\boldsymbol{h}_2,\boldsymbol{D}_4) &= \mathrm{tr}\Bigg[\boldsymbol{W}^v - \boldsymbol{c}_2^T\boldsymbol{c}_2 - 2\sum_{k=0}^{N-1}\boldsymbol{W}^v\boldsymbol{A}_4^{k+1}\boldsymbol{D}_{4k}\\ &\quad + \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\boldsymbol{W}^v\boldsymbol{A}_4^{|k-l|}\boldsymbol{D}_{4k}\boldsymbol{D}_{4l}\Bigg] + (\boldsymbol{c}_2-\boldsymbol{h}_2)(\boldsymbol{c}_2-\boldsymbol{h}_2)^T.\end{aligned} \tag{9}$$

respectively.

It should be noted that l_2-scaling constraints on the local state vector $\boldsymbol{x}_0(i,j)$ involve the local controllability Grammian $\boldsymbol{K} = \boldsymbol{K}^h \oplus \boldsymbol{K}^v$ of the filter in (1), which can be obtained by solving the Lyapunov equations [3]

$$\begin{aligned}\boldsymbol{K}^v &= \boldsymbol{A}_4\boldsymbol{K}^v\boldsymbol{A}_4^T + \boldsymbol{b}_2\boldsymbol{b}_2^T\\ \boldsymbol{K}^h &= \boldsymbol{A}_1\boldsymbol{K}^h\boldsymbol{A}_1^T + \boldsymbol{A}_2\boldsymbol{K}^v\boldsymbol{A}_2^T + \boldsymbol{b}_1\boldsymbol{b}_1^T.\end{aligned} \tag{10}$$

A different yet equivalent local state-space model 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

$$\overline{\boldsymbol{A}} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T},\quad \overline{\boldsymbol{b}} = \boldsymbol{T}^{-1}\boldsymbol{b},\quad \overline{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T}. \tag{11}$$

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

$$(\overline{\boldsymbol{K}})_{ii} = (\boldsymbol{T}^{-1}\boldsymbol{K}\boldsymbol{T}^{-T})_{ii} = 1 \quad \text{for} \quad i = 1,2,\cdots,m+n \tag{12}$$

which essentially normalize the energy of the input-to-state signal flow. We consider the problem of minimizing noise gain of the 2-D filter subject to unity input-to-state energy. Analytically, our objective is to design an optimal coordinate transformation matrix $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ as well as diagonal *high-order* EF matrices $\boldsymbol{D}_{1k}$ for $k = 0,1,\cdots,M-1$ and $\boldsymbol{D}_{4l}$ for $l = 0,1,\cdots,N-1$ that jointly minimize

$$\overline{J}(\boldsymbol{T},\boldsymbol{D}_1,\boldsymbol{D}_4) = \overline{J}^h(\boldsymbol{T}_1,\boldsymbol{D}_1) + \overline{J}^v(\boldsymbol{T}_4,\boldsymbol{D}_4) \tag{13a}$$

subject to the l_2-scaling constraints in (12) where

$$\begin{aligned}
&\overline{J}^h(\boldsymbol{T}_1, \boldsymbol{D}_1) \\
&= \mathrm{tr}\Bigg[\boldsymbol{T}_1^T(\boldsymbol{W}^h - \boldsymbol{c}_1^T\boldsymbol{c}_1)\boldsymbol{T}_1 - 2\sum_{k=0}^{M-1}\boldsymbol{T}_1^T\boldsymbol{W}^h\boldsymbol{A}_1^{k+1}\boldsymbol{T}_1\boldsymbol{D}_{1k} \\
&\quad + \sum_{k=0}^{M-1}\sum_{l=0}^{M-1}\boldsymbol{T}_1^T\boldsymbol{W}^h\boldsymbol{A}_1^{|k-l|}\boldsymbol{T}_1\boldsymbol{D}_{1k}\boldsymbol{D}_{1l}\Bigg] \\
&\overline{J}^v(\boldsymbol{T}_4, \boldsymbol{D}_4) \\
&= \mathrm{tr}\Bigg[\boldsymbol{T}_4^T(\boldsymbol{W}^v - \boldsymbol{c}_2^T\boldsymbol{c}_2)\boldsymbol{T}_4 - 2\sum_{k=0}^{N-1}\boldsymbol{T}_4^T\boldsymbol{W}^v\boldsymbol{A}_4^{k+1}\boldsymbol{T}_4\boldsymbol{D}_{4k} \\
&\quad + \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\boldsymbol{T}_4^T\boldsymbol{W}^v\boldsymbol{A}_4^{|k-l|}\boldsymbol{T}_4\boldsymbol{D}_{4k}\boldsymbol{D}_{4l}\Bigg]
\end{aligned} \tag{13b}$$

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

III. Joint Optimization of High-Order Error Feedback and Realization

To deal with the l_2-scaling constraints in (12), we define

$$\hat{\boldsymbol{T}} = \hat{\boldsymbol{T}}_1 \oplus \hat{\boldsymbol{T}}_4 = (\boldsymbol{T}_1 \oplus \boldsymbol{T}_4)^T(\boldsymbol{K}^h \oplus \boldsymbol{K}^v)^{-\frac{1}{2}} \tag{14}$$

which leads (12) to

$$(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{T}}^{-1})_{ii} = 1 \quad \text{for} \quad i = 1, 2, \cdots, m+n. \tag{15}$$

These constraints are always satisfied if matrices $\hat{\boldsymbol{T}}_1^{-1}$ and $\hat{\boldsymbol{T}}_4^{-1}$ assume the form

$$\begin{aligned}
\hat{\boldsymbol{T}}_1^{-1} &= \left[\frac{\boldsymbol{t}_{11}}{||\boldsymbol{t}_{11}||}, \frac{\boldsymbol{t}_{12}}{||\boldsymbol{t}_{12}||}, \cdots, \frac{\boldsymbol{t}_{1m}}{||\boldsymbol{t}_{1m}||}\right] \\
\hat{\boldsymbol{T}}_4^{-1} &= \left[\frac{\boldsymbol{t}_{41}}{||\boldsymbol{t}_{41}||}, \frac{\boldsymbol{t}_{42}}{||\boldsymbol{t}_{42}||}, \cdots, \frac{\boldsymbol{t}_{4n}}{||\boldsymbol{t}_{4n}||}\right].
\end{aligned} \tag{16}$$

By substituting (14) into (13), we obtain

$$\begin{aligned}
\hat{J}(\hat{\boldsymbol{T}}, \boldsymbol{D}_1, \boldsymbol{D}_4) = \mathrm{tr}&\Bigg[\boldsymbol{U}_0 - \hat{\boldsymbol{c}}_1^T\hat{\boldsymbol{c}}_1 - 2\sum_{k=0}^{M-1}\boldsymbol{U}_{k+1}\boldsymbol{D}_{1k} \\
&+ \sum_{k=0}^{M-1}\sum_{l=0}^{M-1}\boldsymbol{U}_{|k-l|}\boldsymbol{D}_{1k}\boldsymbol{D}_{1l}\Bigg] \\
+\,\mathrm{tr}&\Bigg[\boldsymbol{V}_0 - \hat{\boldsymbol{c}}_2^T\hat{\boldsymbol{c}}_2 - 2\sum_{k=0}^{N-1}\boldsymbol{V}_{k+1}\boldsymbol{D}_{4k} \\
&+ \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\boldsymbol{V}_{|k-l|}\boldsymbol{D}_{4k}\boldsymbol{D}_{4l}\Bigg]
\end{aligned} \tag{17}$$

where

$$\begin{aligned}
\boldsymbol{U}_k &= \hat{\boldsymbol{T}}_1(\boldsymbol{K}^h)^{\frac{1}{2}}\boldsymbol{W}^h\boldsymbol{A}_1^k(\boldsymbol{K}^h)^{\frac{1}{2}}\hat{\boldsymbol{T}}_1^T, \quad \hat{\boldsymbol{c}}_1 = \boldsymbol{c}_1(\boldsymbol{K}^h)^{\frac{1}{2}}\hat{\boldsymbol{T}}_1^T \\
\boldsymbol{V}_k &= \hat{\boldsymbol{T}}_4(\boldsymbol{K}^v)^{\frac{1}{2}}\boldsymbol{W}^v\boldsymbol{A}_4^k(\boldsymbol{K}^v)^{\frac{1}{2}}\hat{\boldsymbol{T}}_4^T, \quad \hat{\boldsymbol{c}}_2 = \boldsymbol{c}_2(\boldsymbol{K}^v)^{\frac{1}{2}}\hat{\boldsymbol{T}}_4^T
\end{aligned}$$

From the foregoing arguments, the problem of obtaining matrices $\boldsymbol{T}$, $\boldsymbol{D}_{1k}$ for $k = 0, 1, \cdots, M-1$ and $\boldsymbol{D}_{4l}$ for $l = 0, 1, \cdots, N-1$ that jointly minimize (13a) subject to l_2-scaling constraints in (12) is now converted into an unconstrained optimization problem of obtaining $\hat{\boldsymbol{T}}$, $\boldsymbol{D}_{1k}$ for $k = 0, 1, \cdots, M-1$ and $\boldsymbol{D}_{4l}$ for $l = 0, 1, \cdots, N-1$ that jointly minimize (17).

Let $\boldsymbol{x}$ be the column vector that collects the variables in $\boldsymbol{t}_{11}, \boldsymbol{t}_{12}, \cdots, \boldsymbol{t}_{1m}$, $\boldsymbol{t}_{41}, \boldsymbol{t}_{42}, \cdots, \boldsymbol{t}_{4n}$, $\boldsymbol{D}_{10}, \boldsymbol{D}_{11}, \cdots, \boldsymbol{D}_{1,M-1}$ and $\boldsymbol{D}_{40}, \boldsymbol{D}_{41}, \cdots, \boldsymbol{D}_{4,N-1}$. Then (17) is a function of $\boldsymbol{x}$, which we denote by $J(\boldsymbol{x})$.

The proposed algorithm starts with the initial point $\boldsymbol{x}_0$ obtained from the assignment $\boldsymbol{T}_1 = \boldsymbol{D}_{1k} = \boldsymbol{I}_m$ and $\boldsymbol{T}_4 = \boldsymbol{D}_{4l} = \boldsymbol{I}_n$ for any k and l. In the kth iteration, a quasi-Newton algorithm updates the most recent point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as [11]

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{d}_k, \tag{18}$$

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

$$|J(\boldsymbol{x}_{k+1}) - J(\boldsymbol{x}_k)| < \varepsilon \tag{19}$$

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

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

$$\begin{aligned}
\boldsymbol{D}_{1k} &= \mathrm{diag}\{\alpha_{k1}, \alpha_{k2}, \cdots, \alpha_{km}\}, \quad k = 0, 1, \ldots, M-1 \\
\boldsymbol{D}_{4l} &= \mathrm{diag}\{\beta_{l1}, \beta_{l2}, \cdots, \beta_{ln}\}, \quad l = 0, 1, \ldots, N-1.
\end{aligned} \tag{20}$$

In this case the gradients of $J(\boldsymbol{x})$ with respect to $\boldsymbol{t}_{11}, \boldsymbol{t}_{12}, \cdots, \boldsymbol{t}_{1m}$ and $\boldsymbol{t}_{41}, \boldsymbol{t}_{42}, \cdots, \boldsymbol{t}_{4n}$ can be evaluated efficiently using the closed-form formulas given below:

$$\begin{aligned}
\frac{\partial J(\boldsymbol{x})}{\partial t_{1ij}} &= 2\boldsymbol{e}_{1j}^T\Big[\boldsymbol{U}_0 - \hat{\boldsymbol{c}}_1^T\hat{\boldsymbol{c}}_1 - \sum_{k=0}^{M-1}(\boldsymbol{U}_{k+1} + \boldsymbol{U}_{k+1}^T)\boldsymbol{D}_{1k} \\
&\quad + \frac{1}{2}\sum_{k=0}^{M-1}\sum_{l=0}^{M-1}(\boldsymbol{U}_{|k-l|} + \boldsymbol{U}_{|k-l|}^T)\boldsymbol{D}_{1k}\boldsymbol{D}_{1l}\Big]\hat{\boldsymbol{T}}_1\boldsymbol{g}_{1ij} \\
\frac{\partial J(\boldsymbol{x})}{\partial t_{4ij}} &= 2\boldsymbol{e}_{4j}^T\Big[\boldsymbol{V}_0 - \hat{\boldsymbol{c}}_2^T\hat{\boldsymbol{c}}_2 - \sum_{k=0}^{N-1}(\boldsymbol{V}_{k+1} + \boldsymbol{V}_{k+1}^T)\boldsymbol{D}_{4k} \\
&\quad + \frac{1}{2}\sum_{k=0}^{N-1}\sum_{l=0}^{N-1}(\boldsymbol{V}_{|k-l|} + \boldsymbol{V}_{|k-l|}^T)\boldsymbol{D}_{4k}\boldsymbol{D}_{4l}\Big]\hat{\boldsymbol{T}}_4\boldsymbol{g}_{4ij}
\end{aligned} \tag{21}$$

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 $\boldsymbol{e}_{4i}$). Furthermore, the gradients of $J(\boldsymbol{x})$ with respect to the diagonal $\boldsymbol{D}_{1k}$ and $\boldsymbol{D}_{4l}$ are given by

$$\begin{aligned}
\frac{\partial J(\boldsymbol{x})}{\partial \alpha_{kp}} &= -2(\boldsymbol{U}_{k+1})_{pp} + 2\sum_{l=0}^{M-1}\alpha_{lp}(\boldsymbol{U}_{|k-l|})_{pp} \\
\frac{\partial J(\boldsymbol{x})}{\partial \beta_{lq}} &= -2(\boldsymbol{V}_{l+1})_{qq} + 2\sum_{k=0}^{N-1}\beta_{kq}(\boldsymbol{V}_{|k-l|})_{qq}
\end{aligned} \tag{22}$$

for $p = 1, 2, \cdots, m$ and $q = 1, 2, \cdots, n$.

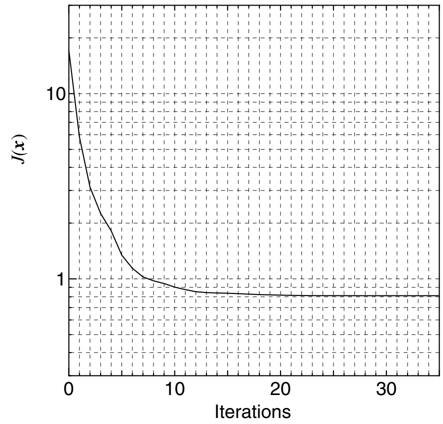

Fig. 1. Profile of $J(\boldsymbol{x})$ during the first 35 iterations.

IV. An Illustrative Example

To illustrate the algorithm proposed above, we consider a 2-D separable-denominator digital filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{3+3}$ in (1) specified by

$$\boldsymbol{A}_1^T = \begin{bmatrix} 0 & 0 & 0.599655 \\ 1 & 0 & -1.836929 \\ 0 & 1 & 2.173645 \end{bmatrix}, \quad \boldsymbol{b}_1 = \begin{bmatrix} 0.047053 \\ 0.062274 \\ 0.060436 \end{bmatrix}$$
$$\boldsymbol{A}_4 = \begin{bmatrix} 0 & 0 & 0.564961 \\ 1 & 0 & -1.887939 \\ 0 & 1 & 2.280029 \end{bmatrix}, \quad \boldsymbol{b}_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$
$$\boldsymbol{A}_2 = \begin{bmatrix} 0.064564 & 0.033034 & 0.012881 \\ 0.091213 & 0.110512 & 0.102759 \\ 0.097256 & 0.151864 & 0.172460 \end{bmatrix}, \quad \boldsymbol{c}_1^T = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$
$$\boldsymbol{c}_2 = \begin{bmatrix} 0.016556 & 0.012550 & 0.008243 \end{bmatrix}, \quad d = 0.019421.$$

Suppose both EF and error feedforward do not exist in (2), i.e., $\boldsymbol{D}_{1k} = \boldsymbol{0}$, $\boldsymbol{D}_{4l} = \boldsymbol{0}$ and $\boldsymbol{h} = \boldsymbol{0}$ for any k and l, then (17) is changed to

$$\hat{J}(\hat{\boldsymbol{T}}, \boldsymbol{0}, \boldsymbol{0}) = \mathrm{tr}[\boldsymbol{U}_{\boldsymbol{0}}] + \mathrm{tr}[\boldsymbol{V}_{\boldsymbol{0}}].$$

For this numerical example, the minimum value of $\hat{J}(\hat{\boldsymbol{T}}, \boldsymbol{0}, \boldsymbol{0})$ with respect to $\hat{\boldsymbol{T}}$ was computed as $\hat{J}_{min}(\hat{\boldsymbol{T}}, \boldsymbol{0}, \boldsymbol{0}) = 7.936533$.

With the EF order set to $M = N = 2$, the quasi-Newton algorithm was applied to minimize $J(\boldsymbol{x})$ with tolerance $\varepsilon = 10^{-8}$ in (19). It took the algorithm 35 iterations to converge to the solution

$$\hat{\boldsymbol{T}}_1 = \begin{bmatrix} 1.041139 & -0.047229 & 0.479350 \\ -0.030124 & 1.091829 & -0.081356 \\ 0.074994 & 0.468043 & 1.106818 \end{bmatrix}$$
$$\hat{\boldsymbol{T}}_4 = \begin{bmatrix} 0.980922 & 0.604023 & 0.247330 \\ -0.733642 & 0.908728 & -0.008450 \\ -1.046870 & 0.019431 & 0.845196 \end{bmatrix}$$
$$\boldsymbol{D}_{10} = \mathrm{diag}\{\, 0.808738 \quad 1.310301 \quad 1.575104 \,\}$$
$$\boldsymbol{D}_{11} = \mathrm{diag}\{\, 0.008971 \quad -0.708546 \quad -0.717346 \,\}$$
$$\boldsymbol{D}_{40} = \mathrm{diag}\{\, 1.553171 \quad 0.669799 \quad 1.121680 \,\}$$
$$\boldsymbol{D}_{41} = \mathrm{diag}\{\, -0.661891 \quad -0.100861 \quad -0.447027 \,\}$$

whose noise gain was found to be $J(\boldsymbol{x}) = 0.811238$. The profile of $J(\boldsymbol{x})$ during the first 35 iterations of the algorithm is depicted in Fig. 1.

The complete simulation results regarding the noise gain $J(\boldsymbol{x}) + (\overline{\boldsymbol{c}} - \boldsymbol{h})(\overline{\boldsymbol{c}} - \boldsymbol{h})^T$ are summarized in Table I where the term "Infinite Precision" refers to the value of $J(\boldsymbol{x})$ derived from the optimal $\hat{\boldsymbol{T}}$, $\boldsymbol{D}_1$ and $\boldsymbol{D}_4$, the term "3-Bit Quantization" means that of $J(\boldsymbol{x}) + (\overline{\boldsymbol{c}} - \boldsymbol{h})(\overline{\boldsymbol{c}} - \boldsymbol{h})^T$ where only each entry of the optimal $\boldsymbol{D}_1$, $\boldsymbol{D}_4$ 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 $J(\boldsymbol{x}) + (\overline{\boldsymbol{c}} - \boldsymbol{h})(\overline{\boldsymbol{c}} - \boldsymbol{h})^T$ where only each entry of the optimal $\boldsymbol{D}_1$, $\boldsymbol{D}_4$ and $\boldsymbol{h}$ was rounded to an integer.

TABLE I. Performance Comparison

(M, N)	Infinite Precision	3-Bit Quantization	Integer Quantization
$(1, 1)$	1.300436	1.351817	3.083187
$(2, 1)$	1.021734	1.067178	2.267652
$(1, 2)$	1.089940	1.134867	3.024715
$(2, 2)$	0.811238	0.850231	2.209052

V. Conclusion

An efficient technique for jointly optimizing *high-order* EF and realization to minimize the effects of roundoff noise subject to l_2-scaling constraints for 2-D separable-denominator digital filters has been developed. Unlike [10], the separately local controllability Grammian and separately local observability Grammian can readily be obtained in closed form by solving Lyapunov equations. Moreover, EF order of the horizontal state does not always need to be equal to that of the vertical state. A numerical example has demonstrated the validity and effectiveness of the proposed technique.

References

[1] 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.

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

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

[4] 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.

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

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

[7] 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.

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

[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 Process.*, vol. 54, pp. 4302-4310, Nov. 2006.

[10] T. Hinamoto, A. Doi and W.-S. Lu, "Jointly optimal high-order error-feedback and realization for roundoff noise minimization in 1-D and 2-D state-space digital filters," *IEEE Trans. Signal Process.*, vol. 61, pp. 5893-5903, Dec. 2013.

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