JOINT OPTIMIZATION OF HIGH-ORDER ERROR FEEDBACK AND REALIZATION FOR ROUNDOFF NOISE MINIMIZAION IN THE FORNASINI-MARCHESINI SECOND MODEL

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*— For two-dimensional (2-D) state-space digital filters described by the Fornasini-Marchesini second local state-space model, the joint optimization of high-order error feedback and realization for minimizing roundoff noise at filter output subject to l_2-scaling constraints is investigated. We present linear-algebraic techniques that convert the problem at hand into an unconstrained optimization problem, and present an efficient quasi-Newton algorithm to solve the unconstrained optimization problem iteratively, in which closed-form formulas are derived for fast and accurate gradient evaluation. A numerical example is presented to illustrate the utility and effectiveness of the proposed algorithm.**

1. INTRODUCTION

In the implementation of IIR digital filters in fixed-point arithmetic, it is of critical significance 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 structures 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],[20]. Separately and jointly optimized scalar or general EF matrix for state-space filters have been explored in [19]. In [20], a quasi-Newton method for joint optimization of general, diagonal, or scalar EF matrix for state-space digital filters is proposed.

In this paper, the problem of jointly optimizing *high-order* EF and realization for 2-D state-space digital filters described by the Fornasini-Marchesini second model [21] to minimize the roundoff noise subject to l_2-scaling constraints is investigated. The constrained optimization problem encountered is converted into an unconstrained optimization problem by using linear-algebraic techniques. An efficient quasi-Newton algorithm [22] is utilized to solve the unconstrained optimization problem at hand. The proposed technique is applied to the case where the *high-order* EF has diagonal matrices. A numerical example is presented to illustrate the proposed algorithm and demonstrate its performance.

2. PROBLEM STATEMENT

Consider a stable, locally controllable and locally observable 2-D state-space digital filter $(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{c}, d)_n$ described by the Fornasini-Marchesini second model [21]

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

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 constant matrices of appropriate dimensions.

By taking into account the quantizations performed before matrix-vector multiplication, an FWL implementation of (1) with *high-order* EF can be obtained as

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

where $\boldsymbol{h}$ is a $1 \times n$ error-feedforward vector, $\boldsymbol{D}_{1k}$ and $\boldsymbol{D}_{2k}$ for $k = 1, 2, \cdots, N$ are referred to as $n \times n$ *high-order* EF diagonal matrices, and $\boldsymbol{e}(i,j) = \tilde{\boldsymbol{x}}(i,j) - \boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,j)]$. The coefficient matrices $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, $\boldsymbol{c}$ and d in (2) are assumed to have exact fractional B_c-bit representations. The FWL local state vector $\tilde{\boldsymbol{x}}(i,j)$ and the output $\tilde{y}(i,j)$ all have 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 $\tilde{\boldsymbol{x}}(i,j)$ to $(B - B_c)$-bit after the multiplications and additions, where the sign bit is not counted. It is assumed that the roundoff error $\boldsymbol{e}(i,j)$ can be modeled as a zero-mean noise process with covariance $\sigma^2\boldsymbol{I}_n$. By subtracting (2) from (1), we obtain

$$\begin{aligned} \Delta\boldsymbol{x}(i,j) &= \boldsymbol{A}_1\Delta\boldsymbol{x}(i-1,j) + \boldsymbol{A}_2\Delta\boldsymbol{x}(i,j-1) \\ &\quad +\boldsymbol{A}_1\boldsymbol{e}(i-1,j) + \boldsymbol{A}_2\boldsymbol{e}(i,j-1) \\ &\quad -\sum_{k=1}^{N}\{\boldsymbol{D}_{1k}\boldsymbol{e}(i-k,j) + \boldsymbol{D}_{2k}\boldsymbol{e}(i,j-k)\} \\ \Delta y(i,j) &= \boldsymbol{c}\Delta\boldsymbol{x}(i,j) + (\boldsymbol{c} - \boldsymbol{h})\boldsymbol{e}(i,j) \end{aligned} \tag{3}$$

978-1-4673-2527-1/12/$31.00 ©2012 IEEE

where $\Delta\boldsymbol{x}(i,j) = \boldsymbol{x}(i,j) - \tilde{\boldsymbol{x}}(i,j)$ and $\Delta y(i,j) = y(i,j) - \tilde{y}(i,j)$. By taking the (z_1, z_2)-transform on both sides of (3) and setting the boundary conditions $\Delta\boldsymbol{x}(i,0) = \Delta\boldsymbol{x}(0,j) = \boldsymbol{0}$ for $i, j = 1, 2, \cdots$, we have

$$\begin{aligned} \Delta Y(z_1,z_2) &= \boldsymbol{H}_e(z_1,z_2)\boldsymbol{E}(z_1,z_2) \\ \boldsymbol{H}_e(z_1,z_2) &= \boldsymbol{c}(\boldsymbol{I}_n - z_1^{-1}\boldsymbol{A}_1 - z_2^{-1}\boldsymbol{A}_2)^{-1} \\ &\quad \cdot\Big(\boldsymbol{I}_n - \sum_{k=1}^{N}\{z_1^{-k}\boldsymbol{D}_{1k} + z_2^{-k}\boldsymbol{D}_{2k}\}\Big) - \boldsymbol{h} \end{aligned} \tag{4}$$

where $\Delta Y(z_1,z_2)$ and $\boldsymbol{E}(z_1,z_2)$ are the (z_1,z_2)-transforms of $\Delta y(i,j)$ and $\boldsymbol{e}(i,j)$, respectively. Defining the transition matrix $\boldsymbol{A}^{(i,j)}$, the noise transfer function $\boldsymbol{H}_e(z_1,z_2)$ can be written as

$$\begin{aligned} \boldsymbol{H}_e(z_1,z_2) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{c}\boldsymbol{A}^{(i,j)}z_1^{-i}z_2^{-j} \\ &\quad \cdot\Big(\boldsymbol{I}_n - \sum_{k=1}^{N}\{z_1^{-k}\boldsymbol{D}_{1k} + z_2^{-k}\boldsymbol{D}_{2k}\}\Big) - \boldsymbol{h} \end{aligned} \tag{5}$$

where

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

We define the normalized noise gain $J_{e1}(\boldsymbol{h},\boldsymbol{D}_1,\boldsymbol{D}_2) = \sigma_{out}^2/\sigma^2$ with $\boldsymbol{D}_r = [\boldsymbol{D}_{r1},\boldsymbol{D}_{r2},\cdots,\boldsymbol{D}_{rN}]$ for $r=1,2$ as

$$\begin{aligned} &J_{e1}(\boldsymbol{h},\boldsymbol{D}_1,\boldsymbol{D}_2) \\ &= \mathrm{tr}\Big[\frac{1}{(2\pi j)^2}\oint_{|z_1|=1}\oint_{|z_2|=1}\boldsymbol{H}_e^*(z_1,z_2)\boldsymbol{H}_e(z_1,z_2)\frac{dz_1}{z_1}\frac{dz_2}{z_2}\Big] \end{aligned} \tag{6}$$

Substituting (5) into (6) yields

$$\begin{aligned} J_{e1}(\boldsymbol{h},\boldsymbol{D}_1,\boldsymbol{D}_2) = \mathrm{tr}\Big[&\boldsymbol{W}_o - 2\sum_{k=1}^{N}\{\boldsymbol{W}'_{k0}\boldsymbol{D}_{1k} + \boldsymbol{W}'_{0k}\boldsymbol{D}_{2k}\} \\ &+ \sum_{k=1}^{N}\sum_{l=1}^{N}\{\boldsymbol{W}'_{k-l,0}\boldsymbol{D}_{1k}\boldsymbol{D}_{1l} + \boldsymbol{W}'_{0,k-l}\boldsymbol{D}_{2k}\boldsymbol{D}_{2l} \\ &+ 2\boldsymbol{W}''_{kl}\boldsymbol{D}_{1k}\boldsymbol{D}_{2l}\} - \boldsymbol{c}^T\boldsymbol{c} + (\boldsymbol{c}-\boldsymbol{h})^T(\boldsymbol{c}-\boldsymbol{h})\Big] \end{aligned} \tag{7}$$

where

$$\begin{aligned} \boldsymbol{W}_o &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}(\boldsymbol{c}\boldsymbol{A}^{(i,j)})^T\boldsymbol{c}\boldsymbol{A}^{(i,j)} \\ \boldsymbol{W}'_{kl} &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}(\boldsymbol{c}\boldsymbol{A}^{(i+k,j+l)})^T\boldsymbol{c}\boldsymbol{A}^{(i,j)} \\ \boldsymbol{W}''_{kl} &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}(\boldsymbol{c}\boldsymbol{A}^{(i+k,j)})^T\boldsymbol{c}\boldsymbol{A}^{(i,j+l)}. \end{aligned}$$

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

$$\boldsymbol{K}_c = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{f}(i,j)\boldsymbol{f}^T(i,j) \tag{8}$$

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

A different yet equivalent local state-space description of (1), $(\overline{\boldsymbol{A}}_1,\overline{\boldsymbol{A}}_2,\overline{\boldsymbol{b}}_1,\overline{\boldsymbol{b}}_2,\overline{\boldsymbol{c}},d)_n$, can be obtained via a coordinate transformation $\overline{\boldsymbol{x}}(i,j) = \boldsymbol{T}^{-1}\boldsymbol{x}(i,j)$ where

$$\begin{aligned} \overline{\boldsymbol{A}}_1 &= \boldsymbol{T}^{-1}\boldsymbol{A}_1\boldsymbol{T}, \quad \overline{\boldsymbol{A}}_2 = \boldsymbol{T}^{-1}\boldsymbol{A}_2\boldsymbol{T} \\ \overline{\boldsymbol{b}}_1 &= \boldsymbol{T}^{-1}\boldsymbol{b}_1, \quad \overline{\boldsymbol{b}}_2 = \boldsymbol{T}^{-1}\boldsymbol{b}_2, \quad \overline{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T}. \end{aligned} \tag{9}$$

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

$$(\overline{\boldsymbol{K}}_c)_{ii} = (\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T})_{ii} = 1, \qquad i = 1,2,\cdots,n. \tag{10}$$

The problem being considered here is to design an optimal coordinate transformation matrix $\boldsymbol{T}$ as well as *high-order* EF diagonal matrices $\boldsymbol{D}_{r1},\boldsymbol{D}_{r2},\cdots,\boldsymbol{D}_{rN}$ for $r=1,2$ that jointly minimize the noise gain

$$\begin{aligned} &J_{e2}(\boldsymbol{T},\boldsymbol{D}_1,\boldsymbol{D}_2) \\ &= \mathrm{tr}\Big[\boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{T} - 2\sum_{k=1}^{N}\{\boldsymbol{T}^T\boldsymbol{W}'_{k0}\boldsymbol{T}\boldsymbol{D}_{1k} + \boldsymbol{T}^T\boldsymbol{W}'_{0k}\boldsymbol{T}\boldsymbol{D}_{2k}\} \\ &\quad +\sum_{k=1}^{N}\sum_{l=1}^{N}\{\boldsymbol{T}^T\boldsymbol{W}'_{k-l,0}\boldsymbol{T}\boldsymbol{D}_{1k}\boldsymbol{D}_{1l} + \boldsymbol{T}^T\boldsymbol{W}'_{0,k-l}\boldsymbol{T}\boldsymbol{D}_{2k}\boldsymbol{D}_{2l} \\ &\qquad + 2\boldsymbol{T}^T\boldsymbol{W}''_{kl}\boldsymbol{T}\boldsymbol{D}_{1k}\boldsymbol{D}_{2l}\} - (\boldsymbol{c}\boldsymbol{T})^T\boldsymbol{c}\boldsymbol{T}\Big]. \end{aligned} \tag{11}$$

subject to l_2-scaling constraints in (10) where the error feed-forward 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 (10), we define

$$\hat{\boldsymbol{T}} = \boldsymbol{T}^T\boldsymbol{K}_c^{-\frac{1}{2}}. \tag{12}$$

Then (10) becomes

$$(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{T}}^{-1})_{ii} = 1, \quad i = 1,2,\cdots,n. \tag{13}$$

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

$$\hat{\boldsymbol{T}}^{-1} = \Big[\frac{\boldsymbol{t}_1}{||\boldsymbol{t}_1||},\frac{\boldsymbol{t}_2}{||\boldsymbol{t}_2||},\cdots,\frac{\boldsymbol{t}_n}{||\boldsymbol{t}_n||}\Big]. \tag{14}$$

Substituting (12) into (11), we obtain

$$\begin{aligned} J_{e3}(\hat{\boldsymbol{T}},\boldsymbol{D}_1,\boldsymbol{D}_2) = \mathrm{tr}\Big[&\hat{\boldsymbol{W}}_o - 2\sum_{k=1}^{N}\{\hat{\boldsymbol{W}}'_{k0}\boldsymbol{D}_{1k} + \hat{\boldsymbol{W}}'_{0k}\boldsymbol{D}_{2k}\} \\ &+\sum_{k=1}^{N}\sum_{l=1}^{N}\{\hat{\boldsymbol{W}}'_{k-l,0}\boldsymbol{D}_{1k}\boldsymbol{D}_{1l} + \hat{\boldsymbol{W}}'_{0,k-l}\boldsymbol{D}_{2k}\boldsymbol{D}_{2l} \\ &+ 2\hat{\boldsymbol{W}}''_{kl}\boldsymbol{D}_{1k}\boldsymbol{D}_{2l}\} - \hat{\boldsymbol{c}}^T\hat{\boldsymbol{c}}\Big] \end{aligned} \tag{15}$$

where

$$\begin{aligned} \hat{\boldsymbol{W}}_o &= \hat{\boldsymbol{T}}\boldsymbol{K}_c^{\frac{1}{2}}\boldsymbol{W}_o\boldsymbol{K}_c^{\frac{1}{2}}\hat{\boldsymbol{T}}^T, \qquad \hat{\boldsymbol{W}}'_{kl} = \hat{\boldsymbol{T}}\boldsymbol{K}_c^{\frac{1}{2}}\boldsymbol{W}'_{kl}\boldsymbol{K}_c^{\frac{1}{2}}\hat{\boldsymbol{T}}^T \\ \hat{\boldsymbol{W}}''_{kl} &= \hat{\boldsymbol{T}}\boldsymbol{K}_c^{\frac{1}{2}}\boldsymbol{W}''_{kl}\boldsymbol{K}_c^{\frac{1}{2}}\hat{\boldsymbol{T}}^T, \qquad \hat{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{K}_c^{\frac{1}{2}}\hat{\boldsymbol{T}}^T. \end{aligned}$$

From the foregoing arguments, the problem of obtaining matrices $\boldsymbol{T}$, $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ that jointly minimize (11) subject to the l_2-scaling constraints in (10) is now converted into an unconstrained optimization problem of obtaining $\hat{\boldsymbol{T}}$, $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ that jointly minimize $J_{e3}(\hat{\boldsymbol{T}}, \boldsymbol{D}_1, \boldsymbol{D}_2)$ in (15).

Let $\boldsymbol{x}$ be the column vector that collects the variables in matrices $[\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n]$, $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$. Then $J_{e3}(\hat{\boldsymbol{T}}, \boldsymbol{D}_1, \boldsymbol{D}_2)$ in (15) is a function of $\boldsymbol{x}$, denoted by $J(\boldsymbol{x})$. The proposed algorithm starts with an initial point $\boldsymbol{x}_0$ obtained from an initial assignment $\boldsymbol{T} = \boldsymbol{D}_{r1} = \boldsymbol{D}_{r2} = \cdots = \boldsymbol{D}_{rN} = \boldsymbol{I}_n$ for $r = 1, 2$. In the kth iteration, a quasi-Newton algorithm updates the most recent point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as [22]

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k, \tag{16}$$

where

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

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

$$|J(\boldsymbol{x}_{k+1}) - J(\boldsymbol{x}_k)| < \varepsilon \tag{17}$$

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

In what follows, we define for $k = 1, 2, \ldots, N$

$$\begin{aligned} \boldsymbol{D}_{1k} &= \text{diag}\{\alpha_{k1}, \alpha_{k2}, \cdots, \alpha_{kn}\} \\ \boldsymbol{D}_{2k} &= \text{diag}\{\beta_{k1}, \beta_{k2}, \cdots, \beta_{kn}\}. \end{aligned} \tag{18}$$

Then the gradient of $J(\boldsymbol{x})$ with respect to $\boldsymbol{T}$ is found to be

$$\begin{aligned} \frac{\partial J(\boldsymbol{x})}{\partial t_{ij}} &= 2\boldsymbol{e}_j^T \Big[\hat{\boldsymbol{W}}_o - \sum_{k=1}^{N} \{(\hat{\boldsymbol{W}}'_{k0} + \hat{\boldsymbol{W}}'^T_{k0})\boldsymbol{D}_{1k} \\ &+ (\hat{\boldsymbol{W}}'_{0k} + \hat{\boldsymbol{W}}'^T_{0k})\boldsymbol{D}_{2k}\} + \frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{N}\{2\,(\hat{\boldsymbol{W}}''_{kl} + \hat{\boldsymbol{W}}''^T_{kl})\boldsymbol{D}_{1k}\boldsymbol{D}_{2l} \\ &+ (\hat{\boldsymbol{W}}'_{k-l,0} + \hat{\boldsymbol{W}}'^T_{k-l,0})\boldsymbol{D}_{1k}\boldsymbol{D}_{1l} + (\hat{\boldsymbol{W}}'_{0,k-l} + \hat{\boldsymbol{W}}'^T_{0,k-l})\boldsymbol{D}_{2k}\boldsymbol{D}_{2l}\} \\ &- \hat{\boldsymbol{c}}^T\hat{\boldsymbol{c}}\Big]\hat{\boldsymbol{T}}\boldsymbol{g}_{ij} \qquad i, j = 1, 2, \cdots, n \end{aligned} \tag{19}$$

where

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

and the gradients of $J(\boldsymbol{x})$ with respect to the EF matrices $\boldsymbol{D}_{1k}$ and $\boldsymbol{D}_{2k}$ are given by

$$\begin{aligned} \frac{\partial J(\boldsymbol{x})}{\partial \alpha_{rp}} &= -2(\hat{\boldsymbol{W}}'_{r0})_{pp} + 2\sum_{l=1}^{N}\beta_{lp}(\hat{\boldsymbol{W}}''_{rl})_{pp} \\ &+ \sum_{l=1}^{N}\alpha_{lp}[(\hat{\boldsymbol{W}}'_{r-l,0})_{pp} + (\hat{\boldsymbol{W}}'_{l-r,0})_{pp}] \\ \frac{\partial J(\boldsymbol{x})}{\partial \beta_{rp}} &= -2(\hat{\boldsymbol{W}}'_{0r})_{pp} + 2\sum_{l=1}^{N}\alpha_{lp}(\hat{\boldsymbol{W}}''_{lr})_{pp} \\ &+ \sum_{l=1}^{N}\beta_{lp}[(\hat{\boldsymbol{W}}'_{0,r-l})_{pp} + (\hat{\boldsymbol{W}}'_{0,l-r})_{pp}] \\ & \qquad r = 1, 2, \cdots, N; \qquad p = 1, 2, \cdots, n. \end{aligned} \tag{20}$$

We remark that using the closed-form formulas given in (19) and (20) allows us to quickly and accurately evaluate gradient $\nabla J(\boldsymbol{x})$, which is a key quantity in updating the iterate via (16), hence ensures high efficiency of the proposed algorithm.

4. AN ILLUSTRATIVE EXAMPLE

As a numerical example, consider a 2-D state-space digital filter $(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{c}, d)_4$ described by

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

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

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

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

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

The local controllability Gramian $\boldsymbol{K}_c$ and the local observability Gramian $\boldsymbol{W}_o$ of the above filter were computed from (7) and (8) with truncation $(0, 0) \leq (i, j) \leq (100, 100)$ as

$$\boldsymbol{K}_c = \begin{bmatrix} 0.00877 & -0.01777 & 0.00506 & -0.02829 \\ -0.01777 & 0.04636 & -0.02382 & 0.06085 \\ 0.00506 & -0.02382 & 0.23071 & -0.45355 \\ -0.02829 & 0.06085 & -0.45355 & 1.05272 \end{bmatrix}$$

$$\boldsymbol{W}_o = 10^3 \begin{bmatrix} 1.52516 & 0.72461 & 0.35244 & 0.16613 \\ 0.72461 & 0.35320 & 0.17607 & 0.08413 \\ 0.35244 & 0.17607 & 0.09200 & 0.04605 \\ 0.16613 & 0.08413 & 0.04605 & 0.02539 \end{bmatrix}.$$

The noise gain of the filter with no error feedforward and no error feedback was then computed from (7) as

$$J_{e1}(\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0}) = \text{tr}[\boldsymbol{W}_o] = 1.995751 \times 10^3.$$

This noise gain was changed to

$$\overline{J}_{e1}(\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0}) = \text{tr}[\boldsymbol{T}_o^T \boldsymbol{W}_o \boldsymbol{T}_o] = 7.769460 \times 10$$

when the l_2-scaling constraints described by

$$(\boldsymbol{T}_o^{-1} \boldsymbol{K}_c \boldsymbol{T}_o^{-T})_{ii} = 1 \ \text{ for } \ i = 1, 2, 3, 4$$

are satisfied where

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

With the EF order set to $N = 2$, the quasi-Newton algorithm was applied to minimize (15) with tolerance $\varepsilon = 10^{-8}$ in (17). It took the algorithm 122 iterations to converge to the solution

$$\hat{\boldsymbol{T}} = \begin{bmatrix} 0.39346 & -0.63325 & 0.88325 & -0.04612 \\ 0.96045 & 0.42245 & 0.17845 & 0.07936 \\ -0.61172 & 0.52495 & 0.64699 & -0.73294 \\ 0.10613 & -0.34455 & 0.39887 & 1.20673 \end{bmatrix}$$

$$\boldsymbol{D}_{11} = \text{diag}\{\, 0.32201 \quad 0.52235 \quad 0.55642 \quad 0.72838 \,\}$$

$$\boldsymbol{D}_{12} = \text{diag}\{\, 0.14448 \quad -0.19827 \quad -0.25002 \quad -0.23110 \,\}$$

$$\boldsymbol{D}_{21}=\text{diag}\{\,0.44095 \quad 1.10838 \quad 0.40658 \quad 0.41164\,\}$$
$$\boldsymbol{D}_{22}=\text{diag}\{\,0.01968 \quad -0.49273 \quad 0.02778 \quad -0.00144\,\}$$

whose noise gain was found to be

$$J_{e3}(\hat{\boldsymbol{T}}, \boldsymbol{D}_1, \boldsymbol{D}_2) = 0.073287.$$

The profile of $J_{e3}(\hat{\boldsymbol{T}}, \boldsymbol{D}_1, \boldsymbol{D}_2)$ during the first 122 iterations of the algorithm is dipicted in Fig. 1.

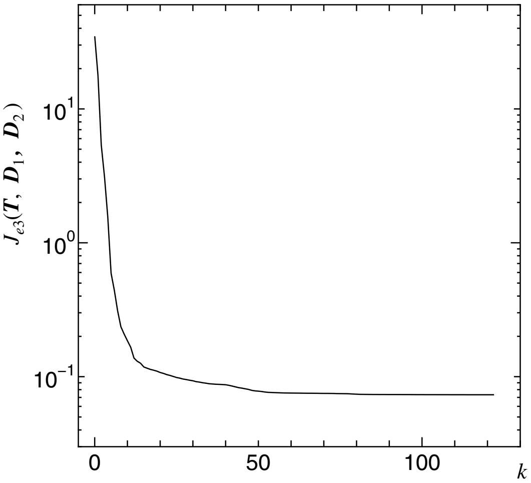

Fig. 1. Profile of $J_{e3}(\hat{\boldsymbol{T}}, \boldsymbol{D}_1, \boldsymbol{D}_2)$ during the first 122 iterations.

The other results regarding the noise gain $J_{e3}(\hat{\boldsymbol{T}}, \boldsymbol{D}_1, \boldsymbol{D}_2)$ in (15) were summarized in Table I where the term ”Infinite Precision” refers to the value of $J_{e3}(\hat{\boldsymbol{T}}, \boldsymbol{D}_1, \boldsymbol{D}_2)$ derived from the optimal $\hat{\boldsymbol{T}}$, $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$, and the term ”3-Bit Quantization” means that of $J_{e3}(\hat{\boldsymbol{T}}, \boldsymbol{D}_1, \boldsymbol{D}_2)$ where only each entry of the optimal matrix $\boldsymbol{D} = [\boldsymbol{D}_1, \boldsymbol{D}_2]$ was rounded to a power-of-two representation with 3 bits after the binary point.

TABLE I
PERFORMANCE COMPARISON

N	$\boldsymbol{D}_1, \boldsymbol{D}_2$	Infinite Precision	3-Bit Quantization
1	Diagonal	0.186190	0.216706
2	Diagonal	0.073287	0.099026

5. CONCLUSION

For 2-D digital filters described by the Fornasini-Marchesini second model, 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 problem at hand into an unconstrained optimization problem. The resultant unconstrained optimization problem has been solved iteratively by applying an efficient quasi-Newton algorithm. Moreover, closed-form formulas for fast evaluation of the gradient of the objective function have been derived. The proposed technique has been applied to the case where *high-order* EF has diagonal matrices. A numerical example has demonstrated the 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] E. Fornasini and G. Marchesini, “Doubly-indexed dynamical systems: State-space models and structural properties,” *Mathematical Systems Theory*, vol. 12, pp. 59-72, 1978.

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