JOINT OPTIMIZATION OF ERROR FEEDBACK AND COORDINATE TRANSFORMATION FOR ROUNDOFF NOISE MINIMIZATION IN 2-D STATE-SPACE DIGITAL FILTERS

Takao Hinamoto, Hiroaki Ohnishi and Wu-Sheng Lu†

Hiroshima University, Higashi-Hiroshima 739-8527, Japan. hinamoto@hiroshima-u.ac.jp
†University of Victoria, Victoria, BC, Canada V8W 3P6. wslu@ece.uvic.ca

ABSTRACT

The joint optimization of an error-feedback matrix and a coordinate-transformation matrix in 2-D state-space digital filters for roundoff noise minimization subject to L_2-norm dynamic-range scaling constraints is investigated. Using linear-algebraic techniques, the problem at hand is converted into an unconstrained optimization problem, and the unconstrained problem obtained is then solved by applying an efficient quasi-Newton algorithm.

I. INTRODUCTION

When implementing IIR digital filters in fixed-point arithmetic, the problem of reducing the effects of roundoff noise (RN) at the filter output is of critical importance. Error feedback (EF) is a useful tool for reducing finite-word-length (FWL) effects in IIR digital filters. Many EF techniques have been reported for 2-D IIR digital filters [1]-[5]. Another useful approach is to construct the 2-D state-space filter structure for the RN gain to be minimized by applying a linear transformation to the state-space coordinates subject to L_2-norm dynamic-range scaling constraints [6],[7]. As a natural extension of the fore-mentioned methods, efforts have been made to develop new mehods that combine the EF and coordinate transformation for achieving better performance in the RN reduction. In [8], jointly-optimized iterative algorithms have also been developed for the filter with a scalar or general EF matrix.

This paper investigates the problem of jointly optimizing the EF and the coordinate transformation in 2-D state-space digital filters so as to minimize the RN subject to L_2-norm dynamic-range scaling constraints. A jointly-optimized iterative technique, relying on an efficient quasi-Newton algorithm [9], is developed for RN minimization subject to the scaling constraints. The proposed technique can be applied to the cases where the error-feedback matrix is a scalar, diagonal, block-diagonal, or general matrix.

II. 2-D STATE-SPACE DIGITAL FILTERS WITH ERROR FEEDBACK

Consider the following local state-space (LSS) model $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{m,n}$ for 2-D IIR digital filters:

$$\begin{aligned} \boldsymbol{x}_{11}(i,j) &= \boldsymbol{A}\boldsymbol{x}(i,j) + \boldsymbol{b}u(i,j) \\ y(i,j) &= \boldsymbol{c}\boldsymbol{x}(i,j) + du(i,j) \end{aligned} \tag{1}$$

where

$$\boldsymbol{x}_{11}(i,j) = \begin{bmatrix} \boldsymbol{x}^h(i+1,j) \\ \boldsymbol{x}^v(i,j+1) \end{bmatrix}, \quad \boldsymbol{x}(i,j) = \begin{bmatrix} \boldsymbol{x}^h(i,j) \\ \boldsymbol{x}^v(i,j) \end{bmatrix}$$

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 \\ \boldsymbol{A}_3 & \boldsymbol{A}_4 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} \boldsymbol{c}_1 & \boldsymbol{c}_2 \end{bmatrix}.$$

Here, $\boldsymbol{x}^h(i,j)$ 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 matrices of appropriate dimensions. The LSS model (1) is assumed stable, separately locally controllable and observable.

Due to finite register sizes, FWL constraints are imposed on the local state vector $\boldsymbol{x}(i,j)$, input, output, and coefficients in the realization $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{m,n}$. Assuming that the quantization is performed before matrix-vector multiplication, the actual FWL filter of (1) with EF can be implemented as

$$\begin{aligned} \tilde{\boldsymbol{x}}_{11}(i,j) &= \boldsymbol{A}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,j)] + \boldsymbol{b}u(i,j) + \boldsymbol{D}\boldsymbol{e}(i,j) \\ \tilde{y}(i,j) &= \boldsymbol{c}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,j)] + du(i,j) \end{aligned} \tag{2}$$

where $\boldsymbol{D}$ is an $(m+n)\times(m+n)$ constant matrix referred to as *error feedback* (EF)*matrix*,

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

and each component of matrices $\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}$, and d assumes an exact fractional B_c-bit representation. The FWL local state vector $\tilde{\boldsymbol{x}}(i,j)$ and output $\tilde{y}(i,j)$ all have a

B-bit fractional representation, while the input $u(i,j)$ is a $(B-B_c)$-bit fraction. The quantizer $\boldsymbol{Q}[\cdot]$ in (2) rounds the B-bit fraction $\tilde{\boldsymbol{x}}(i,j)$ to $(B-B_c)$ bits after the multiplications and additions, where the sign bit is not counted. The quantization error $\boldsymbol{e}(i,j)$ is modeled as a zero-mean noise process of covariance $\sigma^2\boldsymbol{I}_{m+n}$ with

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

Subtracting (2) from (1) yields

$$\begin{aligned} \Delta\boldsymbol{x}_{11}(i,j) &= \boldsymbol{A}\Delta\boldsymbol{x}(i,j) + (\boldsymbol{A}-\boldsymbol{D})\boldsymbol{e}(i,j) \\ \Delta y(i,j) &= \boldsymbol{c}\Delta\boldsymbol{x}(i,j) + \boldsymbol{c}\boldsymbol{e}(i,j) \end{aligned} \tag{3}$$

where

$$\begin{aligned} \Delta\boldsymbol{x}(i,j) &= \boldsymbol{x}(i,j) - \tilde{\boldsymbol{x}}(i,j) \\ \Delta\boldsymbol{x}_{11}(i,j) &= \boldsymbol{x}_{11}(i,j) - \tilde{\boldsymbol{x}}_{11}(i,j) \\ \Delta y(i,j) &= y(i,j) - \tilde{y}(i,j). \end{aligned}$$

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

$$\boldsymbol{G}_D(z_1,z_2) = \boldsymbol{c}(\boldsymbol{Z}-\boldsymbol{A})^{-1}(\boldsymbol{A}-\boldsymbol{D}) + \boldsymbol{c}. \tag{4}$$

For the filter (3) with EF, the noise gain defined by $I(\boldsymbol{D}) = \sigma_{out}^2/\sigma^2$ can be evaluated as

$$\begin{aligned} I(\boldsymbol{D}) &= \frac{1}{(2\pi j)^2}\oint_{\Gamma_1}\oint_{\Gamma_2} \boldsymbol{G}_D(z_1,z_2)\boldsymbol{G}_D^*(z_1,z_2)\frac{dz_1dz_2}{z_1z_2} \\ &= \mathrm{tr}[\boldsymbol{W}_D] \end{aligned} \tag{5}$$

where σ_{out}^2 denotes noise variance at the output, and

$$\boldsymbol{W}_D = \frac{1}{(2\pi j)^2}\oint_{\Gamma_1}\oint_{\Gamma_2} \boldsymbol{G}_D^*(z_1,z_2)\boldsymbol{G}_D(z_1,z_2)\frac{dz_1dz_2}{z_1z_2}$$

with $\Gamma_i = \{z_i : |z_i| = 1\}$ for $i = 1,2$. Utilizing the *2-D Cauchy integral theorem*, the matrix $\boldsymbol{W}_D$ in (5) can be expressed in closed form

$$\boldsymbol{W}_D = (\boldsymbol{A}-\boldsymbol{D})^T\boldsymbol{W}_o(\boldsymbol{A}-\boldsymbol{D}) + \boldsymbol{c}^T\boldsymbol{c} \tag{6}$$

where $\boldsymbol{W}_o$ is called the local observability Gramian of the 2-D filter and defined by

$$\boldsymbol{W}_o = \frac{1}{(2\pi j)^2}\oint_{\Gamma_1}\oint_{\Gamma_2} (\boldsymbol{Z}^*-\boldsymbol{A}^T)^{-1}\boldsymbol{c}^T\boldsymbol{c}(\boldsymbol{Z}-\boldsymbol{A})^{-1}\frac{dz_1dz_2}{z_1z_2}. \tag{7}$$

Matrix $\boldsymbol{W}_o$ in (7) is referred to as the *unit noise matrix* for the 2-D filter (2) with $\boldsymbol{D} = \boldsymbol{0}$, and $\boldsymbol{W}_D$ is viewed as the *unit noise matrix* for the 2-D filter (2) with EF specified by matrix $\boldsymbol{D}$. In the case where there is no EF in the 2-D filter, the noise gain $I(\boldsymbol{D})$ with $\boldsymbol{D} = \boldsymbol{0}$ is expressed as

$$I(\boldsymbol{0}) = \mathrm{tr}[\boldsymbol{A}^T\boldsymbol{W}_o\boldsymbol{A} + \boldsymbol{c}^T\boldsymbol{c}] = \mathrm{tr}[\boldsymbol{W}_o]. \tag{8}$$

It is noted that the L_2-norm scaling constraints on the local state vector $\boldsymbol{x}(i,j)$ involves the local controllability Gramian $\boldsymbol{K}_c$ of the 2-D filter, defined by

$$\boldsymbol{K}_c = \frac{1}{(2\pi j)^2}\oint_{\Gamma_1}\oint_{\Gamma_2} (\boldsymbol{Z}-\boldsymbol{A})^{-1}\boldsymbol{b}\boldsymbol{b}^T(\boldsymbol{Z}^*-\boldsymbol{A}^T)^{-1}\frac{dz_1dz_2}{z_1z_2}. \tag{9}$$

III. JOINT ERROR FEEDBACK AND REALIZATION OPTIMIZATION

A. Probem Statement

Applying a coordinate transformation defined by $\overline{\boldsymbol{x}}(i,j) = \boldsymbol{T}^{-1}\boldsymbol{x}(i,j)$ with $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ transforms the filter $(\boldsymbol{A},\boldsymbol{b},\boldsymbol{c},d)_{m,n}$ to $(\overline{\boldsymbol{A}},\overline{\boldsymbol{b}},\overline{\boldsymbol{c}},d)_{m,n}$ 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 local controllability Gramian $\overline{\boldsymbol{K}}_c$ and local observability Gramian $\overline{\boldsymbol{W}}_o$ in the new realization then satisfy

$$\overline{\boldsymbol{K}}_c = \boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T}, \qquad \overline{\boldsymbol{W}}_o = \boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{T}. \tag{11}$$

If the L_2-norm dynamic-range scaling constraints

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

are imposed on the new realization, then it is known that [16],[17]

$$\min_{\boldsymbol{T}}\ \mathrm{tr}[\overline{\boldsymbol{W}}_o] = \frac{1}{m}\left(\sum_{i=1}^{m}\sigma_{1i}\right)^2 + \frac{1}{n}\left(\sum_{i=1}^{n}\sigma_{4i}\right)^2 \tag{13}$$

where σ_{1i}^2 for $i = 1,2,\cdots,m$ and σ_{4i}^2 for $i = 1,2,\cdots,n$ are the eigenvalues of matrices $\boldsymbol{K}_{1c}\boldsymbol{W}_{1o}$ and $\boldsymbol{K}_{4c}\boldsymbol{W}_{4o}$, respectively, and

$$\boldsymbol{K}_c = \begin{bmatrix} \boldsymbol{K}_{1c} & \boldsymbol{K}_{2c} \\ \boldsymbol{K}_{3c} & \boldsymbol{K}_{4c} \end{bmatrix}.$$

The LSS model $(\overline{\boldsymbol{A}},\overline{\boldsymbol{b}},\overline{\boldsymbol{c}},d)_{m,n}$ satisfying (12) and (13) simultaneously is known as the *optimal realization.*

If a coordinate transformation $\overline{\boldsymbol{x}}(i,j) = \boldsymbol{T}^{-1}\boldsymbol{x}(i,j)$ with $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ is applied to the LSS model (2), then the 2-D filter with EF can be characterized by

$$\begin{aligned} \tilde{\boldsymbol{x}}_{11}(i,j) &= \overline{\boldsymbol{A}}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,j)] + \overline{\boldsymbol{b}}u(i,j) + \boldsymbol{D}\boldsymbol{e}(i,j) \\ \tilde{y}(i,j) &= \overline{\boldsymbol{c}}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,j)] + du(i,j). \end{aligned} \tag{14}$$

This corresponds to (2) in the original realization. In this case, the noise gain $I(\boldsymbol{D},\boldsymbol{T})$ can be expressed as a function of matrices $\boldsymbol{D}$ and $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ in the form

$$I(\boldsymbol{D},\boldsymbol{T}) = \mathrm{tr}[\overline{\boldsymbol{W}}_D] \tag{15}$$

where

$$\overline{\boldsymbol{W}}_D = (\overline{\boldsymbol{A}} - \boldsymbol{D})^T \overline{\boldsymbol{W}}_o (\overline{\boldsymbol{A}} - \boldsymbol{D}) + \overline{\boldsymbol{c}}^T \overline{\boldsymbol{c}}.$$

The problem of RN minimization is to obtain matrices $\boldsymbol{D}$ and $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ which minimize (15) subject to the scaling constraints in (12).

B. Problem Relaxation and Conversion

In order to reduce solution sensitivity, the objective function in (15) is modified to

$$J(\boldsymbol{D}, \boldsymbol{T}) = \mathrm{tr}[(1-\mu)\overline{\boldsymbol{W}}_D + \mu \overline{\boldsymbol{W}}_o] \tag{16}$$

where $0 \le \mu \le 1$ is a scalar that weights the importance of reducing $\mathrm{tr}[\overline{\boldsymbol{W}}_o]$ relative to reducing $\mathrm{tr}[\overline{\boldsymbol{W}}_D]$. Defining

$$\hat{\boldsymbol{T}} = \hat{\boldsymbol{T}}_1 \oplus \hat{\boldsymbol{T}}_4 = (\boldsymbol{T}_1 \oplus \boldsymbol{T}_4)^T (\boldsymbol{K}_{1c} \oplus \boldsymbol{K}_{4c})^{-\frac{1}{2}} \tag{17}$$

it follows that

$$\overline{\boldsymbol{K}}_c = \hat{\boldsymbol{T}}^{-T} \begin{bmatrix} \boldsymbol{I}_m & \boldsymbol{K}_{1c}^{-\frac{1}{2}} \boldsymbol{K}_{2c} \boldsymbol{K}_{4c}^{-\frac{1}{2}} \\ \boldsymbol{K}_{4c}^{-\frac{1}{2}} \boldsymbol{K}_{3c} \boldsymbol{K}_{1c}^{-\frac{1}{2}} & \boldsymbol{I}_n \end{bmatrix} \hat{\boldsymbol{T}}^{-1}. \tag{18}$$

This reduces the scaling constraints in (12) to

$$\begin{aligned} (\hat{\boldsymbol{T}}_1^{-T} \hat{\boldsymbol{T}}_1^{-1})_{ii} = 1, \qquad & i = 1, 2, \cdots, m \\ (\hat{\boldsymbol{T}}_4^{-T} \hat{\boldsymbol{T}}_4^{-1})_{kk} = 1, \qquad & k = 1, 2, \cdots, n. \end{aligned} \tag{19}$$

The constraints in (19) simply state that each column in matrices $\hat{\boldsymbol{T}}_1^{-1}$ and $\hat{\boldsymbol{T}}_4^{-1}$ must be a unity vector. These are satisfied if $\hat{\boldsymbol{T}}_1^{-1}$ and $\hat{\boldsymbol{T}}_4^{-1}$ assume the forms

$$\begin{aligned} \hat{\boldsymbol{T}}_1^{-1} &= \left[\frac{\boldsymbol{t}_{11}}{||\boldsymbol{t}_{11}||}, \frac{\boldsymbol{t}_{12}}{||\boldsymbol{t}_{12}||}, \cdots, \frac{\boldsymbol{t}_{1m}}{||\boldsymbol{t}_{1m}||} \right] \\ \hat{\boldsymbol{T}}_4^{-1} &= \left[\frac{\boldsymbol{t}_{41}}{||\boldsymbol{t}_{41}||}, \frac{\boldsymbol{t}_{42}}{||\boldsymbol{t}_{42}||}, \cdots, \frac{\boldsymbol{t}_{4n}}{||\boldsymbol{t}_{4n}||} \right] \end{aligned} \tag{20}$$

where $\boldsymbol{t}_{1i}$ for $i = 1, 2, \cdots, m$ and $\boldsymbol{t}_{4j}$ for $j = 1, 2, \cdots, n$ are $m \times 1$ and $n \times 1$ real vectors, respectively. In such a case, matrix $\overline{\boldsymbol{W}}_D$ in (15) can be written as

$$\overline{\boldsymbol{W}}_D = \hat{\boldsymbol{T}} [(\hat{\boldsymbol{A}} - \hat{\boldsymbol{T}}^T \boldsymbol{D} \hat{\boldsymbol{T}}^{-T})^T \hat{\boldsymbol{W}}_o (\hat{\boldsymbol{A}} - \hat{\boldsymbol{T}}^T \boldsymbol{D} \hat{\boldsymbol{T}}^{-T}) + \hat{\boldsymbol{C}}] \hat{\boldsymbol{T}}^T \tag{21}$$

where $\hat{\boldsymbol{T}} = \hat{\boldsymbol{T}}_1 \oplus \hat{\boldsymbol{T}}_4$ and

$$\begin{aligned} \hat{\boldsymbol{A}} &= (\boldsymbol{K}_{1c} \oplus \boldsymbol{K}_{4c})^{-\frac{1}{2}} \boldsymbol{A} (\boldsymbol{K}_{1c} \oplus \boldsymbol{K}_{4c})^{\frac{1}{2}} \\ \hat{\boldsymbol{C}} &= (\boldsymbol{K}_{1c} \oplus \boldsymbol{K}_{4c})^{\frac{1}{2}} \boldsymbol{c}^T \boldsymbol{c} (\boldsymbol{K}_{1c} \oplus \boldsymbol{K}_{4c})^{\frac{1}{2}} \\ \hat{\boldsymbol{W}}_o &= (\boldsymbol{K}_{1c} \oplus \boldsymbol{K}_{4c})^{\frac{1}{2}} \boldsymbol{W}_o (\boldsymbol{K}_{1c} \oplus \boldsymbol{K}_{4c})^{\frac{1}{2}}. \end{aligned}$$

Moreover, the objective function in (16) becomes

$$\begin{aligned} J(\boldsymbol{D}, \hat{\boldsymbol{T}}) = {} & (1-\mu)\, \mathrm{tr}[\hat{\boldsymbol{T}} (\hat{\boldsymbol{A}} - \hat{\boldsymbol{T}}^T \boldsymbol{D} \hat{\boldsymbol{T}}^{-T})^T \\ & \cdot \boldsymbol{W}_o (\hat{\boldsymbol{A}} - \hat{\boldsymbol{T}}^T \boldsymbol{D} \hat{\boldsymbol{T}}^{-T}) \hat{\boldsymbol{T}}^T] \\ & + (1-\mu)\, \mathrm{tr}[\hat{\boldsymbol{T}} \hat{\boldsymbol{C}} \hat{\boldsymbol{T}}^T] + \mu\, \mathrm{tr}[\hat{\boldsymbol{T}} \hat{\boldsymbol{W}}_o \hat{\boldsymbol{T}}^T]. \end{aligned} \tag{22}$$

Therefore, the problem of obtaining matrices $\boldsymbol{D}$ and $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ that minimize (16) subject to the scaling constraints in (12) can be converted into an unconstrained optimization problem of obtaining matrices $\boldsymbol{D}$ and $\hat{\boldsymbol{T}} = \hat{\boldsymbol{T}}_1 \oplus \hat{\boldsymbol{T}}_4$ that minimize (22).

C. Optimization Method

Let $\boldsymbol{x}$ be the column vector that collects the variables in matrices $\boldsymbol{D}$ and $\hat{\boldsymbol{T}} = \hat{\boldsymbol{T}}_1 \oplus \hat{\boldsymbol{T}}_4$. Then, $J(\boldsymbol{D}, \hat{\boldsymbol{T}})$ is a function of $\boldsymbol{x}$, denoted by $J(\boldsymbol{x})$. The algorithm starts with a trivial initial point $\boldsymbol{x}_0$ obtained from an initial assignment $\boldsymbol{D} = \hat{\boldsymbol{T}} = \boldsymbol{I}_{m+n}$. In the kth iteration, a quasi-Newton algorithm updates the most recent point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k \tag{23}$$

where [9]

$$\boldsymbol{d}_k = -\boldsymbol{S}_k \nabla J(\boldsymbol{x}_k), \qquad \alpha_k = arg \min_{\alpha} J(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$$

$$\boldsymbol{S}_{k+1} = \boldsymbol{S}_k + \left(1 + \frac{\boldsymbol{\gamma}_k^T \boldsymbol{S}_k \boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_k^T \boldsymbol{\delta}_k}\right) \frac{\boldsymbol{\delta}_k \boldsymbol{\delta}_k^T}{\boldsymbol{\gamma}_k^T \boldsymbol{\delta}_k} - \frac{\boldsymbol{\delta}_k \boldsymbol{\gamma}_k^T \boldsymbol{S}_k + \boldsymbol{S}_k \boldsymbol{\gamma}_k \boldsymbol{\delta}_k^T}{\boldsymbol{\gamma}_k^T \boldsymbol{\delta}_k}$$

$$\boldsymbol{S}_0 = \boldsymbol{I},\ \boldsymbol{\delta}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k,\ \boldsymbol{\gamma}_k = \nabla J(\boldsymbol{x}_{k+1}) - \nabla J(\boldsymbol{x}_k).$$

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

$$|J(\boldsymbol{x}_{k+1}) - J(\boldsymbol{x}_k)| < \varepsilon \tag{24}$$

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

Case 1: $\boldsymbol{D}$ is a general matrix

From (22), the optimal choice of $\boldsymbol{D}$ is given by

$$\boldsymbol{D} = \hat{\boldsymbol{T}}^{-T} \hat{\boldsymbol{A}} \hat{\boldsymbol{T}}^T \tag{25}$$

which leads to

$$J(\hat{\boldsymbol{T}}^{-T} \hat{\boldsymbol{A}} \hat{\boldsymbol{T}}^T, \hat{\boldsymbol{T}}) = \mathrm{tr}[\hat{\boldsymbol{T}} \{(1-\mu)\hat{\boldsymbol{C}} + \mu \hat{\boldsymbol{W}}_o\} \hat{\boldsymbol{T}}^T]. \tag{26}$$

Then, the elements in vector $\boldsymbol{x}$ consist of $\hat{\boldsymbol{T}} = \hat{\boldsymbol{T}}_1 \oplus \hat{\boldsymbol{T}}_4$ and the gradients of $J(\boldsymbol{x})$ are found to be

$$\begin{aligned} \frac{\partial J(\boldsymbol{x})}{\partial t_{ij}} &= \lim_{\Delta \to \infty} \frac{J(\hat{\boldsymbol{T}}_{ij}) - J(\hat{\boldsymbol{T}})}{\Delta} \\ &= 2\boldsymbol{e}_j^T \hat{\boldsymbol{T}} [(1-\mu)\hat{\boldsymbol{C}} + \mu \hat{\boldsymbol{W}}_o] \hat{\boldsymbol{T}}^T \hat{\boldsymbol{T}} \boldsymbol{g}_{ij} \end{aligned} \tag{27}$$

$$(1 \le i, j \le m) \ \text{ or } \ (m+1 \le i, j \le m+n).$$

where $\hat{\boldsymbol{T}}_{ij}$ is the matrix obtained from $\hat{\boldsymbol{T}}$ with a perturbed (i, j)th component, and is given by [10]

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

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

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

Case 2: $\boldsymbol{D}$ is a block-diagonal matrix

$$\boldsymbol{D} = \boldsymbol{D}_1 \oplus \boldsymbol{D}_4 \tag{28}$$

where $\boldsymbol{D}_1$ and $\boldsymbol{D}_4$ are $m \times m$ and $n \times n$ matrices, respectively. The gradients of $J(\boldsymbol{x})$ can be derived as

$$\begin{aligned} \frac{\partial J(\boldsymbol{x})}{\partial t_{ij}} &= 2\beta_1 + (1-\mu)(\beta_2 - \beta_3) \\ \frac{\partial J(\boldsymbol{x})}{\partial d_{ij}} &= 2\boldsymbol{e}_j^T (1-\mu)\hat{\boldsymbol{T}}\hat{\boldsymbol{W}}_o(\hat{\boldsymbol{T}}^T \boldsymbol{D} - \hat{\boldsymbol{A}}\hat{\boldsymbol{T}}^T)\boldsymbol{e}_i \end{aligned} \tag{29}$$

where

$$\begin{aligned} \beta_1 &= \boldsymbol{e}_j^T \hat{\boldsymbol{T}} [\, (1-\mu)(\hat{\boldsymbol{A}}^T \hat{\boldsymbol{W}}_o \hat{\boldsymbol{A}} + \hat{\boldsymbol{C}}\,) + \mu \hat{\boldsymbol{W}}_o] \hat{\boldsymbol{T}}^T \hat{\boldsymbol{T}}\, \boldsymbol{g}_{ij} \\ \beta_2 &= \boldsymbol{e}_j^T \hat{\boldsymbol{T}}\hat{\boldsymbol{W}}_o \hat{\boldsymbol{T}}^T \hat{\boldsymbol{T}} \boldsymbol{D}\boldsymbol{D}^T \hat{\boldsymbol{T}} \boldsymbol{g}_{ij} \\ \beta_3 &= \boldsymbol{e}_j^T \hat{\boldsymbol{T}} (\hat{\boldsymbol{A}}^T \hat{\boldsymbol{W}}_o \hat{\boldsymbol{T}}^T \boldsymbol{D} + \hat{\boldsymbol{W}}_o \hat{\boldsymbol{A}} \hat{\boldsymbol{T}}^T \boldsymbol{D}^T)\, \boldsymbol{g}_{ij} \end{aligned}$$

with $\boldsymbol{g}_{ij}$ defined in (27). Here, $d_{ij} \in \boldsymbol{D}_1 \oplus \boldsymbol{D}_4$ such that $d_{ij} \in \boldsymbol{D}_1$ for $(1,1) \leq (i,j) \leq (m,m)$ and $d_{ij} \in \boldsymbol{D}_4$ for $(m+1, m+1) \leq (i,j) \leq (m+n, m+n)$.

Case 3: $\boldsymbol{D}$ is a diagonal matrix

$$\boldsymbol{D} = \mathrm{diag}\{d_{11}, d_{22}, \cdots, d_{m+n,m+n}\} \tag{30}$$

which leads to

$$\frac{\partial J(\boldsymbol{x})}{\partial d_{ii}} = 2\boldsymbol{e}_i^T (1-\mu)\hat{\boldsymbol{T}}\hat{\boldsymbol{W}}_o(\hat{\boldsymbol{T}}^T \boldsymbol{D} - \hat{\boldsymbol{A}}\hat{\boldsymbol{T}}^T)\boldsymbol{e}_i \tag{31}$$

where $1 \leq i \leq m+n$. In this case, $\partial J(\boldsymbol{x})/\partial t_{ij}$ is the same as in (29).

Case 4: $\boldsymbol{D}_1 = \alpha \boldsymbol{I}_m$ and $\boldsymbol{D}_4 = \beta \boldsymbol{I}_n$ with scalars α, β

The gradients of $J(\boldsymbol{x})$ can be calculated using

$$\begin{aligned} \frac{\partial J(\boldsymbol{x})}{\partial \alpha} &= 2\boldsymbol{e}_1^T (1-\mu)\hat{\boldsymbol{T}}\hat{\boldsymbol{W}}_o(\hat{\boldsymbol{T}}^T \boldsymbol{D} - \hat{\boldsymbol{A}}\hat{\boldsymbol{T}}^T)\boldsymbol{e}_1 \\ \frac{\partial J(\boldsymbol{x})}{\partial \beta} &= 2\boldsymbol{e}_{m+1}^T (1-\mu)\hat{\boldsymbol{T}}\hat{\boldsymbol{W}}_o(\hat{\boldsymbol{T}}^T \boldsymbol{D} - \hat{\boldsymbol{A}}\hat{\boldsymbol{T}}^T)\boldsymbol{e}_{m+1} \end{aligned} \tag{32}$$

and $\partial J(\boldsymbol{x})/\partial t_{ij}$ is computed using (29).

IV. CONCLUSION

The joint optimization of a error feedback matrix and a coordinate-transformation matrix in 2-D state-space digital filters for roundoff noise minimization subject to L_2-norm dynamic-range scaling constraints has been investigated. It has been shown that the problem at hand can be converted into an unconstrained optimization problem by using linear algebraic techniques. An efficient quasi-Newton algorithm has been employed to solve the unconstrained optimization problem iteratively. It has been clarified that the proposed technique can be applied to the cases where the error feedback matrix is a scalar, diagonal, block-diagonal, or general matrix. Our computer simulation results have demonstrated the effectiveness of the proposed technique compared with the existing method.

1. REFERENCES

[1] T. Hinamoto, S. Karino and N. Kuroda, "Error spectrum shaping in 2-D digital filters," *Proc. IEEE Int. Symp. Circuits Syst.* (ISCAS'95), vol. 1, pp. 348-351, May 1995.

[2] P. Agathoklis and C. Xiao, "Low roundoff noise structures for 2-D filters," *Proc. IEEE Int. Symp. Circuits Syst.* (ISCAS'96), vol. 2, pp. 352-355, May 1996.

[3] T. Hinamoto, S. Karino and N. Kuroda, "2-D state-space digital filters with error spectrum shaping," *Proc. IEEE Int. Symp. Circuits Syst.* (ISCAS'96), vol. 2, pp. 766-769, May 1996.

[4] T. Hinamoto, N. Kuroda and T. Kuma, "Error feedback for noise reduction in 2-D digital filers with quadrantally symmetric or antisymmetric coefficients," *Proc. IEEE Int. Symp. Circuits Syst.* (ISCAS'97), vol. 4, pp. 2461-2464, June 1997.

[5] T. Hinamoto, S. Karino, N. Kuroda and T. Kuma, "Error spectrum shaping in two-dimensional recursive digital filters," *IEEE Trans. Circuits Syst.*, vol. 46, pp. 1203-1215, Oct. 1999.

[6] M. Kawamata and T. Higuchi, "A unified study on the roundoff noise in 2-D state-space digital filters," *IEEE Trans. Circuits Syst.*, vol. 33, pp. 724-730, July 1986.

[7] W.-S. Lu and A. Antoniou, "Synthesis of 2-D state-space fixed-point digital filter structures with minimum roundoff noise," *IEEE Trans. Circuits Syst.*, vol. 33, pp. 965-973, Oct. 1986.

[8] T. Hinamoto, K. Higashi and W.-S. Lu, "Separate/joint optimization of error feedback and coordinate transformation for roundoff noise minimization in two-dimensional state-space digital filters," *IEEE Trans. Signal Processing*, vol. 51, 2003 (to appear).

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

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