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].

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. A 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.

II. PROBLEM STATEMENT

Consider a 2-D stable, separately locally controllable and separately locally observable state-space digital filter $(A, b, c, d)$ described by

$$x(i, j) = Ax(i, j) + bu(i, j)$$

(1)

where $x(i, j)$ is an $m \times 1$ horizontal state vector, $x'(i, j)$ is an $n \times 1$ vertical state vector, $u(i, j)$ is a scalar input, and $A$, $A_1$, $A_2$, $A_1$, $b$, $c_1$, $c_2$ 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

$$\bar{x}(i, j) = AQ[\bar{x}(i, j)] + bu(i, j)$$

$$+ \sum_{k=0}^{M-1} D_{1k} \otimes 0 e^{-k(i, j)} + \sum_{i=0}^{N-1} [0 \otimes D_{u}] e^{-i(i, j)}$$

$$\bar{y}(i, j) = cAQ[\bar{x}(i, j)] + du(i, j) + he_{0}(i, j)$$

(2)

where $h$ is a $1 \times (m + n)$ error-feedforward vector, $D_{1k}$’s and $D_{u}$’s are $m \times m$ and $n \times n$ high-order EF matrices, respectively, and

$$e^{-k(i, j)} = [e^{h(1)}(i, j), e^{h(1)}(i, j - k)]$$

$$e_{k}(i, j) = \bar{x}(i, j) - Q[\bar{x}(i, j)]$$

The coefficient matrices $A$, $b$, $c$ and $d$ in (2) are assumed to have exact fractional $B_{r}$-bit representations. The FWL horizontal (vertical) state vector $\bar{x}^{h}(i, j)$ ($\bar{x}'(i, j)$) and each output $\bar{y}(i, j)$ has $B_{r}$-bit fractional representations, while the input $u(i, j)$ is a $(B - B_{r})$-bit fraction. The quantizer $Q_{i}$ in (2) rounds the $B_{r}$-bit fraction $\bar{x}(i, j)$ to $(B - B_{r})$-bit after the multiplications and additions, where the sign bit is not counted. Assume that the roundoff error $e_{0}(i, j)$ can be modeled as
a Gaussian random process with zero mean and covariance \( \sigma^2 I_n \). By subtracting (2) from (1), it follows that
\[
\Delta x_1(i,j) = A \Delta x_0(i,j) + A e_0(i,j)
\]
\[
- \sum_{k=0}^{M-1} [D_{1k} \oplus \emptyset] e_{-k}(i,j) - \sum_{l=0}^{N-1} [\emptyset \oplus D_{4l}] e_{-l}(i,j)
\]
\[
\Delta y(i,j) = c \Delta x_0(i,j) + (c-h) e_0(i,j)
\]
where \( \Delta x_k(i,j) = x_k(i,j) - \hat{x}_k(i,j) \) for \( k = 0, 1 \) and \( \Delta y(i,j) = y(i,j) - \hat{y}(i,j) \). By taking \((z_1, z_2)\)-transform on both sides of (3) and setting the boundary conditions to be zero, we obtain
\[
\Delta Y(z_1, z_2) = [H^h_e(z_1), H^v_e(z_1, z_2')] E_0(z_1, z_2)
\]
(4a)
where \( \Delta Y(z_1, z_2) \) and \( E_0(z_1, z_2) \) are \((z_1, z_2)\)-transforms of \( \Delta y(i,j) \) and \( e_0(i,j) \), respectively, and
\[
H^h_e(z_1) = c_1 \sum_{i=0}^{\infty} \left( A_i^1 - \sum_{k=0}^{M-1} A_i^{-k-1} D_{1k} \right) z_1^{-i} - h_1
\]
\[
H^v_e(z_1, z_2) = [c_2 + c_1 (z_1 I_m - A_1^{-1}) A_2] \sum_{i=0}^{\infty} \left( A_i^1 - \sum_{k=0}^{M-1} A_i^{-k-1} D_{1k} \right) z_2^{-i} - h_2
\]
with \( A_1^1 = 0 \) and \( A_1^1 = 0 \) for \( i < 0 \).

Based on above analysis, we can now define the normalized noise gain of the 2-D filter as
\[
J^h_0(h_1, D_1, D_4) = J^h_0(h_1, D_1) + J^v_0(h_2, D_4) = \sigma^2_{out}/\sigma^2
\]
(5a)
where
\[
J^h_0(h_1, D_1) = \text{tr} \left[ \frac{1}{2\pi j} \oint_{|z_1|=1} [H^h_e(z_1)]^* H^h_e(z_1) \frac{dz_1}{z_1} \right]
\]
\[
J^v_0(h_2, D_4) = \text{tr} \left[ \frac{1}{2\pi j} \oint_{|z_1|=1} \oint_{|z_2|=1} [H^v_e(z_1, z_2)]^* H^v_e(z_1, z_2) \frac{dz_1 dz_2}{z_1 z_2} \right]
\]
\[
D_1 = [D_{10}, D_{11}, \ldots, D_{1,M-1}]
\]
\[
D_4 = [D_{40}, D_{41}, \ldots, D_{4,N-1}]
\]
and \( \sigma^2_{out} \) is noise variance at the filter output. By substituting (4b) into (5b), we have
\[
J^h_0(h_1, D_1) = \text{tr} \left[ W^h - \sum_{k=0}^{M-1} \left\{ D_{1k}^T W^h A_i^{k+1} + (A_i^1)^{k+1} W^h D_{1k} \right\} \right.
\]
\[
+ \sum_{k=0}^{M-1} \sum_{l=0}^{M-1} D_{1k}^T \left\{ W^h A_i^{-l} + (A_i^1)^{-l} W^h \right\} D_{1l} \Bigg]
\]
\[
- \sum_{k=0}^{M-1} D_{1k}^T W^h D_{1k} - 2h_1^T c_1 + h_1^T h_1 \right]
\]
where \( W^h \) is the horizontal observability Gramian of the filter in (1) that can be obtained by solving the Lyapunov equation
\[
W^h = A_i^1 W^h A_i + c_1^T c_1
\]
and
\[
J^v_0(h_2, D_4) = \text{tr} \left[ W^v - \sum_{k=0}^{N-1} \left\{ D_{4k}^T W^v A_i^{k+1} + (A_i^1)^{k+1} W^v D_{4k} \right\} \right.
\]
\[
+ \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} D_{4k}^T \left\{ W^v A_i^{-l} + (A_i^1)^{-l} W^v \right\} D_{4l} \Bigg]
\]
\[
- \sum_{k=0}^{N-1} D_{4k}^T W^v D_{4k} - 2h_2^T c_2 + h_2^T h_2 \right]
\]
where \( W^v \) is the vertical observability Gramian of the filter in (1) that can be obtained by solving the Lyapunov equation
\[
W^v = A_i^1 W^v A_i + A_i^2 W^v A_2 + c_2^T c_2.
\]
If the high-order EF matrices \( D_{1k} \) for \( k = 0, 1, \ldots, M-1 \) and \( D_{4l} \) for \( l = 0, 1, \ldots, N-1 \) are diagonal then (6) and (7) can be considerably simplified to
\[
J^h(h_1, D_1) = \text{tr} \left[ W^h - c_1^T c_1 - 2 \sum_{k=0}^{M-1} W^h A_i^{k+1} D_{1k} \right.
\]
\[
+ \sum_{k=0}^{M-1} \sum_{l=0}^{M-1} W^h A_i^{-l} D_{1k} D_{1l} \Bigg] + (c_1 - h_1)(c_1 - h_1)^T
\]
(8)
and
\[
J^v(h_2, D_4) = \text{tr} \left[ W^v - c_2^T c_2 - 2 \sum_{k=0}^{N-1} W^v A_i^{k+1} D_{4k} \right.
\]
\[
+ \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} W^v A_i^{-l} D_{4k} D_{4l} \Bigg] + (c_2 - h_2)(c_2 - h_2)^T.
\]
(9)
respectively.

It should be noted that \( l_2\)-scaling constraints on the local state vector \( x_0(i,j) \) involve the local controllability Grammian \( K = K^h \oplus K^v \) of the filter in (1), which can be obtained by solving the Lyapunov equations [3]
\[
K^h = A_i K^h A_i^T + b_i b_i^T
\]
\[
K^v = A_i K^v A_i^T + A_i K^v A_i^T + b_i b_i^T.
\]
(10)
A different yet equivalent local state-space model of (1), \( \langle \mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D} \rangle_n \), can be obtained via a coordinate transformation \( x_0(i,j) = T^{-1} x_0(i,j) \) with \( T = T_1 \oplus T_4 \) where
\[
\mathcal{A} = T^{-1} \mathcal{A}, \quad \mathcal{B} = T^{-1} \mathcal{B}, \quad \mathcal{C} = e^T. \quad (11)
\]
The \( l_2\)-scaling constraints are imposed on the local state vector \( x_0(i,j) \) so that
\[
\langle \mathcal{K} \rangle_{ij} = (T^{-1} \mathcal{K} T^{-1} x_0)_{ij} = 1 \quad \text{for} \quad i = 1, 2, \ldots, m+n \quad (12)
\]
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 \( T = T_1 \oplus T_4 \) as well as diagonal high-order EF matrices \( D_{1k} \) for \( k = 0, 1, \ldots, M-1 \) and \( D_{4l} \) for \( l = 0, 1, \ldots, N-1 \) that jointly minimize
\[
J(T, D_1, D_4) = J^h(T_1, D_1) + J^v(T_4, D_4)
\]
(13a)
subject to the $l_2$-scaling constraints in (12) where

$$
\mathcal{J}'(T_1, D_1) = \left[ T_1^T (W^h - c_1^T c_1) T_1 - 2 \sum_{k=0}^{M-1} T_1^T W^h A_i^{k+1} T_1 D_{1k} + \sum_{k=0}^{M-1} \sum_{i=0}^{N-1} T_1^T W^h A_i^{k-i} T_1 D_{1k} D_{1l} \right]
$$

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

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

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

$$\hat{T} = \hat{T}_1 \oplus \hat{T}_4 = (T_1 \oplus T_4)^T (K^h \oplus K^v)^{-\frac{1}{2}}$$

which leads to (12) to

$$
(T^T T^{-1})_{ii} = 1 \text{ for } i = 1, 2, \ldots, m + n.
$$

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

$$
\hat{T}_1^{-1} = \left[ \frac{t_{11}}{||t_{11}||}, \frac{t_{12}}{||t_{12}||}, \ldots, \frac{t_{1m}}{||t_{1m}||} \right]
$$

$$
\hat{T}_4^{-1} = \left[ \frac{t_{41}}{||t_{41}||}, \frac{t_{42}}{||t_{42}||}, \ldots, \frac{t_{4n}}{||t_{4n}||} \right].
$$

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

$$
\hat{J}(\hat{T}, D_1, D_4) = \left[ U_0 - c_1^T \hat{c}_1 - 2 \sum_{k=0}^{M-1} U_{k+1} D_{1k} + \sum_{k=0}^{M-1} \sum_{i=0}^{N-1} U_{[k-i]} D_{1k} D_{1l} \right]
$$

$$
+ \left[ V_0 - c_2^T \hat{c}_2 - 2 \sum_{k=0}^{N-1} V_{k+1} D_{4k} + \sum_{k=0}^{N-1} \sum_{i=0}^{N-1} V_{[k-i]} D_{4k} D_{4l} \right],
$$

where

$$U_k = T_1 (K^h)^{\frac{1}{2}} W^h A_i^{k} (K^h)^{\frac{1}{2}} T_1^T, \quad c_1 = c_1 (K^h)^{\frac{1}{2}} T_1^T,
$$

$$V_k = T_4 (K^v)^{\frac{1}{2}} W^v A_i^{k} (K^v)^{\frac{1}{2}} T_4^T, \quad c_2 = c_2 (K^v)^{\frac{1}{2}} T_4^T.
$$

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

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

The proposed algorithm starts with the initial point $x_0$ obtained from the assignment $T_1 = D_{1k} = I_n$ and $T_1 = D_{4l} = I_n$ for any $k$ and $l$. In the $k$th iteration, a quasi-Newton algorithm updates the most recent point $x_k$ to point $x_{k+1}$ as

$$
x_{k+1} = x_k + \alpha_k d_k,
$$

where

$$
d_k = -S_k \nabla J(x_k), \quad \alpha_k = \arg \min J(x_k + \alpha d_k)
$$

and $S_k = S_k + (1 + \frac{\gamma^T_k S_k \gamma_k}{\delta_k \delta_k}) \frac{\delta \gamma}{\delta_k \delta_k} \frac{\delta \gamma k}{\delta_k \delta_k} \frac{\delta \gamma k}{\delta_k \delta_k}$

$$
S_0 = I, \quad \delta k = x_{k+1} - x_k, \quad \gamma_k = -\nabla J(x_k + 1) - \nabla J(x_k).
$$

Here, $\nabla J(x)$ is the gradient of $J(x)$ with respect to $x$, and $S_k$ is a positive-definite approximation of the inverse Hessian matrix of $J(x_k)$. This iteration process continues until $|J(x_{k+1}) - J(x_k)| < \varepsilon$.

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

$$D_{1k} = \text{diag}\{\alpha_{k1}, \alpha_{k2}, \ldots, \alpha_{kn}\}, \quad k = 0, 1, \ldots, M - 1
$$

$$D_{4l} = \text{diag}\{\beta_{l1}, \beta_{l2}, \ldots, \beta_{ln}\}, \quad l = 0, 1, \ldots, N - 1.
$$

(20)

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

$$
\frac{\partial J(x)}{\partial t_{ij}} = 2e_i^T \left[ U_0 - c_1^T \hat{c}_1 - 2 \sum_{k=0}^{M-1} (U_{k+1} + U_{k+1}) D_{1k} \right]
$$

$$+ \frac{1}{2} \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} (U_{[k-l]} + U_{[k-l]}) D_{1k} D_{1l}$$

$$\frac{\partial J(x)}{\partial t_{4j}} = 2e_{i(j)}^T \left[ V_0 - c_2^T \hat{c}_2 - 2 \sum_{k=0}^{N-1} (V_{k+1} + V_{k+1}) D_{4k} \right]
$$

$$+ \frac{1}{2} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} (V_{[k-l]} + V_{[k-l]}) D_{4k} D_{4l}$$

(21)

where $t_{ij}$ is the $i$th element of vector $t_{1j}$ ($t_{4j}$) and $g_{i(j)}$ is the $i$th element of vector $g_{i1j}$ ($g_{i4j}$) and $e_{i1j}$ ($e_{i4j}$) ($i, j \leq m$).

(22)
IV. AN ILLUSTRATIVE EXAMPLE

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

\[
\begin{align*}
A^T_1 &= \begin{bmatrix} 0 & 0 & 0.590655 \\ 1 & 0 & -1.836929 \\ 0 & 1 & 2.173645 \end{bmatrix}, & b_1 &= \begin{bmatrix} 0.047053 \\ 0.062274 \\ 0.060436 \end{bmatrix} \\
A_4 &= \begin{bmatrix} 0 & 0 & 0.564961 \\ 1 & 0 & -1.887939 \\ 0 & 1 & 2.280029 \end{bmatrix}, & b_2 &= \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \\
A_2 &= \begin{bmatrix} 0.064564 & 0.033034 & 0.012881 \\ 0.091213 & 0.110512 & 0.102759 \\ 0.097256 & 0.151864 & 0.172460 \end{bmatrix}, & c_1^T &= \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \\
c_2 &= \begin{bmatrix} 0.016556 \\ 0.012550 \\ 0.008243 \end{bmatrix}, & d &= 0.019421.
\end{align*}
\]

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

\[\dot{J}(\hat{T}, 0, 0) = \text{tr}[U_0] + \text{tr}[V_0].\]

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

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

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

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

The complete simulation results regarding the noise gain \(J(x) + (\bar{c} - h)(\bar{c} - h)^T\) are summarized in Table I where the term “Infinite Precision” refers to the value of \(J(x)\) derived from the optimal \(\hat{T}, D_1\) and \(D_4\), the term “3-Bit Quantization” means that of \((\bar{c} - h)(\bar{c} - h)^T\) where only each entry of the optimal \(D_1, D_4\) and \(h\) was rounded to a power-of-two representation with 3 bits after the binary point, and the term “Integer Quantization” means that of \((\bar{c} - h)(\bar{c} - h)^T\) where only each entry of the optimal \(D_1, D_4\) and \(h\) was rounded to an integer.

<table>
<thead>
<tr>
<th>Table I. Performance Comparison</th>
<th>Infinite Precision</th>
<th>3-Bit Quantization</th>
<th>Integer Quantization</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M, N)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1, 1)</td>
<td>1.300436</td>
<td>1.351817</td>
<td>3.083187</td>
</tr>
<tr>
<td>(2, 1)</td>
<td>1.021734</td>
<td>1.067178</td>
<td>2.267652</td>
</tr>
<tr>
<td>(1, 2)</td>
<td>1.089940</td>
<td>1.134867</td>
<td>3.024715</td>
</tr>
<tr>
<td>(2, 2)</td>
<td>0.811238</td>
<td>0.850231</td>
<td>2.209052</td>
</tr>
</tbody>
</table>

V. CONCLUSION

An efficient technique for jointly optimizing high-order EF and realization to minimize the effects of roundoff noise subject to \(\ell_2\)-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