Roundoff Noise Minimization in Closed-Loop Systems Using Optimal Error Feedback and Realization

Takao Hinamoto[1†], Akimitsu Doi[1] and Wu-Sheng Lu[2]

[1]Hiroshima Institute of Technology, Hiroshima, Japan
(E-mail: hinamoto@ieee.org, doi@cc.it-hiroshima.ac.jp)
[2]University of Victoria, Victoria, Canada
(E-mail: wslu@ece.uvic.ca)

Abstract: We investigate the problem of optimizing error-feedback and realization to minimize the roundoff noise subject to l_2-scaling constraints in a system in which full-order state observer feedback is used in place of system state feedback. In this paper, the optimal error-feedback matrix is determined in closed-form in advance. As a result, joint optimization of error-feedback and realization is no longer required. An iterative optimization technique for obtaining the optimal coordinate transformation matrix that minimizes the roundoff noise subject to l_2-scaling constraints is explored by emplying a quasi-Newton algorithm. We also describe a separate optimization technique for comparison purposes. The results of numerical experiments demonstrate the validity and effectiveness of the proposed technique.

Keywords: l_2-scaling, Error feedback, Quasi-Newton method, Realization, Roundoff noise minimization, State observer

1. INTRODUCTION

In fixed-point recursive digital filters, roundoff noise arises due to finite-precision nature of computer arithmetic, and roundoff noise in turn degrades the filters' performance. Error feedback (EF) is well known as an effective method for reducing finite-word-length (FWL) effects in recursive digital filters, and many EF methods have been proposed in the past [1]-[5]. In addition, it is known that the roundoff noise significantly depends on the internal description of the recursive filter, and techniques for realizing a state-space model structure that minimizes the roundoff noise at the filter output subject to l_2-scaling constraints have been explored by appropriately constructing a linear transformation for state-space coordinates [6],[7]. As a natural extension of the aforementioned methods, efforts have been made to develop new methods that combine EF and realization for achieving better performance [8],[9]. Another relevant development was initiated in [10] where roundoff noise in a closed-loop system equipped with a full-order state observer feedback system is analyzed and an algorithm for realizing a full-order state observer feedback system which minimizes the roundoff noise subject to l_2-scaling constraints is proposed. It is noted that no EF is incorporated into the algorithm in [10].

This paper investigates the problem of optimizing EF and realization for roundoff noise minimization subject to l_2-scaling constraints in a system in which full-order state observer feedback is used instead of system state feedback. An iterative optimization technique for obtaining the optimal solution that minimizes the roundoff noise subject to l_2-scaling constraints is explored by utilizing a quasi-Newton algorithm. Numerical experiments are included to demonstrate the validity and effectiveness of the proposed technique. It is stated that an alternative analytical optimization technique will appear this year [11].

† Takao Hinamoto is the presenter of this paper.

2. NOISE GAIN AS FUNCTION OF ERROR FEEDBACK MATRIX

Suppose that a stable, controllable and observable linear discrete-time system $(\boldsymbol{F}_o, \boldsymbol{g}_o, \boldsymbol{h}_o)_n$ is described by

$$\begin{aligned} \boldsymbol{x}(k+1) &= \boldsymbol{F}_o\boldsymbol{x}(k) + \boldsymbol{g}_o u(k) \\ y(k) &= \boldsymbol{h}_o\boldsymbol{x}(k) \end{aligned} \tag{1}$$

where $\boldsymbol{x}(k)$ is an $n \times 1$ state-variable vector, $u(k)$ is a scalar input, $y(k)$ is a scalar output, and $\boldsymbol{F}_o, \boldsymbol{g}_o$, and $\boldsymbol{h}_o$ are real constant matrices of $n \times n$, $n \times 1$, and $1 \times n$, respectively. The transfer function of the linear discrete-time system in (1) is expressed as

$$H_o(z) = \boldsymbol{h}_o(z\boldsymbol{I}_n - \boldsymbol{F}_o)^{-1}\boldsymbol{g}_o. \tag{2}$$

Then the full-order state observer and associated state feedback law are modelled as

$$\begin{aligned} \hat{\boldsymbol{x}}(k+1) &= \boldsymbol{E}_o\hat{\boldsymbol{x}}(k) + \boldsymbol{g}_o u(k) + \boldsymbol{b}_o y(k) \\ u(k) &= -\boldsymbol{f}_o\hat{\boldsymbol{x}}(k) + r(k) \end{aligned} \tag{3}$$

where $\hat{\boldsymbol{x}}(k)$ is an $n \times 1$ state-variable vector in the full-order state observer with an $n \times 1$ gain vector $\boldsymbol{b}_o$, $r(k)$ is a scalar reference signal, $\boldsymbol{f}_o$ is a $1 \times n$ state-feedback coefficient vector and $\boldsymbol{E}_o = \boldsymbol{F}_o - \boldsymbol{b}_o\boldsymbol{h}_o$. It is obvious that the full-order state observer and associated state feedback system in (3) are equivalent to

$$\begin{aligned} \hat{\boldsymbol{x}}(k+1) &= \boldsymbol{\Phi}_o\hat{\boldsymbol{x}}(k) + \boldsymbol{g}_o r(k) + \boldsymbol{b}_o y(k) \\ u(k) &= -\boldsymbol{f}_o\hat{\boldsymbol{x}}(k) + r(k) \end{aligned} \tag{4}$$

where $\boldsymbol{\Phi}_o = \boldsymbol{F}_o - \boldsymbol{b}_o\boldsymbol{h}_o - \boldsymbol{g}_o\boldsymbol{f}_o$.

Next, we consider the FWL implementation of the system due to the quantizations that are carried out before matrix-vector multiplication. At the same time, we assume that an EF is integrated into the system for noise reduction purposes. The model under these circumstances becomes

$$\begin{aligned} \tilde{\boldsymbol{x}}(k+1) &= \boldsymbol{\Phi}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(k)] + \boldsymbol{g}r(k) + \boldsymbol{b}y(k) + \boldsymbol{C}\boldsymbol{e}(k) \\ u(k) &= -\boldsymbol{f}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(k)] + r(k) \end{aligned} \tag{5}$$

where $\boldsymbol{e}(k) = \tilde{\boldsymbol{x}}(k) - \boldsymbol{Q}[\tilde{\boldsymbol{x}}(k)]$ and $\boldsymbol{C}$ is an $n \times n$ EF matrix. The coefficient matrices $\boldsymbol{\Phi}$, $\boldsymbol{g}$, $\boldsymbol{b}$, and $\boldsymbol{f}$ in (5) are representations of $\boldsymbol{\Phi}_o$, $\boldsymbol{g}_o$, $\boldsymbol{b}_o$, and $\boldsymbol{f}_o$, respectively, with exact fractional B_c-bit. The FWL state-variable vector $\tilde{\boldsymbol{x}}(k)$ and each output $u(k)$ in (5) has B-bit fractional representations, while the reference input $r(k)$ and the signal $y(k)$ are $(B - B_c)$-bit fractions. The quantizer $\boldsymbol{Q}[\cdot]$ in (5) rounds the B-bit fraction $\tilde{\boldsymbol{x}}(k)$ to $(B - B_c)$-bit after the multiplications and additions, where the sign bit is not counted. It is assumed that the quantization error vector $\boldsymbol{e}(k)$ can be modeled as a Gaussian random process with zero mean and covariance $\sigma^2 \boldsymbol{I}_n$.

The closed-loop system that consists of the systems in (1) and (5) is then described by

$$\begin{aligned}\begin{bmatrix} \boldsymbol{x}(k+1) \\ \tilde{\boldsymbol{x}}(k+1) \end{bmatrix} &= \begin{bmatrix} \boldsymbol{F}_o & -\boldsymbol{g}_o\boldsymbol{f} \\ \boldsymbol{b}\boldsymbol{h}_o & \boldsymbol{\Phi} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(k) \\ \tilde{\boldsymbol{x}}(k) \end{bmatrix} \\ &\quad + \begin{bmatrix} \boldsymbol{g}_o \\ \boldsymbol{g} \end{bmatrix} r(k) + \begin{bmatrix} \boldsymbol{g}_o\boldsymbol{f} \\ \boldsymbol{C} - \boldsymbol{\Phi} \end{bmatrix} \boldsymbol{e}(k) \qquad (6) \\ y(k) &= \begin{bmatrix} \boldsymbol{h}_o & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(k) \\ \tilde{\boldsymbol{x}}(k) \end{bmatrix}.\end{aligned}$$

From (6), the transfer function from the quantization error vector $\boldsymbol{e}(k)$ to the output $y(k)$ can be written as

$$\boldsymbol{H}_C(z) = \tilde{\boldsymbol{h}}(z\boldsymbol{I}_{2n} - \tilde{\boldsymbol{F}})^{-1}\tilde{\boldsymbol{G}} \qquad (7)$$

where

$$\tilde{\boldsymbol{F}} = \begin{bmatrix} \boldsymbol{F}_o & -\boldsymbol{g}_o\boldsymbol{f} \\ \boldsymbol{b}\boldsymbol{h}_o & \boldsymbol{\Phi} \end{bmatrix}, \quad \tilde{\boldsymbol{G}} = \begin{bmatrix} \boldsymbol{g}_o\boldsymbol{f} \\ \boldsymbol{C} - \boldsymbol{\Phi} \end{bmatrix}, \quad \tilde{\boldsymbol{h}} = \begin{bmatrix} \boldsymbol{h}_o & \boldsymbol{0} \end{bmatrix}.$$

Based on the model developed above, we now define the normalized noise gain $J_1(\boldsymbol{C}) = \sigma_{out}^2/\sigma^2$ in terms of the transfer function $\boldsymbol{H}_C(z)$ in (7) as

$$J_1(\boldsymbol{C}) = \mathrm{tr}\left[\frac{1}{2\pi j}\oint_{|z|=1} \boldsymbol{H}_C^*(z)\boldsymbol{H}_C(z)\frac{dz}{z}\right]. \qquad (8)$$

To derive a formula for evaluating the normalized noise gain $J_1(\boldsymbol{C})$, we replace $\boldsymbol{f}$, $\boldsymbol{b}$, and $\boldsymbol{\Phi}$ in matrices $\tilde{\boldsymbol{F}}$ and $\tilde{\boldsymbol{G}}$ by their infinite-precision counterparts $\boldsymbol{f}_o$, $\boldsymbol{b}_o$, and $\boldsymbol{\Phi}_o$, respectively. Then an approximate version of $\boldsymbol{H}_C(z)$ in (7) is found to be

$$\begin{aligned}\boldsymbol{H}_C(z) &\simeq \tilde{\boldsymbol{h}}\boldsymbol{S}(z\boldsymbol{I}_{2n} - \boldsymbol{S}\tilde{\boldsymbol{F}}\boldsymbol{S})^{-1}\boldsymbol{S}\tilde{\boldsymbol{G}} \\ &= \tilde{\boldsymbol{h}}(z\boldsymbol{I}_{2n} - \boldsymbol{\Psi})^{-1}\boldsymbol{V}(z\boldsymbol{I}_n - \boldsymbol{C})\end{aligned} \qquad (9)$$

where

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{I}_n & -\boldsymbol{I}_n \end{bmatrix}, \boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{F}_o - \boldsymbol{g}_o\boldsymbol{f}_o & \boldsymbol{g}_o\boldsymbol{f}_o \\ \boldsymbol{0} & \boldsymbol{E}_o \end{bmatrix}, \boldsymbol{V} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I}_n \end{bmatrix}.$$

Substituting (9) into (8) yields

$$\begin{aligned}J_1(\boldsymbol{C}) = \mathrm{tr}\Big[&\boldsymbol{C}^T\boldsymbol{Y}_4\boldsymbol{C} - (\boldsymbol{f}_o^T\boldsymbol{g}_o^T\boldsymbol{Y}_3^T + \boldsymbol{E}_o^T\boldsymbol{Y}_4)\boldsymbol{C} \\ &-\boldsymbol{C}^T(\boldsymbol{Y}_3\boldsymbol{g}_o\boldsymbol{f}_o + \boldsymbol{Y}_4\boldsymbol{E}_o) + \boldsymbol{Y}_4\Big]\end{aligned} \qquad (10)$$

where

$$\begin{bmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 \\ \boldsymbol{Y}_3 & \boldsymbol{Y}_4 \end{bmatrix} = \boldsymbol{\Psi}^T \begin{bmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 \\ \boldsymbol{Y}_3 & \boldsymbol{Y}_4 \end{bmatrix} \boldsymbol{\Psi} + \begin{bmatrix} \boldsymbol{h}_o^T \\ \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{h}_o & \boldsymbol{0} \end{bmatrix}.$$

To reach an adequate formulation for the problem at hand, we derive an explicit expression for the controllability Gramian $\boldsymbol{K}$ for state-variable vector $\tilde{\boldsymbol{x}}(k)$ in system (6), because $\boldsymbol{K}$ is a critical quantity involved in the l_2-scaling constraints on $\tilde{\boldsymbol{x}}(k)$. To this end, the transfer function from reference input $r(k)$ to state-variable vector $[\boldsymbol{x}(k)^T, \tilde{\boldsymbol{x}}(k)^T]^T$ is derived by setting $\boldsymbol{e}(k) = \boldsymbol{0}$ in (6) and taking z-transform of (6), i.e.

$$\begin{bmatrix} \boldsymbol{X}(z) \\ \tilde{\boldsymbol{X}}(z) \end{bmatrix} = (z\boldsymbol{I}_{2n} - \tilde{\boldsymbol{F}})^{-1} \begin{bmatrix} \boldsymbol{g}_o \\ \boldsymbol{g} \end{bmatrix} R(z) \qquad (11)$$

where $\boldsymbol{X}(z)$ and $\tilde{\boldsymbol{X}}(z)$ denote the z-transforms of $\boldsymbol{x}(k)$ and $\tilde{\boldsymbol{x}}(k)$, respectively. In order to derive a formula for evaluating a good approximate version of the controllability Gramian, we replace matrices $\boldsymbol{\Phi}$, $\boldsymbol{g}$, $\boldsymbol{f}$ and $\boldsymbol{b}$ in (11) by $\boldsymbol{\Phi}_o$, $\boldsymbol{g}_o$, $\boldsymbol{f}_o$ and $\boldsymbol{b}_o$, respectively, which yields

$$\boldsymbol{S}\begin{bmatrix} \boldsymbol{X}(z) \\ \tilde{\boldsymbol{X}}(z) \end{bmatrix} = (z\boldsymbol{I}_{2n} - \boldsymbol{S}\tilde{\boldsymbol{F}}\boldsymbol{S})^{-1}\boldsymbol{S}\begin{bmatrix} \boldsymbol{g}_o \\ \boldsymbol{g}_o \end{bmatrix} R(z). \qquad (12)$$

Hence

$$\tilde{\boldsymbol{X}}(z) = \boldsymbol{X}(z) = \boldsymbol{F}(z)R(z) \qquad (13)$$

where $\boldsymbol{F}(z) = [z\boldsymbol{I}_n - (\boldsymbol{F}_o - \boldsymbol{g}_o\boldsymbol{f}_o)]^{-1}\boldsymbol{g}_o$. The controllability Gramian can now be defined as

$$\boldsymbol{K} = \frac{1}{2\pi j}\oint_{|z|=1} \boldsymbol{F}(z)\boldsymbol{F}^*(z)\frac{dz}{z} \qquad (14)$$

which can be obtained by solving the Lyapunov equation

$$\boldsymbol{K} = (\boldsymbol{F}_o - \boldsymbol{g}_o\boldsymbol{f}_o)\boldsymbol{K}(\boldsymbol{F}_o - \boldsymbol{g}_o\boldsymbol{f}_o)^T + \boldsymbol{g}_o\boldsymbol{g}_o^T. \qquad (15)$$

We are now in a position to consider a family of different yet equivalent state-space descriptions of (4), which can be obtained via coordinate transformation $\overline{\boldsymbol{x}}(k) = \boldsymbol{T}^{-1}\hat{\boldsymbol{x}}(k)$ as $(\overline{\boldsymbol{\Phi}}_o, \overline{\boldsymbol{g}}_o, \overline{\boldsymbol{b}}_0, \overline{\boldsymbol{f}}_o)_n$, where

$$\begin{aligned}\overline{\boldsymbol{\Phi}}_o &= \boldsymbol{T}^{-1}\boldsymbol{\Phi}_o\boldsymbol{T}, &\quad \overline{\boldsymbol{g}}_o &= \boldsymbol{T}^{-1}\boldsymbol{g}_o \\ \overline{\boldsymbol{b}}_o &= \boldsymbol{T}^{-1}\boldsymbol{b}_o, &\quad \overline{\boldsymbol{f}}_o &= \boldsymbol{f}_o\boldsymbol{T}.\end{aligned} \qquad (16)$$

Accordingly, the controllability and observability Gramians relating to $(\overline{\boldsymbol{\Phi}}_o, \overline{\boldsymbol{g}}_o, \overline{\boldsymbol{b}}_o, \overline{\boldsymbol{f}}_o)_n$ can be expressed as

$$\overline{\boldsymbol{K}} = \boldsymbol{T}^{-1}\boldsymbol{K}\boldsymbol{T}^{-T}, \quad \overline{\boldsymbol{Y}}_i = \boldsymbol{T}^T\boldsymbol{Y}_i\boldsymbol{T} \qquad (17)$$

for $1 \le i \le 4$, respectively. In order to prevent the new state-variables from overflow oscillations, the l_2-scaling constraints are imposed on the state-variable vector $\overline{\boldsymbol{x}}(k)$ so that

$$(\overline{\boldsymbol{K}})_{ii} = (\boldsymbol{T}^{-1}\boldsymbol{K}\boldsymbol{T}^{-T})_{ii} = 1 \ \text{ for } i = 1, 2, \cdots, n. \qquad (18)$$

With an equivalent state-space realization as specified in (16), the normalized noise gain is found to be

$$\begin{aligned}J_2(\boldsymbol{C}, \boldsymbol{T}) = \mathrm{tr}\Big[\boldsymbol{T}^T\boldsymbol{Y}_4\boldsymbol{T}\boldsymbol{C}\boldsymbol{C}^T\Big] + \mathrm{tr}\Big[\boldsymbol{T}^T\boldsymbol{Y}_4\boldsymbol{T}\Big] \\ - 2\,\mathrm{tr}\Big[\boldsymbol{T}^T(\boldsymbol{Y}_3\boldsymbol{g}_o\boldsymbol{f}_o + \boldsymbol{Y}_4\boldsymbol{E}_o)^T\boldsymbol{T}\boldsymbol{C}\Big].\end{aligned} \qquad (19)$$

The problem considered here is to design the EF matrix $\boldsymbol{C}$ and coordinate transformation matrix $\boldsymbol{T}$ to minimize $J_2(\boldsymbol{C},\boldsymbol{T})$ in (19) subject to the l_2-scaling constraints in (18).

3. OPTIMIZATION OF ERROR FEEDBACK AND REALIZATION

3.1. A Separate Optimization Technique

To perform the first step, define the Lagrange function

$$J(\boldsymbol{P},\lambda)=\text{tr}[\boldsymbol{Y}_4\boldsymbol{P}]+\lambda\,(\text{tr}[\boldsymbol{K}\boldsymbol{P}^{-1}]-n) \tag{20}$$

where $\boldsymbol{P}=\boldsymbol{T}\boldsymbol{T}^T$. The optimal coordinate transformation matrix $\boldsymbol{T}$ can be determined by solving the equations

$$\frac{\partial J(\boldsymbol{P},\lambda)}{\partial \boldsymbol{P}}=\boldsymbol{0},\quad \frac{\partial J(\boldsymbol{P},\lambda)}{\partial \lambda}=0 \tag{21}$$

which lead to $\boldsymbol{P}\boldsymbol{Y}_4\boldsymbol{P}=\lambda\boldsymbol{K}$ and $\text{tr}[\boldsymbol{K}\boldsymbol{P}^{-1}]=n$. Then by some manupulations it follows that

$$\boldsymbol{P}=\frac{1}{n}\Big(\sum_{i=1}^{n}\theta_i\Big)\boldsymbol{Y}_4^{-\frac{1}{2}}\Big[\boldsymbol{Y}_4^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{Y}_4^{\frac{1}{2}}\Big]^{\frac{1}{2}}\boldsymbol{Y}_4^{-\frac{1}{2}} \tag{22}$$

where θ_i^2 denotes the eigenvalue of $\boldsymbol{K}\boldsymbol{Y}_4$. Substituting (22) into (20) yields the minimum value of $J(\boldsymbol{P},\lambda)$ as

$$\min_{\boldsymbol{P},\lambda} J(\boldsymbol{P},\lambda)=\frac{1}{n}\Big(\sum_{i=1}^{n}\theta_i\Big)^2. \tag{23}$$

Referring to (22), the optimal coordinate transformation matrix $\boldsymbol{T}$ that minimizes (20) is obtained in a closed form as

$$\boldsymbol{T}=\frac{1}{\sqrt{n}}\Big(\sum_{i=1}^{n}\theta_i\Big)^{\frac{1}{2}}\boldsymbol{Y}_4^{-\frac{1}{2}}\Big[\boldsymbol{Y}_4^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{Y}_4^{\frac{1}{2}}\Big]^{\frac{1}{4}}\boldsymbol{V}_o \tag{24}$$

where $\boldsymbol{V}_o$ is an arbitrary $n\times n$ orthogonal matrix. From (24) we obtain

$$\overline{\boldsymbol{K}}=n\Big(\sum_{i=1}^{n}\theta_i\Big)^{-1}\boldsymbol{V}_o^T\Big[\boldsymbol{Y}_4^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{Y}_4^{\frac{1}{2}}\Big]^{\frac{1}{2}}\boldsymbol{V}_o \tag{25}$$

Next, we choose the $n\times n$ orthogonal matrix $\boldsymbol{V}_o$ such that matrix $\overline{\boldsymbol{K}}$ in (25) satisfies the l_2-scaling constraints in (18). To this end, we perform the eigenvalue-eigenvector decomposition

$$\Big[\boldsymbol{Y}_4^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{Y}_4^{\frac{1}{2}}\Big]^{\frac{1}{2}}=\boldsymbol{Q}\,\text{diag}\{\theta_1,\theta_2,\cdots,\theta_n\}\boldsymbol{Q}^T \tag{26}$$

where $\boldsymbol{Q}\boldsymbol{Q}^T=\boldsymbol{I}_n$. As a result, it follows that

$$n\Big(\sum_{i=1}^{n}\theta_i\Big)^{-1}\Big[\boldsymbol{Y}_4^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{Y}_4^{\frac{1}{2}}\Big]^{\frac{1}{2}}=\boldsymbol{Q}\boldsymbol{\Lambda}^{-2}\boldsymbol{Q}^T \tag{27}$$

where

$$\boldsymbol{\Lambda}=\text{diag}\{\lambda_1,\lambda_2,\cdots,\lambda_n\},\quad \lambda_i=\Big(\frac{\theta_1+\theta_2+\cdots+\theta_n}{n\theta_i}\Big)^{\frac{1}{2}}.$$

Now, an $n\times n$ orthogonal matrix $\boldsymbol{Z}$ such that

$$(\boldsymbol{Z}\boldsymbol{\Lambda}^{-2}\boldsymbol{Z}^T)_{ii}=1\ \text{ for }\ i=1,2,\cdots,n \tag{28}$$

can be obtained by numerical manipulations [7, p.278]. By choosing $\boldsymbol{V}_o=\boldsymbol{Q}\boldsymbol{Z}^T$ in (24), the optimal coordinate transformation matrix $\boldsymbol{T}$ satisfying (18) and (23) can be constructed as

$$\boldsymbol{T}=\frac{1}{\sqrt{n}}\Big(\sum_{i=1}^{n}\theta_i\Big)^{\frac{1}{2}}\boldsymbol{Y}_4^{-\frac{1}{2}}\Big[\boldsymbol{Y}_4^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{Y}_4^{\frac{1}{2}}\Big]^{\frac{1}{4}}\boldsymbol{Q}\boldsymbol{Z}^T. \tag{29}$$

Using (19), we compute $\partial J_2(\boldsymbol{C},\boldsymbol{T})/\partial\boldsymbol{C}$ and set it to null that leads to

$$\boldsymbol{C}=\boldsymbol{T}^{-1}\boldsymbol{Y}_4^{-1}(\boldsymbol{Y}_3\boldsymbol{g}_o\boldsymbol{f}_o+\boldsymbol{Y}_4\boldsymbol{E}_o)\boldsymbol{T}. \tag{30}$$

This provides an analytic method for reducing the round-off noise at the output of a closed-loop system subject to l_2-scaling constraints by using EF with a general matrix $\boldsymbol{C}$.

3.2. An Iterative Optimization Technique

The normalized noise gain in (10) is equivalent to

$$\begin{aligned}J_1(\boldsymbol{C})=\text{tr}\,\big[&(\boldsymbol{M}-\boldsymbol{C})^T\boldsymbol{Y}_4(\boldsymbol{M}-\boldsymbol{C})\\&+\boldsymbol{Y}_4-\boldsymbol{M}^T\boldsymbol{Y}_4\boldsymbol{M}\big]\end{aligned} \tag{31}$$

where $\boldsymbol{M}=\boldsymbol{E}_o+\boldsymbol{L}\boldsymbol{g}_o\boldsymbol{f}_o$ and $\boldsymbol{L}$ is an $n\times n$ matrix satisfying $\boldsymbol{Y}_4\boldsymbol{L}=\boldsymbol{Y}_3$. With an equivalent state-space realization as specified in (16), the normalized noise gain in (31) is changed to

$$\begin{aligned}J_2(\boldsymbol{C},\boldsymbol{T})=\text{tr}\,\big[&(\boldsymbol{M}\boldsymbol{T}-\boldsymbol{T}\boldsymbol{C})^T\boldsymbol{Y}_4(\boldsymbol{M}\boldsymbol{T}-\boldsymbol{T}\boldsymbol{C})\\&+\boldsymbol{T}^T(\boldsymbol{Y}_4-\boldsymbol{M}^T\boldsymbol{Y}_4\boldsymbol{M})\boldsymbol{T}\big].\end{aligned} \tag{32}$$

Because the constraints in (18) are nonconvex and highly nonlinear with respect to matrix $\boldsymbol{T}$, it is beneficial if these constraints can be eliminated so as to work with an unconstrained problem that admits fast Newton-like algorithms. To this end, we start by defining

$$\hat{\boldsymbol{T}}=\boldsymbol{T}^T\boldsymbol{K}^{-\frac{1}{2}} \tag{33}$$

which leads (18) to

$$(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{T}}^{-1})_{ii}=1\ \text{ for }\ i=1,2,\cdots,n. \tag{34}$$

These constraints are always satisfied if the matrix $\hat{\boldsymbol{T}}^{-1}$ assumes the form

$$\hat{\boldsymbol{T}}^{-1}=\Big[\frac{\boldsymbol{t}_1}{||\boldsymbol{t}_1||},\frac{\boldsymbol{t}_2}{||\boldsymbol{t}_2||},\cdots,\frac{\boldsymbol{t}_n}{||\boldsymbol{t}_n||}\Big]. \tag{35}$$

By virtue of (33), the normalized noise gain in (32) can be expressed as

$$\begin{aligned}&J_3(\boldsymbol{C},\hat{\boldsymbol{T}})\\&=\text{tr}[(\hat{\boldsymbol{M}}\hat{\boldsymbol{T}}^T-\hat{\boldsymbol{T}}^T\boldsymbol{C})^T\hat{\boldsymbol{Y}}_4(\hat{\boldsymbol{M}}\hat{\boldsymbol{T}}^T-\hat{\boldsymbol{T}}^T\boldsymbol{C})]\\&\quad+\text{tr}[\hat{\boldsymbol{T}}(\hat{\boldsymbol{Y}}_4-\hat{\boldsymbol{M}}^T\hat{\boldsymbol{Y}}_4\hat{\boldsymbol{M}})\hat{\boldsymbol{T}}^T]\end{aligned} \tag{36}$$

where $\hat{\boldsymbol{M}} = \boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{M}\boldsymbol{K}^{\frac{1}{2}}$ and $\hat{\boldsymbol{Y}}_4 = \boldsymbol{K}^{\frac{1}{2}}\boldsymbol{Y}_4\boldsymbol{K}^{\frac{1}{2}}$. From (36) it is evident that the optimal choice of $\boldsymbol{C}$ is given by

$$\boldsymbol{C} = \hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{M}}\hat{\boldsymbol{T}}^{T} \tag{37}$$

which leads to

$$J_3(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{M}}\hat{\boldsymbol{T}}^{T}, \hat{\boldsymbol{T}}) = \mathrm{tr}[\hat{\boldsymbol{T}}(\hat{\boldsymbol{Y}}_4 - \hat{\boldsymbol{M}}^{T}\hat{\boldsymbol{Y}}_4\hat{\boldsymbol{M}})\hat{\boldsymbol{T}}^{T}]. \tag{38}$$

The above derivation clearly indicates that the joint optimization of EF and realization is not needed because the optimal $\boldsymbol{C}$ has already been determined by the closed-form in (37), hence it is sufficient to optimize matrix $\hat{\boldsymbol{T}}$ by minimizing $J_3(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{M}}\hat{\boldsymbol{T}}^{T}, \hat{\boldsymbol{T}})$ in (38).

The solution of minimizing $J_3(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{M}}\hat{\boldsymbol{T}}^{T}, \hat{\boldsymbol{T}})$ in (38) was found to be fairly sensitive to how the EF matrix $\boldsymbol{C}$ was quantized in the system implementation. As a remedy for this sensitivity issue, we consider a regularized minimization problem with the objective function given by

$$\begin{aligned} J_o(\boldsymbol{x}) &= \mu\,\mathrm{tr}[\hat{\boldsymbol{T}}\hat{\boldsymbol{Y}}_4\hat{\boldsymbol{T}}^{T}] \\ &\quad + (1-\mu)\,\mathrm{tr}[\hat{\boldsymbol{T}}(\hat{\boldsymbol{Y}}_4 - \hat{\boldsymbol{M}}^{T}\hat{\boldsymbol{Y}}_4\hat{\boldsymbol{M}})\hat{\boldsymbol{T}}^{T}] \\ &= \mathrm{tr}[\hat{\boldsymbol{T}}\{\hat{\boldsymbol{Y}}_4 - (1-\mu)\,\hat{\boldsymbol{M}}^{T}\hat{\boldsymbol{Y}}_4\hat{\boldsymbol{M}}\}\hat{\boldsymbol{T}}^{T}] \end{aligned} \tag{39}$$

where $\boldsymbol{x}$ is defined by $\boldsymbol{x} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T, \cdots, \boldsymbol{t}_n^T)^T$ and μ is a regularization parameter in the range $0 < \mu \leq 1$ that can be selected to make an appropriate trade-off between minimizing (38) and reducing $\mathrm{tr}[\hat{\boldsymbol{T}}\hat{\boldsymbol{Y}}_4\hat{\boldsymbol{T}}^{T}]$.

The proposed algorithm starts with an initial point $\boldsymbol{x}_0$ obtained from the assignment $\hat{\boldsymbol{T}} = \boldsymbol{C} = \boldsymbol{I}_n$. In the kth iteration, a quasi-Newton algorithm updates the most recent point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as [12]

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{d}_k \tag{40}$$

where

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

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

$$|J_o(\boldsymbol{x}_{k+1}) - J_o(\boldsymbol{x}_k)| < \varepsilon \tag{41}$$

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

The gradients of $J_o(\boldsymbol{x})$ with respect to $\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n$ are found to be

$$\frac{\partial J_o(\boldsymbol{x})}{\partial t_{ij}} = 2\boldsymbol{e}_j^T\hat{\boldsymbol{T}}\{\hat{\boldsymbol{Y}}_4 - (1-\mu)\,\hat{\boldsymbol{M}}^{T}\hat{\boldsymbol{Y}}_4\hat{\boldsymbol{M}}\}\hat{\boldsymbol{T}}^{T}\hat{\boldsymbol{T}}\boldsymbol{\Delta}_{ij} \tag{42}$$

where t_{ij} is the ith element of column vector $\boldsymbol{t}_j$, and

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

4. NUMERICAL EXPERIMENTS

Consider a linear discrete-time system specified by

$$\boldsymbol{F}_o = \begin{bmatrix} 0 & 0 & 0.775585 \\ 1 & 0 & -2.534177 \\ 0 & 1 & 2.758200 \end{bmatrix}^T, \quad \boldsymbol{g}_o = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$
$$\boldsymbol{h}_o = \begin{bmatrix} 0.0022 & 0.0044 & 0.0022 \end{bmatrix}.$$

Suppose that the full-order state observer has the poles at $z = 0.4532$, 0.5761 and 0.8437, and that the poles of the regulator are placed at $z = 0.9067$, 0.7523 and 0.6231. Then, the coefficient vector of the full-order state observer and the gain vector of the state feedback were found to be

$$\begin{aligned} \boldsymbol{b}_o &= 100\begin{bmatrix} 0.818859 & 1.010891 & 1.182995 \end{bmatrix}^T \\ \boldsymbol{f}_o &= \begin{bmatrix} 0.350562 & -0.818344 & 0.476100 \end{bmatrix}. \end{aligned}$$

The controllability and observability Gramians $\boldsymbol{K}$ and $\boldsymbol{Y}_4$ relating to the closed-loop discrete-time system in (6) were computed from (15) and (10) as

$$\boldsymbol{K} = 10^2\begin{bmatrix} 4.510584 & 4.474816 & 4.374469 \\ 4.474816 & 4.510584 & 4.474816 \\ 4.374469 & 4.474816 & 4.510584 \end{bmatrix}$$
$$\boldsymbol{Y}_4 = 10\begin{bmatrix} 0.469714 & -1.031460 & 0.556758 \\ -1.031460 & 2.265722 & -1.223243 \\ 0.556758 & -1.223243 & 0.660514 \end{bmatrix}.$$

By performing l_2-scaling to the system with

$$\boldsymbol{T}_o = 10\ \mathrm{diag}\{2.123814, 2.123814, 2.123814\}$$

so that $(\boldsymbol{T}_o^{-1}\boldsymbol{K}\boldsymbol{T}_o^{-T})_{ii} = 1$ is satisfied for $i = 1, 2, 3$, the normalized noise gain of the system without EF was computed from (10) as

$$J_1(\boldsymbol{0}) = \mathrm{tr}[\boldsymbol{T}_o^T\boldsymbol{Y}_4\boldsymbol{T}_o] = 1.531772 \times 10^3.$$

The quasi-Newton algorithm was applied to minimize (39) with $\mu = 0.01$ and tolerance $\varepsilon = 10^{-8}$ in (41). The initial normalized noise gain was found to be $J_o(\boldsymbol{x}) = 10.084792 \times 10^{-3}$ and it took the algorithm 15 iterations to converge to the solution

$$\hat{\boldsymbol{T}} = \begin{bmatrix} 1.957936 & 0.179654 & 0.105801 \\ -1.161786 & 1.125098 & 1.523983 \\ 0.649386 & 0.333143 & 2.019611 \end{bmatrix}$$
$$\boldsymbol{C} = \begin{bmatrix} 1.040052 & 0.157842 & -0.093433 \\ -0.581603 & 0.837674 & 0.134128 \\ -0.535618 & -0.025389 & 1.049268 \end{bmatrix}$$

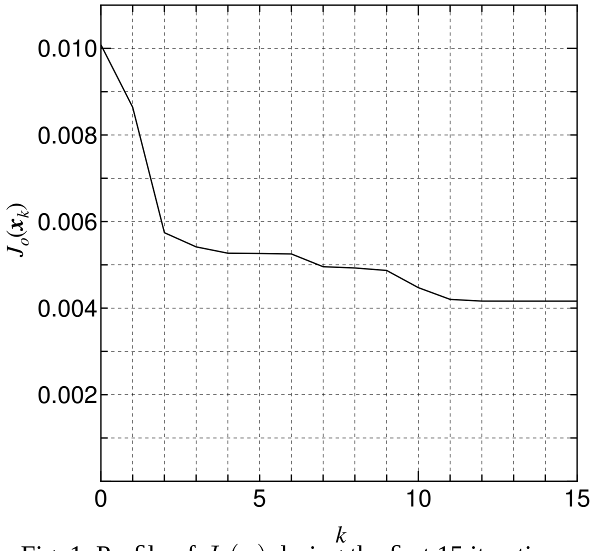

Fig. 1 Profile of $J_o(\boldsymbol{x})$ during the first 15 iterations.

with which the normalized noise gain was reduced to $J_o(\boldsymbol{x}) = 4.161830 \times 10^{-3}$. The profile of $J_o(\boldsymbol{x})$ during the first 15 iterations of the algorithm is dipicted in Fig. 1. The performance of the iterative algorithm in terms of the normalized noise gain $J_3(\boldsymbol{C}, \hat{\boldsymbol{T}})$ in (36) was compared with that achieved by a separate optimization technique for two implementation settings. The numerical results for the two implementation settings are summarized in Table 1 where the column with "Infinite Precision" shows the value of $J_3(\boldsymbol{C}, \hat{\boldsymbol{T}})$ derived from the optimal $\hat{\boldsymbol{T}}$ and $\boldsymbol{C}$. The column with "Integer Quantization" means that of $J_3(\boldsymbol{C}, \hat{\boldsymbol{T}})$ where each entry of the optimal $\boldsymbol{C}$ was rounded to an integer and $\hat{\boldsymbol{T}}$ still remains optimal. From the table, it is observed that the iterative optimization technique consistently outperforms the separate optimization technique.

Table 1 Performance comparison of two roundoff noise minimization techniques.

Optimization	Infinite Precision	Integer Quantization
Separate	5.132825×10^{-4}	39.236503×10^{-4}
Iterative	1.926278×10^{-4}	25.238441×10^{-4}

5. CONCLUSION

This paper has investigated the optimization problem of EF and realization to minimize the roundoff noise subject to l_2-scaling constraints in a system in which full-order state observer feedback is used in place of system state feedback. It has been shown that joint optimization of EF and realization is not needed when utilizing a general-matrix EF loop, because the optimal EF matrix, in closed-form, can be determined in advance. An efficient iterative technique for obtaining the optimal coordinate transformation matrix that minimizes the roundoff noise has been developed by employing a quasi-Newtom algorithm. Numerical experiments have demonstrated the validity and effectiveness of the proposed technique and has demonstrated its performance.

REFERENCES

[1] 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 Process., vol. 30, no. 6, pp. 963-973, Dec. 1982.

[2] W. E. Higgins and D. C. Munson, Optimal and suboptimal error spectrum shaping for cascade-form digital filters, IEEE Trans. Circuits Syst., vol. 31, no. 5, pp. 429-437, May 1984.

[3] P. P. Vaidyanathan, On error-spectrum shaping in state-space digital filters, IEEE Trans. Circuits Syst., vol. 32, no. 1, pp. 88-92, Jan. 1985.

[4] D. Williamson, Roundoff noise minimization and pole-zero sensitivity in fixed-point digital filters using residue feedback, IEEE Trans. Acoust., Speech, Signal Process., vol. 34, no. 5, pp. 1210-1220, Oct. 1986.

[5] T. I. Laakso and I. O. Hartimo, Noise reduction in recursive digital filters using high-order error feedback, IEEE Trans. Signal Processing, vol. 40, no. 5, pp. 1096-1107, May 1992.

[6] C. T. Mullis and R. A. Roberts, Synthesis of minimum roundoff noise fixed point digital filters, IEEE Trans. Circuits Syst., vol. 23, no. 9, pp. 551-562, Sep. 1976.

[7] S. Y. Hwang, Minimum uncorrelated unit noise in state-space digital filtering, IEEE Trans. Acoust., Speech, Signal Process., vol. 25, no. 4, pp. 273-281, Aug. 1977.

[8] T. Hinamoto, H. Ohnishi, and W.-S. Lu, Roundoff noise minimization of state-space digital filters using separate and joint error feedback/coordinate transformation optimization, IEEE Trans. Circuits Syst. I, vol. 50, no. 1, pp. 23-33, Jan. 2003.

[9] 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, no. 6, pp. 2135-2145, Jun. 2005.

[10] G. Li and M. Gevers, Optimal finite precision implementation of a state-estimate feedback controller, IEEE Trans. Circuits Syst., vol. 37, no. 12, pp. 1487-1498, Dec. 1990.

[11] T. Hinamoto, A. Doi and W.-S. Lu, Optimal error feedback and realization for roundoff noise minimization in linear discrete-time systems with full-order state observer feedback, in Proc. IEEE Int. Symp. Circuits Syst. (ISCAS'15), Lisbon, Portugal, May 2015 (to appear).

[12] R. Fletcher, Practical Methods of Optimization 2^{nd} ed. New York, NY, USA: Wiley, 1987.