Joint Optimization of High-Order Error Feedback and Realization for Roundoff Noise Minimization in State-Estimate Feedback Controllers

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*— Joint optimization of high-order error-feedback and state-space realization to minimize roundoff noise at the output of a closed-loop system with a state-estimate feedback controller subject to l_2-scaling constraints is investigated. The problem is converted into an unconstrained optimization problem by using linear-algebraic techniques. The unconstrained optimization problem at hand is then solved iteratively by employing an efficient quasi-Newton algorithm with closed-form formulas for key gradient evaluation. A numerical example is presented to illustrate the utility of the proposed technique.**

1. INTRODUCTION

Owing to finite-precision nature of computer arithmetic, roundoff noise in fixed-point IIR digital filters occurs that in turn degrades the filter's performance. Error feedback (EF) is known as an effective technique for reducing finite-word-lengh (FWL) effects in IIR digital filters, and many EF methods have been proposed in the past [1]-[5]. On the other hand, it is known that the roundoff noise is critically dependent on the internal structure of the IIR filter, and several techniques for constructing a state-space filter structure that minimizes the roundoff noise gain at the filter output subject to l_2-scaling constraints have been explored by appropriately choosing a linear transformation to 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 state-space realization for achieving better performance [8]-[10].

Another relevant development was initiated in [11] in early 1990's, where roundoff noise in a closed-loop system equipped with a state-estimate feedback controller is analyzed and an algorithm for realizing a state-estimate feedback controller which minimizes the roundoff noise subject to l_2-scaling constraints is proposed. However, no EF is incorporated into the algorithm in [11]. Recently, the joint optimization problem of constant EF and state-space realization for a state-estimate feedback controller has been considered to minimize the roundoff noise gain at the output of the closed-loop system with a state-estimate feedback controller subject to l_2-scaling constraints [12].

In this paper, joint optimization of *high-order* EF and realization for minimizing roundoff noise in the closed-loop system with a state-estimate feedback controller subject to l_2-scaling constraints is investigated. The novelty of the proposed method has largely to do with the introduction of *high-order* EF to the closed-loop system with a state-estimate feedback controller. We convert the constrained optimization problem encountered into an unconstrained optimization problem by using linear-algebraic techniques. An efficient quasi-Newton algorithm is utilized to solve the unconstrained optimization problem at hand. The proposed technique is applied to the case where *high-order* EF matrices are diagonal. A numerical example is presented to illustrate the validity and effectiveness of the proposed technique.

2. SYSTEM MODEL AND PROBLEM FORMULATION

Consider a stable, controllable and observable linear discrete-time system $(\boldsymbol{A}_o, \boldsymbol{b}_o, \boldsymbol{c}_o)_n$ described by

$$\begin{aligned} \boldsymbol{x}(k+1) &= \boldsymbol{A}_o\boldsymbol{x}(k) + \boldsymbol{b}_o u(k) \\ y(k) &= \boldsymbol{c}_o\boldsymbol{x}(k) \end{aligned} \tag{1}$$

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 $\boldsymbol{A}_o, \boldsymbol{b}_o$ and $\boldsymbol{c}_o$ are real constant matrices of appropriate dimensions. The transfer function of the linear system in (1) is given by

$$H_o(z) = \boldsymbol{c}_o(z\boldsymbol{I}_n - \boldsymbol{A}_o)^{-1}\boldsymbol{b}_o \tag{2}$$

then the observer and associated controller are modelled as

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

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 $\boldsymbol{g}_o$, $r(k)$ is a scalar reference signal, $\boldsymbol{k}_o$ is a $1 \times n$ state-feedback coefficient vector, and $\boldsymbol{F}_o = \boldsymbol{A}_o - \boldsymbol{g}_o\boldsymbol{c}_o$. Obviously, the observer-controller in (3) is equivalent to

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

where $\boldsymbol{R}_o = \boldsymbol{F}_o - \boldsymbol{b}_o\boldsymbol{k}_o$.

In the next step of system modelling, we consider the FWL implementation of the system due to the quantizations that are performed before matrix-vector multiplication. At the same time, assume that a high-order EF is integrated into the system for noise reduction purposes. The model under these circumstances becomes

$$\begin{aligned} \tilde{\boldsymbol{x}}(k+1) &= \boldsymbol{R}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(k)] + \boldsymbol{b}r(k) + \boldsymbol{g}y(k) + \sum_{i=0}^{N-1}\boldsymbol{D}_i\boldsymbol{e}(k-i) \\ u(k) &= -\boldsymbol{k}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(k)] + r(k) \end{aligned} \tag{5}$$

978-1-4799-4132-2/14/$31.00 ©2014 IEEE

where

$$\boldsymbol{e}(k) = \tilde{\boldsymbol{x}}(k) - \boldsymbol{Q}[\tilde{\boldsymbol{x}}(k)]$$

and $\boldsymbol{D}_i$ for $i = 0, 1, \cdots, N-1$ are $n \times n$ EF matrices. The coefficient matrices $\boldsymbol{R}$, $\boldsymbol{b}$, $\boldsymbol{g}$, and $\boldsymbol{k}$ in (5) are representations of $\boldsymbol{R}_o$, $\boldsymbol{b}_o$, $\boldsymbol{g}_o$, and $\boldsymbol{k}_o$, respectively, with exact fractional Bc-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 roundoff error $\boldsymbol{e}(k)$ 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

$$\begin{aligned} \begin{bmatrix} \boldsymbol{x}(k+1) \\ \tilde{\boldsymbol{x}}(k+1) \end{bmatrix} &= \begin{bmatrix} \boldsymbol{A}_o & -\boldsymbol{b}_o\boldsymbol{k} \\ \boldsymbol{g}\boldsymbol{c}_o & \boldsymbol{R} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(k) \\ \tilde{\boldsymbol{x}}(k) \end{bmatrix} + \begin{bmatrix} \boldsymbol{b}_o \\ \boldsymbol{b} \end{bmatrix} r(k) \\ &\quad + \begin{bmatrix} \boldsymbol{b}_o\boldsymbol{k} \\ \boldsymbol{D}_0 - \boldsymbol{R} \end{bmatrix} \boldsymbol{e}(k) + \sum_{i=1}^{N-1} \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{D}_i \end{bmatrix} \boldsymbol{e}(k-i) \\ y(k) &= \begin{bmatrix} \boldsymbol{c}_o & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(k) \\ \tilde{\boldsymbol{x}}(k) \end{bmatrix}. \end{aligned} \tag{6}$$

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

$$\boldsymbol{G}_D(z) = \tilde{\boldsymbol{c}}(z\boldsymbol{I}_{2n} - \tilde{\boldsymbol{A}})^{-1}\tilde{\boldsymbol{B}}(z) \tag{7}$$

where

$$\tilde{\mathbf{A}} = \begin{bmatrix} \boldsymbol{A}_o & -\boldsymbol{b}_o\boldsymbol{k} \\ \boldsymbol{g}\boldsymbol{c}_o & \boldsymbol{R} \end{bmatrix}, \qquad \tilde{\boldsymbol{c}} = \begin{bmatrix} \boldsymbol{c}_o & \boldsymbol{0} \end{bmatrix}$$

$$\tilde{\boldsymbol{B}}(z) = \boldsymbol{B}_0 + \boldsymbol{B}_1 z^{-1} + \cdots + \boldsymbol{B}_{N-1} z^{-(N-1)}$$

with

$$\boldsymbol{B}_0 = \begin{bmatrix} \boldsymbol{b}_o\boldsymbol{k} \\ \boldsymbol{D}_0 - \boldsymbol{R} \end{bmatrix}, \quad \boldsymbol{B}_i = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{D}_i \end{bmatrix} \text{ for } 1 \le i \le N-1.$$

Based on the model developed above, we now define the normalized noise gain $J_1(\boldsymbol{D}) = \sigma_{out}^2/\sigma^2$ with $\boldsymbol{D} = [\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}]$ in terms of the transfer function $\boldsymbol{G}_D(z)$ in (7) as

$$J_1(\boldsymbol{D}) = \text{tr}\left[\frac{1}{2\pi j}\oint_{|z|=1} \boldsymbol{G}_D^*(z)\boldsymbol{G}_D(z)\frac{dz}{z}\right]. \tag{8}$$

In order to derive a computationally tractable formula for evaluating the noise gain $J_1(\boldsymbol{D})$, we replace $\boldsymbol{k}$, $\boldsymbol{g}$, and $\boldsymbol{R}$ in matrices $\tilde{\boldsymbol{A}}$ and $\boldsymbol{B}_0$ by their infinite-precision counterparts $\boldsymbol{k}_o$, $\boldsymbol{g}_o$, and $\boldsymbol{R}_o$, respectively. If we define

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{I}_n & -\boldsymbol{I}_n \end{bmatrix} \tag{9}$$

then an approximate version of $\boldsymbol{G}_D(z)$ in (7) is found to be

$$\begin{aligned} \boldsymbol{G}_D(z) &\simeq \tilde{\boldsymbol{c}}\boldsymbol{S}(z\boldsymbol{I}_{2n} - \boldsymbol{S}^{-1}\tilde{\boldsymbol{A}}\boldsymbol{S})^{-1}\boldsymbol{S}^{-1}\tilde{\boldsymbol{B}}(z) \\ &= \tilde{\boldsymbol{c}}(z\boldsymbol{I}_{2n} - \boldsymbol{\Phi})^{-1}\left(\begin{bmatrix} \boldsymbol{b}_o\boldsymbol{k}_o \\ \boldsymbol{F}_o - \boldsymbol{D}_0 \end{bmatrix} - \sum_{i=1}^{N-1} \boldsymbol{B}_i z^{-i}\right) \\ &= \boldsymbol{c}_o(z\boldsymbol{I}_n - \boldsymbol{A}_o + \boldsymbol{b}_o\boldsymbol{k}_o)^{-1}\boldsymbol{b}_o\boldsymbol{k}_o(z\boldsymbol{I}_n - \boldsymbol{F}_o)^{-1} \\ &\quad \cdot\left(z\boldsymbol{I}_n - \sum_{i=0}^{N-1}\boldsymbol{D}_i z^{-i}\right) \\ &= \tilde{\boldsymbol{c}}(z\boldsymbol{I}_{2n} - \boldsymbol{\Phi})^{-1}\boldsymbol{U}\left(z\boldsymbol{I}_n - \sum_{i=0}^{N-1}\boldsymbol{D}_i z^{-i}\right) \end{aligned} \tag{10}$$

where

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{A}_o - \boldsymbol{b}_o\boldsymbol{k}_o & \boldsymbol{b}_o\boldsymbol{k}_o \\ \boldsymbol{0} & \boldsymbol{F}_o \end{bmatrix}, \qquad \boldsymbol{U} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I}_n \end{bmatrix}.$$

Under the assumption that marices $\boldsymbol{D}_i$'s are diagonal for $i = 0, 1, \cdots, N-1$, substituting (10) into (8) yields

$$\begin{aligned} J_1(\boldsymbol{D}) &= \text{tr}\,[\boldsymbol{W}_4] - 2\sum_{i=0}^{N-1}\text{tr}\,[\boldsymbol{\Psi}_4(i+1,0)\boldsymbol{D}_i] \\ &\quad + \sum_{i=0}^{N-1}\sum_{l=0}^{N-1}\text{tr}\,[\boldsymbol{\Psi}_4(i,l)\boldsymbol{D}_i\boldsymbol{D}_l] \end{aligned} \tag{11}$$

where

$$\boldsymbol{W} = \boldsymbol{\Phi}^T\boldsymbol{W}\boldsymbol{\Phi} + \begin{bmatrix} \boldsymbol{c}_o & \boldsymbol{0} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{c}_o & \boldsymbol{0} \end{bmatrix}$$

$$\boldsymbol{\Psi}(i,l) = (\boldsymbol{\Phi}^T)^i\boldsymbol{W}\boldsymbol{\Phi}^l = \begin{bmatrix} \boldsymbol{\Psi}_1(i,l) & \boldsymbol{\Psi}_2(i,l) \\ \boldsymbol{\Psi}_3(i,l) & \boldsymbol{\Psi}_4(i,l) \end{bmatrix}$$

with

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{W}_2 \\ \boldsymbol{W}_3 & \boldsymbol{W}_4 \end{bmatrix}.$$

In order 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 the 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.

$$\begin{bmatrix} \boldsymbol{X}(z) \\ \tilde{\boldsymbol{X}}(z) \end{bmatrix} = (z\boldsymbol{I}_{2n} - \tilde{\boldsymbol{A}})^{-1} \begin{bmatrix} \boldsymbol{b}_o \\ \boldsymbol{b} \end{bmatrix} R(z) \tag{12}$$

where $\boldsymbol{X}(z)$ and $\tilde{\boldsymbol{X}}(z)$ denote the z-transforms of $\boldsymbol{x}(k)$ and $\tilde{\boldsymbol{x}}(k)$, respectively. To derive a computationally tractable formula for evaluating a good approximate version of the controllability Gramian, we replace matrices $\boldsymbol{R}$, $\boldsymbol{b}$, $\boldsymbol{k}$ and $\boldsymbol{g}$ in (12) by $\boldsymbol{R}_o$, $\boldsymbol{b}_o$, $\boldsymbol{k}_o$ and $\boldsymbol{g}_o$, respectively, which yields

$$\boldsymbol{S}^{-1}\begin{bmatrix} \boldsymbol{X}(z) \\ \tilde{\boldsymbol{X}}(z) \end{bmatrix} = (z\boldsymbol{I}_{2n} - \boldsymbol{S}^{-1}\tilde{\boldsymbol{A}}\boldsymbol{S})^{-1}\boldsymbol{S}^{-1}\begin{bmatrix} \boldsymbol{b}_o \\ \boldsymbol{b}_o \end{bmatrix} R(z). \tag{13}$$

Hence

$$\tilde{\boldsymbol{X}}(z) = \boldsymbol{X}(z) = \boldsymbol{F}(z)R(z) \tag{14}$$

where

$$\boldsymbol{F}(z) = [z\boldsymbol{I}_n - (\boldsymbol{A}_o - \boldsymbol{b}_o\boldsymbol{k}_o)]^{-1}\boldsymbol{b}_o.$$

The controllability Gramian can now be defined as

$$\boldsymbol{K}=\frac{1}{2\pi j}\oint_{|z|=1}\boldsymbol{F}(z)\boldsymbol{F}^*(z)\frac{dz}{z} \tag{15}$$

which can be obtained by solving the Lyapunov equation

$$\boldsymbol{K}=(\boldsymbol{A}_o-\boldsymbol{b}_o\boldsymbol{k}_o)\boldsymbol{K}(\boldsymbol{A}_o-\boldsymbol{b}_o\boldsymbol{k}_o)^T+\boldsymbol{b}_o\boldsymbol{b}_o^T. \tag{16}$$

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 $(\overline{\boldsymbol{R}}_o,\overline{\boldsymbol{b}}_o,\overline{\boldsymbol{g}}_0,\overline{\boldsymbol{k}}_o)_n$, where

$$\begin{aligned}\overline{\boldsymbol{R}}_o&=\boldsymbol{T}^{-1}\boldsymbol{R}_o\boldsymbol{T}, & \overline{\boldsymbol{b}}_o&=\boldsymbol{T}^{-1}\boldsymbol{b}_o\\ \overline{\boldsymbol{g}}_o&=\boldsymbol{T}^{-1}\boldsymbol{g}_o, & \overline{\boldsymbol{k}}_o&=\boldsymbol{k}_o\boldsymbol{T}.\end{aligned} \tag{17}$$

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

$$\overline{\boldsymbol{K}}=\boldsymbol{T}^{-1}\boldsymbol{K}\boldsymbol{T}^{-T},\quad \overline{\boldsymbol{W}}_i=\boldsymbol{T}^T\boldsymbol{W}_i\boldsymbol{T},\quad 1\le i\le 4. \tag{18}$$

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

$$(\overline{\boldsymbol{K}})_{ii}=(\boldsymbol{T}^{-1}\boldsymbol{K}\boldsymbol{T}^{-T})_{ii}=1\quad\text{for}\quad i=1,2,\cdots,n. \tag{19}$$

With an equivalent state-space realization as specified in (17) and assuming the use of high-order diagonal EF matrices, the normalized noise gain is found to be

$$\begin{aligned}J_2(\boldsymbol{D},\boldsymbol{T})=&\,\mathrm{tr}\Big[\boldsymbol{T}^T\boldsymbol{W}_4\boldsymbol{T}\Big]-2\sum_{i=0}^{N-1}\mathrm{tr}\Big[\boldsymbol{T}^T\boldsymbol{\Psi}_4(i+1,0)\boldsymbol{T}\boldsymbol{D}_i\Big]\\&+\sum_{i=0}^{N-1}\sum_{l=0}^{N-1}\mathrm{tr}\Big[\boldsymbol{T}^T\boldsymbol{\Psi}_4(i,l)\boldsymbol{T}\boldsymbol{D}_i\boldsymbol{D}_l\Big]\end{aligned} \tag{20}$$

The problem considered here is to jointly design the high-order EF matrices $\boldsymbol{D}_0,\boldsymbol{D}_1,\cdots,\boldsymbol{D}_{N-1}$ and coordinate transformation matrix $\boldsymbol{T}$ to minimize $J_2(\boldsymbol{D},\boldsymbol{T})$ in (20) subject to the l_2-scaling constraints in (19).

3. A JOINT OPTIMIZATION TECHNIQUE

We define

$$\hat{\boldsymbol{T}}=\boldsymbol{T}^T\boldsymbol{K}^{-\frac{1}{2}} \tag{21}$$

which leads (19) to

$$(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{T}}^{-1})_{ii}=1\quad\text{for}\quad i=1,2,\cdots,n. \tag{22}$$

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

$$\hat{\boldsymbol{T}}^{-1}=\left[\frac{\boldsymbol{t}_1}{||\boldsymbol{t}_1||},\frac{\boldsymbol{t}_2}{||\boldsymbol{t}_2||},\cdots,\frac{\boldsymbol{t}_n}{||\boldsymbol{t}_n||}\right]. \tag{23}$$

Substituting (21) into (20), we obtain

$$\begin{aligned}J_3(\boldsymbol{D},\hat{\boldsymbol{T}})=&\,\mathrm{tr}\Big[\hat{\boldsymbol{T}}\hat{\boldsymbol{W}}_4\hat{\boldsymbol{T}}^T\Big]-2\sum_{i=0}^{N-1}\mathrm{tr}\Big[\hat{\boldsymbol{T}}\hat{\boldsymbol{\Psi}}_4(i+1,0)\hat{\boldsymbol{T}}^T\boldsymbol{D}_i\Big]\\&+\sum_{i=0}^{N-1}\sum_{l=0}^{N-1}\mathrm{tr}\Big[\hat{\boldsymbol{T}}\hat{\boldsymbol{\Psi}}_4(i,l)\hat{\boldsymbol{T}}^T\boldsymbol{D}_i\boldsymbol{D}_l\Big]\end{aligned} \tag{24}$$

where

$$\hat{\boldsymbol{\Psi}}_4(i,l)=\boldsymbol{K}^{\frac{1}{2}}\boldsymbol{\Psi}_4(i,l)\boldsymbol{K}^{\frac{1}{2}},\qquad \hat{\boldsymbol{W}}_4=\boldsymbol{K}^{\frac{1}{2}}\boldsymbol{W}_4\boldsymbol{K}^{\frac{1}{2}}.$$

Following the foregoing arguments, the problem of obtaining matrices $\boldsymbol{T}$ and $\boldsymbol{D}_0,\boldsymbol{D}_1,\cdots,\boldsymbol{D}_{N-1}$ that jointly minimize (20) subject to the l_2-scaling constraints in (19) is now converted into an unconstrained optimization problem of obtaining $\hat{\boldsymbol{T}}$ and $\boldsymbol{D}_0,\boldsymbol{D}_1,\cdots,\boldsymbol{D}_{N-1}$ that jointly minimize (24).

Let $\boldsymbol{x}$ be the column vector that collects the variables in matrices $[\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_n]$ and $\boldsymbol{D}_0,\boldsymbol{D}_1,\cdots,\boldsymbol{D}_{N-1}$. Then $J_3(\boldsymbol{D},\hat{\boldsymbol{T}})$ in (24) is a function of $\boldsymbol{x}$, denoted by $J(\boldsymbol{x})$. The proposed algorithm starts with an initial point $\boldsymbol{x}_0$ obtained from the assignment $\hat{\boldsymbol{T}}=\boldsymbol{D}_0=\boldsymbol{D}_1=\cdots=\boldsymbol{D}_{N-1}=\boldsymbol{I}_n$. In the kth iteration, a quasi-Newton algorithm updates the most recent point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as [13]

$$\boldsymbol{x}_{k+1}=\boldsymbol{x}_k+\alpha_k\boldsymbol{d}_k \tag{25}$$

where

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

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

$$|J(\boldsymbol{x}_{k+1})-J(\boldsymbol{x}_k)|<\varepsilon \tag{26}$$

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

4. AN ILLUSTRATIVE EXAMPLE

Consider a linear discrete-time system specified by

$$\boldsymbol{A}_o=\begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ 0.339377 & -1.152652 & 1.520167\end{bmatrix},\quad \boldsymbol{b}_o=\begin{bmatrix}0\\0\\1\end{bmatrix}$$

$$\boldsymbol{c}_o=\begin{bmatrix}0.093253 & 0.128620 & 0.314713\end{bmatrix}.$$

Suppose that the full-order observer has the poles at $z=0.1532$, 0.2861 and 0.1137, and that the poles of the regulator are placed at $z=0.5067$, 0.6023 and 0.4331. Then the coefficient vector of the full-order observer and the gain vector of the state feedback were computed as

$$\begin{aligned}\boldsymbol{g}_o&=\begin{bmatrix}-0.006436 & 3.683651 & 5.083920\end{bmatrix}^T\\ \boldsymbol{k}_o&=\begin{bmatrix}0.471552 & -0.367158 & 3.062267\end{bmatrix}.\end{aligned}$$

The controllability and observability Gramians $\boldsymbol{K}$ and $\boldsymbol{W}_4$ relating to the closed-loop system in (6) were computed from (16) and (11) as

$$\boldsymbol{K}=\begin{bmatrix}10.140736 & -9.324397 & 7.646117\\ -9.324397 & 10.140736 & -9.324397\\ 7.646117 & -9.324397 & 10.140736\end{bmatrix}$$

$$\boldsymbol{W}_4=10\begin{bmatrix}0.946720 & 2.004689 & 1.053033\\ 2.004689 & 4.432667 & 2.355278\\ 1.053033 & 2.355278 & 1.389340\end{bmatrix}.$$

When performing l_2-scaling to the system with

$$\boldsymbol{T}_o = \text{diag}\{3.184452, 3.184452, 3.184452\}$$

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 (11) as

$$J_1(\mathbf{0}) = \text{tr}\left[\boldsymbol{T}_o^T \boldsymbol{W}_4 \boldsymbol{T}_o\right] = 686.398739.$$

In the case where $N = 1$ and matrix $\boldsymbol{D}_0$ is diagonal, the quasi-Newton algorithm was applied to minimize (24) with tolerance $\varepsilon = 10^{-8}$ in (26). It took the algorithm 19 iterations to converge to the solution

$$\hat{\boldsymbol{T}} = \begin{bmatrix} 1.294114 & -0.168488 & 0.007707 \\ 0.371342 & 0.867987 & 0.470162 \\ -0.779773 & -0.169966 & 0.981896 \end{bmatrix}$$

$$\boldsymbol{D}_0 = \text{diag}\{\, 0.061401 \quad -0.954205 \quad -0.615047 \,\}$$

with $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}}) = 12.705365$.

In the case where $N = 2$ and matrices $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$ are diagonal, the quasi-Newton algorithm was applied to minimize (24) with tolerance $\varepsilon = 10^{-8}$ in (26). It took the algorithm 24 iterations to converge to the solution

$$\hat{\boldsymbol{T}} = \begin{bmatrix} 1.305265 & -0.172325 & 0.015749 \\ 0.360830 & 0.851958 & 0.512722 \\ -0.806130 & -0.188264 & 0.975634 \end{bmatrix}$$

$$\boldsymbol{D}_0 = \text{diag}\{\, 0.06453466 - 1.761331 \quad -0.650203 \,\}$$

$$\boldsymbol{D}_1 = \text{diag}\{\, -0.027322 \quad -0.837318 \quad -0.053387 \,\}$$

with $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}}) = 11.703374$.

In the case where $N = 3$ and matrices $\boldsymbol{D}_0$, $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are diagonal, the quasi-Newton algorithm was applied to minimize (24) with tolerance $\varepsilon = 10^{-8}$ in (26). It took the algorithm 105 iterations to converge to the solution

$$\hat{\boldsymbol{T}} = \begin{bmatrix} 0.327718 & 0.245254 & 1.713434 \\ 0.016465 & 1.176168 & 1.222032 \\ -0.843787 & 0.103998 & 1.016202 \end{bmatrix}$$

$$\boldsymbol{D}_0 = \text{diag}\{\, 2.662090 \quad -3.970640 \quad 1.436650 \,\}$$

$$\boldsymbol{D}_1 = \text{diag}\{\, -3.288568 \quad -5.584430 \quad 8.055175 \,\}$$

$$\boldsymbol{D}_2 = \text{diag}\{\, -9.873059 \quad -2.751501 \quad 7.632869 \,\}$$

with $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}}) = 2.509133$.

The other results regarding the normalized noise gain $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}})$ in (24) were summarized with those achieved by the separate optimization method of [8] in Table I where "Infinite Precision" shows the value of $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}})$ derived from the optimal $\hat{\boldsymbol{T}}$ and $\boldsymbol{D}$, and "3-Bit Quantization (Integer Quantization)" means the value of $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}})$ where each entry of the optimal $\boldsymbol{D}$ was rounded to a power-of-two representation with 3 bits after the binary point (an integer) and then matrix $\hat{\boldsymbol{T}}$ was updated to minimize $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}})$ with such a $\boldsymbol{D}$ fixed.

TABLE I
PERFORMANCE COMPARISON

N	Optimization	Infinite Precision	3-Bit Quantization	Integer Quantization
1	Separate	14.742567	14.807933	15.337683
	Joint	12.705365	12.769499	13.214089
2	Separate	14.097350	14.147798	15.028691
	Joint	11.703374	11.788775	12.558440
3	Separate	12.798210	12.892739	14.552876
	Joint	2.509133	2.552543	6.100339

5. CONCLUSION

The joint optimization problem of high-order EF and state-space realization to minimize the effects of roundoff noise in the closed-loop system with a state-estimate feedback controller subject to l_2-scaling constraints has been investigated. The probelm at hand has been 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. The proposed technique has been applied to the case where high-order EF matrices are diagonal. The validity and effectiveness of the proposed technique has been demonstrated by a numerical example.

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, 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, pp. 429-437, May 1984.

[3] P. P. Vaidyanathan, "On error-spectrum shaping in state-space digital filters," *IEEE Trans. Circuits Syst.*, vol. 32, 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, 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, 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, pp. 551-562, Sept. 1976.

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

[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," *IEEE Trans. Circuits Syst. I*, vol. 50, 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, pp. 2135-2145, June 2005.

[10] T. Hinamoto, A. Doi and W.-S. Lu, "Jointly optimal high-order error feedback and realization for roundoff noise minimization in 1-D and 2-D state-space digital filters," *IEEE Trans. Signal Process.*, vol. 61, pp. 5893-5904, Dec. 2013.

[11] G. Li and M. Gevers, "Optimal finite precision implementation of a state-estimate feedback controller," *IEEE Trans. Circuits Syst.*, vol. 37, pp. 1487-1498, Dec. 1990.

[12] T. Hinamoto, K. Kawai, M. Nakamoto and W.-S. Lu, "Jointly optimized error-feedback and realization for roundoff noise minimization in state-estimate feedback digital controllers," in *Proc. 16th European Signal Processing Conference (EUSIPCO'08)*, CD-ROM, Aug. 2008.

[13] R. Fletcher, *Practical Methods of Optimization*, 2nd ed. New York, NY, USA: Wiley, 1987.