JOINTLY OPTIMIZED ERROR-FEEDBACK AND REALIZATION FOR ROUNDOFF NOISE MINIMIZATION IN STATE-ESTIMATE FEEDBACK DIGITAL CONTROLLERS

Takao Hinamoto, Keijiro Kawai and Masayoshi Nakamoto

Graduate School of Engineering, Hiroshima University
Higashi-Hiroshima 739-8527, Japan
phone:+81-82-424-7672, fax:+81-82-422-7195
email:{hinamoto, msy}@hiroshima-u.ac.jp

Wu-Sheng Lu

Dept. of Elect. & Comput. Engineering,
University of Victoria
Victoria, B.C., V8W 3P6, Canada
phone:+1-250-721-8692, fax:+1-250-721-6052
email: wslu@ece.uvic.ca

ABSTRACT

The joint optimization problem of error-feedback and realization for the closed-loop system with a state-estimate feedback digital controller is investigated where the main objective is to minimize the effects of roundoff noise at the closed-loop system output subject to l_2-norm dynamic-range scaling constraints. It is shown that the problem can be 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. Analytical details are given as to how the proposed technique can be applied to the cases where the error-feedback matrix is a general, diagonal, or scalar matrix. A numerical example is presented to illustrate the utility of the proposed technique.

1. INTRODUCTION

Due to the finite precision nature of computer arithmetic, the output roundoff noise of a fixed-point IIR digital filter usually arises. This noise is critically dependent on the internal structure of an IIR digital filter [1],[2]. Error feedback (EF) is known as an effective technique for reducing the output roundoff noise in an IIR digital filter [3]-[5]. Williamson [6] has reduced the output roundoff noise more effectively by choosing the filter structure and applying EF to the filter. Lu and Hinamoto [7] have developed a jointly optimized technique of EF and realization to minimize the effects of roundoff noise at the filter output subject to l_2-scaling constraints. Li and Gevers [8] have analyzed the output roundoff noise of the closed-loop system with a state-estimate feedback controller, and presented an algorithm for realizing the state-estimate feedback controller with minimum output roundoff noise under l_2-norm dynamic-range scaling constraints. Hinamoto and Yamamoto [9] have proposed a method for applying EF to a given closed-loop system with a state-estimate feedback controller.

This paper investigates the problem of jointly optimizing EF and realization for the closed-loop system with a state-estimate feedback controller so as to minimize the output roundoff noise subject to l_2-norm dynamic-range scaling constraints. To this end, an iterative technique which relies on an efficient quasi-Newton algorithm [10] is developed. Our computer simulation results demonstrate the validity and effectiveness of the proposed technique.

2. ROUNDOFF NOISE ANALYSIS

Let a stable, controllable and observable linear discrete-time system be 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

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

If a regulator is designed by using the full-order state observer, we obtain a state-estimate feedback controller as

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

where $\tilde{\boldsymbol{x}}(k)$ is an $n\times 1$ state-variable vector in the full-order state observer, $\boldsymbol{g}_o$ is an $n\times 1$ gain vector chosen so that all the eigenvalues of $\boldsymbol{F}_o = \boldsymbol{A}_o - \boldsymbol{g}_o\boldsymbol{c}_o$ are inside the unit circle in the complex plane, $\boldsymbol{k}_o$ is a $1\times n$ state-feedback gain vector chosen so that each of the eigenvalues of $\boldsymbol{A}_o - \boldsymbol{b}_o\boldsymbol{k}_o$ is at a desirable location within the unit circle, $r(k)$ is a scalar reference signal, and $\boldsymbol{R}_o = \boldsymbol{F}_o - \boldsymbol{b}_o\boldsymbol{k}_o$.

Performing quantization before matrix-vector multiplication, we can express the finite-word-length (FWL) implementation of (3) with error feedback as

$$\begin{aligned} \hat{\boldsymbol{x}}(k+1) &= \boldsymbol{R}\,\boldsymbol{Q}[\hat{\boldsymbol{x}}(k)] + \boldsymbol{b}\,r(k) + \boldsymbol{g}\,y(k) + \boldsymbol{D}\,\boldsymbol{e}(k) \\ u(k) &= -\boldsymbol{k}\,\boldsymbol{Q}[\hat{\boldsymbol{x}}(k)] + r(k) \end{aligned} \tag{4}$$

where $\boldsymbol{e}(k) = \hat{\boldsymbol{x}}(k) - \boldsymbol{Q}[\hat{\boldsymbol{x}}(k)]$ and $\boldsymbol{D}$ is an $n\times n$ error feedback matrix. All coefficient matrices $\boldsymbol{R}$, $\boldsymbol{b}$, $\boldsymbol{g}$ and $\boldsymbol{k}$

are assumed to have an exact fractional B_c bit representation. The FWL state-variable vector $\hat{\boldsymbol{x}}(k)$ and signal $u(k)$ all have a B bit fractional representation, while the reference input $r(k)$ is a $(B-B_c)$ bit fraction. The vector quantizer $\boldsymbol{Q}[\cdot]$ in (4) rounds the B bit fraction $\tilde{\boldsymbol{x}}(k)$ to $(B-B_c)$ bits after completing 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 zero-mean noise process with covariance $\sigma^2 \boldsymbol{I}_n$.

The closed-loop control system consisting of the linear system in (1) and the state-estimate feedback controller in (4) is illustrated in Fig. 1. This closed-loop

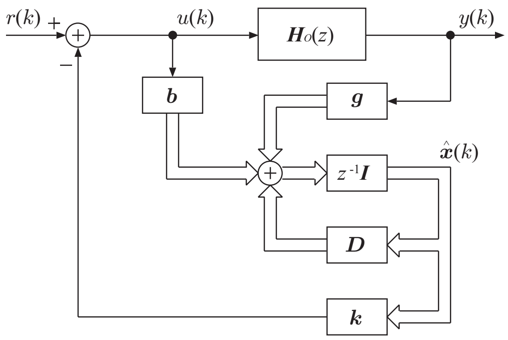

Fig. 1. The closed-loop control system with a state-estimate feedback controller.

system is described by

$$\begin{bmatrix} \boldsymbol{x}(k+1) \\ \hat{\boldsymbol{x}}(k+1) \end{bmatrix} = \overline{\boldsymbol{A}} \begin{bmatrix} \boldsymbol{x}(k) \\ \hat{\boldsymbol{x}}(k) \end{bmatrix} + \overline{\boldsymbol{b}} r(k) + \overline{\boldsymbol{B}} \boldsymbol{e}(k)$$
$$y(k) = \overline{\boldsymbol{c}} \begin{bmatrix} \boldsymbol{x}(k) \\ \hat{\boldsymbol{x}}(k) \end{bmatrix} \tag{5}$$

where

$$\overline{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{A}_o & -\boldsymbol{b}_o \boldsymbol{k} \\ \boldsymbol{g}\boldsymbol{c}_o & \boldsymbol{R} \end{bmatrix}, \qquad \overline{\boldsymbol{b}} = \begin{bmatrix} \boldsymbol{b}_o \\ \boldsymbol{b} \end{bmatrix}$$
$$\overline{\boldsymbol{B}} = \begin{bmatrix} \boldsymbol{b}_o \boldsymbol{k} \\ \boldsymbol{D} - \boldsymbol{R} \end{bmatrix}, \qquad \overline{\boldsymbol{c}} = [\boldsymbol{c}_o \quad \boldsymbol{0}].$$

Let the transfer function from the roundoff noise $\boldsymbol{e}(k)$ to the output $y(k)$ in (5) be defined by $\boldsymbol{G}_D(z)$. Then

$$\boldsymbol{G}_D(z) = \overline{\boldsymbol{c}}\,(z\boldsymbol{I}_{2n} - \overline{\boldsymbol{A}})^{-1} \overline{\boldsymbol{B}}. \tag{6}$$

The noise gain $J(\boldsymbol{D}) = \sigma_{out}^2/\sigma^2$ is then computed as

$$J(\boldsymbol{D}) = \mathrm{tr}[\boldsymbol{W}_D] \tag{7}$$

with

$$\boldsymbol{W}_D = \frac{1}{2\pi j} \oint_{|z|=1} \boldsymbol{G}_D^*(z) \boldsymbol{G}_D(z) \frac{dz}{z} \tag{8}$$

where σ_{out}^2 stands for the noise variance of the output. For tractability, we evaluate $J(\boldsymbol{D})$ in (7) by replacing $\boldsymbol{R}$, $\boldsymbol{b}$, $\boldsymbol{g}$ and $\boldsymbol{k}$ by $\boldsymbol{R}_o$, $\boldsymbol{b}_o$, $\boldsymbol{g}_o$ and $\boldsymbol{k}_o$, respectively. Define

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

the transfer function $\boldsymbol{G}_D(z)$ in (6) can be expressed as

$$\begin{aligned} \boldsymbol{G}_D(z) &= \overline{\boldsymbol{c}} \boldsymbol{S} (z\boldsymbol{I}_{2n} - \boldsymbol{S}^{-1} \overline{\boldsymbol{A}} \boldsymbol{S})^{-1} \boldsymbol{S}^{-1} \overline{\boldsymbol{B}} \\ &= \overline{\boldsymbol{c}} (z\boldsymbol{I}_{2n} - \Phi)^{-1} \begin{bmatrix} \boldsymbol{b}_o \boldsymbol{k}_o \\ \boldsymbol{F}_o - \boldsymbol{D} \end{bmatrix} \\ &= \boldsymbol{c}_o (z\boldsymbol{I}_n - \boldsymbol{A}_o + \boldsymbol{b}_o \boldsymbol{k}_o)^{-1} \boldsymbol{b}_o \boldsymbol{k}_o (z\boldsymbol{I}_n - \boldsymbol{F}_o)^{-1} \\ &\quad \cdot (z\boldsymbol{I}_n - \boldsymbol{D}) \\ &= \overline{\boldsymbol{c}} (z\boldsymbol{I}_{2n} - \Phi)^{-1} \boldsymbol{U} (z\boldsymbol{I}_n - \boldsymbol{D}) \end{aligned} \tag{10}$$

where

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

It is noted that the stability of the closed-loop control system is determined by the eigenvalues of matrix $\overline{\boldsymbol{A}}$ in (5), or equivalently, those of matrix Φ in (10). This means that neither of the quantization error $\boldsymbol{e}(k)$ and the error-feedback matrix $\boldsymbol{D}$ affects the stability.

Substituting (10) into matrix $\boldsymbol{W}_D$ in (8) gives

$$\begin{aligned} \boldsymbol{W}_D = {} & (\boldsymbol{b}_0 \boldsymbol{k}_0)^T \boldsymbol{W}_1 \boldsymbol{b}_0 \boldsymbol{k}_0 + (\boldsymbol{b}_0 \boldsymbol{k}_0)^T \boldsymbol{W}_2 (\boldsymbol{F}_0 - \boldsymbol{D}) \\ & + (\boldsymbol{F}_0 - \boldsymbol{D})^T \boldsymbol{W}_3 \boldsymbol{b}_0 \boldsymbol{k}_0 \\ & + (\boldsymbol{F}_0 - \boldsymbol{D})^T \boldsymbol{W}_4 (\boldsymbol{F}_0 - \boldsymbol{D}) \end{aligned} \tag{11}$$

where

$$\boldsymbol{W} = \Phi^T \boldsymbol{W} \Phi + \overline{\boldsymbol{c}}^T \overline{\boldsymbol{c}}, \qquad \boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{W}_2 \\ \boldsymbol{W}_3 & \boldsymbol{W}_4 \end{bmatrix}.$$

Since $\boldsymbol{W}$ is positive semidefinite, it can be shown that there exists an $n \times n$ matrix $\boldsymbol{P}$ such that $\boldsymbol{W}_3 = \boldsymbol{W}_4 \boldsymbol{P}$. In addition, (11) can be written by virtue of $\boldsymbol{W}_2 = \boldsymbol{W}_3^T$ as

$$\begin{aligned} \boldsymbol{W}_D = {} & (\boldsymbol{F}_0 + \boldsymbol{P}\boldsymbol{b}_0 \boldsymbol{k}_0 - \boldsymbol{D})^T \boldsymbol{W}_4 (\boldsymbol{F}_0 + \boldsymbol{P}\boldsymbol{b}_0 \boldsymbol{k}_0 - \boldsymbol{D}) \\ & + (\boldsymbol{b}_0 \boldsymbol{k}_0)^T (\boldsymbol{W}_1 - \boldsymbol{P}^T \boldsymbol{W}_4 \boldsymbol{P}) \boldsymbol{b}_0 \boldsymbol{k}_0. \end{aligned} \tag{12}$$

Alternatively, applying z-transform to the first equation in (5) under the assumption that $\boldsymbol{e}(k) = \boldsymbol{0}$, we obtain

$$\begin{bmatrix} \boldsymbol{X}(z) \\ \hat{\boldsymbol{X}}(z) \end{bmatrix} = (z\boldsymbol{I} - \overline{\boldsymbol{A}})^{-1} \overline{\boldsymbol{b}} R(z) \tag{13}$$

where $\boldsymbol{X}(z)$, $\hat{\boldsymbol{X}}(z)$ and $R(z)$ represent the z-transforms of $\boldsymbol{x}(k)$, $\hat{\boldsymbol{x}}(k)$ and $r(k)$, respectively. Replacing $\boldsymbol{R}$, $\boldsymbol{b}$, $\boldsymbol{k}$ and $\boldsymbol{g}$ by $\boldsymbol{R}_o$, $\boldsymbol{b}_o$, $\boldsymbol{k}_o$ and $\boldsymbol{g}_o$, respectively, and then using

$$\boldsymbol{S}^{-1} \begin{bmatrix} \boldsymbol{X}(z) \\ \hat{\boldsymbol{X}}(z) \end{bmatrix} = (z\boldsymbol{I}_{2n} - \boldsymbol{S}^{-1} \overline{\boldsymbol{A}} \boldsymbol{S})^{-1} \boldsymbol{S}^{-1} \overline{\boldsymbol{b}}$$

yield

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

where

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

The controllability Gramian $\boldsymbol{K}$ defined by

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

can be obtained by solving the Lyapunov equation

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

3. ROUNDOFF NOISE MINIMIZATION

Consider the system in (4) with $\boldsymbol{D} = \boldsymbol{0}$ and denote it by $(\boldsymbol{R}, \boldsymbol{b}, \boldsymbol{g}, \boldsymbol{k})_n$. By applying a coordinate transformation $\tilde{\boldsymbol{x}}'(k) = \boldsymbol{T}^{-1}\hat{\boldsymbol{x}}(k)$ to the above system $(\boldsymbol{R}, \boldsymbol{b}, \boldsymbol{g}, \boldsymbol{k})_n$, we obtain a new realization characterized by $(\tilde{\boldsymbol{R}}, \tilde{\boldsymbol{b}}, \tilde{\boldsymbol{g}}, \tilde{\boldsymbol{k}})_n$ where

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

For the system in (17), the counterparts of $\boldsymbol{W}_i$ for $i = 1, 2, 3, 4$ are given by

$$\tilde{\boldsymbol{W}}_i = \boldsymbol{T}^T\boldsymbol{W}_i\boldsymbol{T} \tag{18}$$

and the corresponding noise gain is given by

$$J(\boldsymbol{D}, \boldsymbol{T}) = \mathrm{tr}[\tilde{\boldsymbol{W}}_D] \tag{19}$$

where $\tilde{\boldsymbol{W}}_D$ can be obtained using (11) as

$$\begin{aligned} \tilde{\boldsymbol{W}}_D = &\left[\boldsymbol{T}^{-1}(\boldsymbol{F}_0 + \boldsymbol{P}\boldsymbol{b}_0\boldsymbol{k}_0)\boldsymbol{T} - \boldsymbol{D}\right]^T \\ &\cdot\boldsymbol{T}^T\boldsymbol{W}_4\boldsymbol{T}\left[\boldsymbol{T}^{-1}(\boldsymbol{F}_0 + \boldsymbol{P}\boldsymbol{b}_0\boldsymbol{k}_0)\boldsymbol{T} - \boldsymbol{D}\right] \\ &+\boldsymbol{T}^T(\boldsymbol{b}_0\boldsymbol{k}_0)^T(\boldsymbol{W}_1 - \boldsymbol{P}^T\boldsymbol{W}_4\boldsymbol{P})\boldsymbol{b}_0\boldsymbol{k}_0\boldsymbol{T}. \end{aligned}$$

In addition, (15) can be written as

$$\tilde{\boldsymbol{K}} = \boldsymbol{T}^{-1}\boldsymbol{K}\boldsymbol{T}^{-T}. \tag{20}$$

As a result, the output roundoff noise minimization problem amounts to obtaining matrices $\boldsymbol{D}$ and $\boldsymbol{T}$ which jointly minimize $J(\boldsymbol{D}, \boldsymbol{T})$ in (19) subject to the l_2-scaling constraints specified by

$$(\tilde{\boldsymbol{K}})_{ii} = (\boldsymbol{T}^{-1}\boldsymbol{K}\boldsymbol{T}^{-T})_{ii} = 1, \quad i = 1, 2, \cdots, n. \tag{21}$$

To deal with (21), we define

$$\hat{\boldsymbol{T}} = \boldsymbol{T}^T\boldsymbol{K}^{-\frac{1}{2}}. \tag{22}$$

Then the l_2-scaling constraints in (21) can be written as

$$(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{T}}^{-1})_{ii} = 1, \quad i = 1, 2, \cdots, n. \tag{23}$$

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

$$\hat{\boldsymbol{T}}^{-1} = \left[\frac{\boldsymbol{t}_1}{||\boldsymbol{t}_1||}, \frac{\boldsymbol{t}_2}{||\boldsymbol{t}_2||}, \cdots, \frac{\boldsymbol{t}_n}{||\boldsymbol{t}_n||}\right]. \tag{24}$$

Substituting (22) into (19), we obtain

$$\begin{aligned} J(\boldsymbol{D}, \hat{\boldsymbol{T}}) = \mathrm{tr}\Big[&\hat{\boldsymbol{T}}(\hat{\boldsymbol{A}} - \hat{\boldsymbol{T}}^T\boldsymbol{D}\hat{\boldsymbol{T}}^{-T})^T\hat{\boldsymbol{W}}_4 \\ &\cdot(\hat{\boldsymbol{A}} - \hat{\boldsymbol{T}}^T\boldsymbol{D}\hat{\boldsymbol{T}}^{-T})\hat{\boldsymbol{T}}^T + \hat{\boldsymbol{T}}\hat{\boldsymbol{C}}\hat{\boldsymbol{T}}^T\Big] \end{aligned} \tag{25}$$

where

$$\begin{aligned} \hat{\boldsymbol{A}} &= \boldsymbol{K}^{-\frac{1}{2}}(\boldsymbol{F}_0 + \boldsymbol{P}\boldsymbol{b}_0\boldsymbol{k}_0)\boldsymbol{K}^{\frac{1}{2}}, \quad \hat{\boldsymbol{W}}_4 = \boldsymbol{K}^{\frac{1}{2}}\boldsymbol{W}_4\boldsymbol{K}^{\frac{1}{2}} \\ \hat{\boldsymbol{C}} &= \boldsymbol{K}^{\frac{1}{2}}(\boldsymbol{b}_0\boldsymbol{k}_0)^T(\boldsymbol{W}_1 - \boldsymbol{P}^T\boldsymbol{W}_4\boldsymbol{P})\boldsymbol{b}_0\boldsymbol{k}_0\boldsymbol{K}^{\frac{1}{2}}. \end{aligned}$$

From the foregoing arguments, the problem of obtaining matrices $\boldsymbol{D}$ and $\boldsymbol{T}$ that minimize (19) subject to the scaling constraints in (21) is now converted into an unconstrained optimization problem of obtaining $\boldsymbol{D}$ and $\hat{\boldsymbol{T}}$ that jointly minimize $J(\boldsymbol{D}, \hat{\boldsymbol{T}})$ in (25).

Let $\boldsymbol{x}$ be the column vector that collects the variables in matrices $\boldsymbol{D}$ and $[\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n]$. Then $J(\boldsymbol{D}, \hat{\boldsymbol{T}})$ is a function of $\boldsymbol{x}$, denoted by $J(\boldsymbol{x})$. The proposed algorithm starts with an initial point $\boldsymbol{x}_0$ obtained from an initial assignment $\boldsymbol{D} = \hat{\boldsymbol{T}} = \boldsymbol{I}_n$. In the kth iteration, a quasi-Newton algorithm updates the most recent point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as [10]

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{d}_k, \tag{26}$$

where

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

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

$$|J(\boldsymbol{x}_{k+1}) - J(\boldsymbol{x}_k)| < \varepsilon \tag{27}$$

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

In what follows, we derive closed-form expressions of $\nabla J(\boldsymbol{x})$ for the cases where $\boldsymbol{D}$ assumes the form of a general, diagonal, or scalar matrix.

1) Case 1: $\boldsymbol{D}$ Is a General Matrix: From (25), the optimal choice of $\boldsymbol{D}$ is given by

$$\boldsymbol{D} = \hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{A}}\hat{\boldsymbol{T}}^T, \tag{28}$$

which leads to

$$J(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{A}}\hat{\boldsymbol{T}}^T, \hat{\boldsymbol{T}}) = \mathrm{tr}\left[\hat{\boldsymbol{T}}\hat{\boldsymbol{C}}\hat{\boldsymbol{T}}^T\right]. \tag{29}$$

In this case, the number of elements in vector $\boldsymbol{x}$ consisting of $\hat{\boldsymbol{T}}$ is equal to n^2 and the gradient of $J(\boldsymbol{x})$ is found to be

$$\begin{aligned} \frac{\partial J(\boldsymbol{x})}{\partial t_{ij}} &= \lim_{\Delta \to 0}\frac{J(\hat{\boldsymbol{T}}_{ij}) - J(\hat{\boldsymbol{T}})}{\Delta} \\ &= 2\boldsymbol{e}_j^T\hat{\boldsymbol{T}}\hat{\boldsymbol{C}}\hat{\boldsymbol{T}}^T\hat{\boldsymbol{T}}\boldsymbol{g}_{ij}, \qquad i, j = 1, 2, \cdots, n \end{aligned} \tag{30}$$

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

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

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

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

2) Case 2: $\boldsymbol{D}$ Is a Diagonal Matrix: Here, matrix $\boldsymbol{D}$ assumes the form

$$\boldsymbol{D} = \mathrm{diag}\{d_1, d_2, \cdots, d_n\}. \tag{31}$$

In this case, (25) becomes

$$J(\boldsymbol{D}, \hat{\boldsymbol{T}}) = \mathrm{tr}\left[\hat{\boldsymbol{T}} \boldsymbol{M}_d \hat{\boldsymbol{T}}^T\right] \tag{32}$$

where

$$\begin{aligned} \boldsymbol{M}_d = {} & \hat{\boldsymbol{C}} + \hat{\boldsymbol{A}}^T \hat{\boldsymbol{W}}_4 \hat{\boldsymbol{A}} + \hat{\boldsymbol{W}}_4 \hat{\boldsymbol{T}}^T \boldsymbol{D}^2 \hat{\boldsymbol{T}}^{-T} \\ & - \hat{\boldsymbol{A}}^T \hat{\boldsymbol{W}}_4 \hat{\boldsymbol{T}}^T \boldsymbol{D} \hat{\boldsymbol{T}}^{-T} - \hat{\boldsymbol{W}}_4 \hat{\boldsymbol{A}} \hat{\boldsymbol{T}}^T \boldsymbol{D} \hat{\boldsymbol{T}}^{-T}. \end{aligned}$$

It follows that

$$\begin{aligned} \frac{\partial J(\boldsymbol{x})}{\partial t_{ij}} &= 2 \boldsymbol{e}_j^T \hat{\boldsymbol{T}} \boldsymbol{M}_d \hat{\boldsymbol{T}}^T \hat{\boldsymbol{T}} \boldsymbol{g}_{ij}, \quad i,j = 1,2,\cdots,n \\ \frac{\partial J(\boldsymbol{x})}{\partial d_i} &= 2 \boldsymbol{e}_i^T (\boldsymbol{D} \hat{\boldsymbol{T}} - \hat{\boldsymbol{T}} \hat{\boldsymbol{A}}^T) \hat{\boldsymbol{W}}_4 \hat{\boldsymbol{T}}^T \boldsymbol{e}_i, \quad i = 1,2,\cdots,n. \end{aligned} \tag{33}$$

3) Case 3: $\boldsymbol{D}$ Is a Scalar Matrix: It is assumed here that $\boldsymbol{D} = \alpha \boldsymbol{I}_n$ with a scalar α. The gradient of $J(\boldsymbol{x})$ can then be calculated as

$$\begin{aligned} \frac{\partial J(\boldsymbol{x})}{\partial t_{ij}} &= 2 \boldsymbol{e}_j^T \hat{\boldsymbol{T}} \boldsymbol{M}_s \hat{\boldsymbol{T}}^T \hat{\boldsymbol{T}} \boldsymbol{g}_{ij}, \quad i,j = 1,2,\cdots,n \\ \frac{\partial J(\boldsymbol{x})}{\partial \alpha} &= \mathrm{tr}\left[\hat{\boldsymbol{T}} (2\alpha \hat{\boldsymbol{W}}_4 - \hat{\boldsymbol{A}}^T \hat{\boldsymbol{W}}_4 - \hat{\boldsymbol{W}}_4 \hat{\boldsymbol{A}}) \hat{\boldsymbol{T}}^T\right] \end{aligned} \tag{34}$$

where

$$\boldsymbol{M}_s = (\hat{\boldsymbol{A}} - \alpha \boldsymbol{I}_n)^T \hat{\boldsymbol{W}}_4 (\hat{\boldsymbol{A}} - \alpha \boldsymbol{I}_n) + \hat{\boldsymbol{C}}.$$

4. A NUMERICAL EXAMPLE

In this section we illustrate the proposed method by considering a linear discrete-time system specified by

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

$$\boldsymbol{c}_o = [\ 0.093253 \quad 0.128620 \quad 0.314713\ ].$$

Suppose that the poles of the observer and regulator in the system are required to be located at $z = 0.1532$, 0.2861, 0.1137, and $z = 0.5067$, 0.6023, 0.4331, respectively. This can be achieved by choosing

$$\boldsymbol{k}_o = [\ 0.471552 \quad -0.367158 \quad 3.062267\ ]$$

$$\boldsymbol{g}_o = [\ -0.006436 \quad 3.683651 \quad 5.083920\ ]^T.$$

Performing the l_2-scaling to the state-estimate feedback controller, we obtain $J(\boldsymbol{0}) = 686.4121$ in (7) where $\boldsymbol{D} = \boldsymbol{0}$. Next, the controller is transformed into the optimal realization that minimizes $J(\boldsymbol{0})$ in (7) under the l_2-scaling constraints. This leads to $J_{min}(\boldsymbol{0}) = 28.6187$. Finally, EF and state-variable coordinate transformation are applied to the above optimal realization so as to jointly minimize the output roundoff noise. The profiles of $J(\boldsymbol{x})$ during the first 20 iteration for the cases of $\boldsymbol{D}$ being a general, diagonal, and scalar matrix are depicted in Fig. 2.

1) Case 1: $\boldsymbol{D}$ Is a General Matrix: The quasi-Newton algorithm was applied to minimize (25). It took the algorithm 20 iterations to converge to the solution

$$\boldsymbol{D} = \begin{bmatrix} 0.211191 & -3.078211 & -3.344596 \\ -1.321589 & 1.897308 & 3.243515 \\ 1.917916 & -1.890027 & -3.807473 \end{bmatrix}$$

$$\boldsymbol{T} = \begin{bmatrix} -11.039974 & -43.683697 & -30.131793 \\ -3.231505 & 8.919473 & 9.118205 \\ 2.620911 & 6.462685 & 7.032260 \end{bmatrix}$$

and the minimized noise gain was found to be $J(\boldsymbol{D}, \hat{\boldsymbol{T}}) = 4.8823$. Next, the above optimal EF matrix $\boldsymbol{D}$ was rounded to a power-of-two representation with 3 bits after the binary point, which resulted in

$$\boldsymbol{D}_{3bit} = \begin{bmatrix} 0.250 & -3.125 & -3.375 \\ -1.375 & 1.875 & 3.250 \\ 1.875 & -1.875 & -3.750 \end{bmatrix}$$

and a noise gain $J(\boldsymbol{D}_{3bit}, \hat{\boldsymbol{T}}) = 23.4873$. Furthermore, when the optimal EF matrix $\boldsymbol{D}$ was rounded to the integer representation

$$\boldsymbol{D}_{int} = \begin{bmatrix} 0 & -3 & -3 \\ -1 & 2 & 3 \\ 2 & -2 & -4 \end{bmatrix},$$

the noise gain was found to be $J(\boldsymbol{D}_{int}, \hat{\boldsymbol{T}}) = 293.0187$.

2) Case 2: $\boldsymbol{D}$ Is a Diagonal Matrix: Again, the quasi-Newton algorithm was applied to minimize $J(\boldsymbol{D}, \hat{\boldsymbol{T}})$ in (25) for a diagonal EF matrix $\boldsymbol{D}$. It took the algorithm 20 iterations to converge to the solution

$$\boldsymbol{D} = \mathrm{diag}\{0.050638, -0.608845, -0.951572\}$$

$$\boldsymbol{T} = \begin{bmatrix} 3.588878 & 0.735966 & 0.010417 \\ -2.457241 & 0.728171 & 0.556762 \\ 1.514232 & -2.058856 & 0.142204 \end{bmatrix}$$

and the minimized noise gain was found to be $J(\boldsymbol{D}, \hat{\boldsymbol{T}}) = 12.7097$. Next, the above optimal diagonal EF matrix $\boldsymbol{D}$ was rounded to a power-of-two representation with 3 bits after the binary point to yield $\boldsymbol{D}_{3bit} = \mathrm{diag}\{0.000, -0.625, -1.000\}$, which leads to a noise gain $J(\boldsymbol{D}_{3bit}, \hat{\boldsymbol{T}}) = 12.7722$. Furthermore, when the optimized diagonal EF matrix $\boldsymbol{D}$ was rounded to the integer representation $\boldsymbol{D}_{int} = \mathrm{diag}\{0, -1, -1\}$, the noise gain was found to be $J(\boldsymbol{D}_{int}, \hat{\boldsymbol{T}}) = 13.7535$.

3) Case 3: $\boldsymbol{D}$ Is a Scalar Matrix: In this case, the quasi-Newton algorithm was applied to minimize (25)

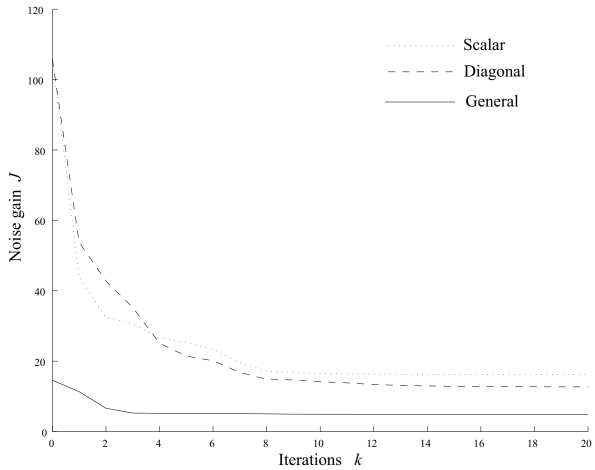

Fig. 2. Profiles of iterative noise gain minimization.

for $\boldsymbol{D} = \alpha \boldsymbol{I}_3$ with a scalar α. The algorithm converges after 20 iterations to converge to the solution

$$\boldsymbol{D} = -0.779678\, \boldsymbol{I}_3$$

$$\boldsymbol{T} = \begin{bmatrix} 3.252790 & -0.081745 & -0.198376 \\ -1.717225 & 1.220068 & -0.792487 \\ 0.546599 & -0.854316 & 2.295944 \end{bmatrix}$$

and the minimized noise gain was found to be $J(\boldsymbol{D}, \hat{\boldsymbol{T}}) = 16.2006$. Next, the EF matrix $\boldsymbol{D} = \alpha \boldsymbol{I}_3$ was rounded to a power-of-two representation with 3 bits after the binary point as well as an integer representation. It was found that these representations were given by $\boldsymbol{D}_{3bit} = \mathrm{diag}\{0.750, 0.750, 0.750\}$ and $\boldsymbol{D}_{int} = \mathrm{diag}\{1, 1, 1\}$, respectively. The corresponding noise gains were obtained as $J(\boldsymbol{D}_{3bit}, \hat{\boldsymbol{T}}) = 16.2370$ and $J(\boldsymbol{D}_{int}, \hat{\boldsymbol{T}}) = 18.2063$, respectively.

The above simulation results in terms of noise gain $J(\boldsymbol{D}, \hat{\boldsymbol{T}})$ in (25) are summarized in Table 1. For comparison purpose, their counterparts obtained using the method in [9] are also included in the table, where the minimization of the roundoff noise was carried out using EF and state-variable coordinate transformation, but in a separate manner. From the table, it is observed that the proposed joint optimization offers improved reduction in roundoff noise gain for the cases of a scalar EF matrix and a diagonal EF matrix when compared with those obtained by using *separate* optimization. However, in the case of a general EF matrix, the optimal solution with infinite precision appears to be quite sensitive to the parameter perturbations.

More reduction of the noise gain might be possible by re-designing the coordinate transformation matrix $\boldsymbol{T}$ for the optimally quantized $\boldsymbol{D}$.

5. CONCLUSION

The joint optimization problem of EF and realization to minimize the effects of roundoff noise of 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 cases where EF matrix is a general, diagonal, or scalar matrix. The effectiveness for the cases of a scalar EF matrix and a diagonal EF matrix compared with the existing method [9] has been illustrated by a numerical example.

Table 1: Noise gain $J(\boldsymbol{D}, \hat{\boldsymbol{T}})$ for different EF schemes.

Error-Feedback Scheme	Accuracy of $\boldsymbol{D}$		
	Infinite Precision	3 Bit Quantization	Integer Quantization
$\boldsymbol{D} = \boldsymbol{0}$ Separate	28.6187		
Scalar Separate [9]	20.1235	20.1810	26.0527
Scalar Joint	16.2006	16.2370	18.2063
Diagonal Separate [9]	16.4104	16.4547	17.4039
Diagonal Joint	12.7097	12.7722	13.7535
General Separate [9]	11.6352	11.7054	16.5814
General Joint	4.8823	23.4873	293.0187

REFERENCES

[1] C. T. Mullis and R. A. Roberts, "Synthesis of minimum roundoff noise fixed point digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-23, pp. 551-562, Sept. 1976.

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

[3] W. E. Higgins and D. C. Munson, "Optimal and suboptimal error-spectrum shaping for cascade-form digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 429-437, May 1984.

[4] 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.

[5] P. P. Vaidyanathan, "On error-spectrum shaping in state-space digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 88-92, Jan. 1985.

[6] D. Williamson, "Roundoff noise minimization and pole-zero sensitivity in fixed-point digital filters using residue feedback," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-34, pp. 1210-1220, Oct. 1986.

[7] W.-S. Lu and T. Hinamoto, "Jointly optimized error-feedback and realization for roundoff noise minimization in state-space digital filters," *IEEE Trans. Signal Processing*, vol. 53, pp. 2135-2145, June 2005.

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

[9] T. Hinamoto and S. Yamamoto, "Error spectrum shaping in closed-loop systems with state-estimate feedback controller," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS'02)*, May 2002, vol. 1, pp. 289-292.

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