Roundoff Noise Minimization in State-Space Discrete-Time Systems Using Joint Optimization of High-Order Error Feedback and Realization

Takao Hinamoto, *Life Fellow, IEEE*, Akimitsu Doi, *Member, IEEE*, and Wu-Sheng Lu, *Life Fellow, IEEE*

***Abstract*— This paper deals with the joint optimization problem of high-order error-feedback and state-space realization for minimizing roundoff noise at the output of a closed-loop system with a state-estimate feedback regulator subject to l_2-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. A case study is presented to demonstrate the validity and effectiveness of the proposed technique.**

***Index Terms*— l_2-scaling, high-order error-feedback, joint optimization, roundoff noise minimization**

I. INTRODUCTION

Due to finite-precision nature of computer arithmetic, roundoff noise in fixed-point IIR digital filters occurs that in turn degrades the filters' 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]-[10]. It is known that the roundoff noise is critically dependent on the internal structure of the IIR filter, and several techniques for realizing a state-space filter structure that minimizes the roundoff noise at the filter output subject to l_2-scaling constraints have been explored by appropriately choosing a linear transformation to state-space coordinates [11]-[14]. 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 [15],[16]. More recently, the use of high-order EF is found to offer improved noise-reduction performance for state-space digital filters [17]. Another relevant development was initiated in [18] where roundoff noise in a closed-loop system equipped with a state-estimate feedback regulator is analyzed and an algorithm for constructing a state-estimate feedback regulator 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 [18]. Recently, the joint optimization problem of static EF and state-space realization for a state-estimate feedback regulator has been considered so as to minimize the roundoff noise at the output of the closed-loop system with a state-estimate feedback regulator subject to l_2-scaling constraints [19]. Furthermore, the l_2-sensitivity of a closed-loop system with respect to the coefficients of a state-estimate feedback regulator has also been analyzed, and two iterative techniques for constructing an optimal state-estimate feedback regulator that minimizes the l_2-sensitivity subject to l_2-scaling constraints have been explored [20].

In this paper, roundoff noise in a closed-loop system equipped with a state-estimate feedback regulator is analyzed in connection with *high-order* EF, and the joint optimization problem of *high-order* EF and realization for minimizing roundoff noise in the closed-loop system with a state-estimate feedback regulator 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 regulator. The *high-order* EF, as will be demonstrated later in the paper, is responsible for a considerable performance enhancement in noise gain reduction relative to the system that employs static EF. A state-space model for quantized discrete-time systems with *high-order* EF is established, a formula for roundoff noise gain is derived, and our objective—the joint optimization of *high-order* EF and state-space realization for the roundoff noise minimization subject to l_2-scaling conditions is formulated as a constrained optimization problem. 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. An iterative technique is also developed for obtaining the optimal solution such that static EF with a general matrix minimizes the roundoff noise at the system's output subject to l_2-scaling constraints. A case study is presented to illustrate the validity and effectiveness of the proposed algorithm and demonstrate its performance.

Throughout, $\boldsymbol{I}_n$ denotes the identity matrix of dimension $n \times n$. The transpose (conjugate transpose) of a matrix $\boldsymbol{A}$ is indicated by $\boldsymbol{A}^T$ ($\boldsymbol{A}^*$). tr$[\boldsymbol{A}]$ is used to denote the trace of a square matrix $\boldsymbol{A}$. Moreover, bold uppercase, bold lowercase and plain lowercase are used to make the distinction between matrices, vectors and scalar values, respectively.

Manuscript received February xx, 2014; revised xxxx xx, 2014; accepted xxxx xx, 2014.

T. Hinamoto and A. Doi are with Hiroshima Institute of Technology, Hiroshima, 731-5193, Japan (e-mails: hinamoto@ieee.org, doi@cc.it-hiroshima.ac.jp, Phone:+81-82-921-4338, Fax:+81-82-921-8978)

W.-S. Lu is with the Depertment of Electrical and Computer Engineering, University of Victoria, Victoria, B.C, Canada, V8W 3P6. (e-mail: wslu@ece.uvic.ca, Phone:+1-250-721-8692, Fax:+1-250-721-6052)

II. System Model and Problem Formulation

A. A state-space model of quantized discrete-time system with high-order EF

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

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

then the observer and associated regulator 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-regulator 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

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

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 $\sigma^2\boldsymbol{I}_n$ characterized by

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

A block diagram of a state-estimate feedback regulator with first-order EF is illustrated in Fig. 1.

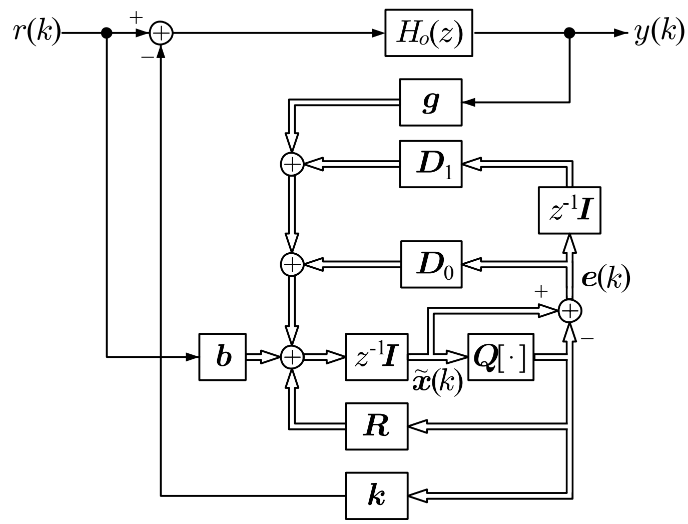

Fig. 1. Block diagram of a state-estimate feedback regulator with first-order EF.

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

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

B. A formula for noise-gain evaluation

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

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

In order to derive a formula for evaluating the normalized 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

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

Substituting (10) into (8) yields

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

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

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

C. Problem formulation

In order to reach an adequate formulation for the problem at hand, what remains to be done is to 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, we derive the transfer function from reference input $r(k)$ to state-variable vector $[\boldsymbol{x}(k)^T, \tilde{\boldsymbol{x}}(k)^T]^T$ by setting $\boldsymbol{e}(k) = \boldsymbol{0}$ in (6) and then taking z-transform of (6). This gives

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

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 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

$$\overline{\boldsymbol{K}} = \boldsymbol{T}^{-1}\boldsymbol{K}\boldsymbol{T}^{-T}, \quad \overline{\boldsymbol{W}}_i = \boldsymbol{T}^T\boldsymbol{W}_i\boldsymbol{T} \quad \text{for} \quad i = 1,2,3,4. \tag{18}$$

respectively. 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), the normalized noise gain is found to be

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

Formula (20) provides an analytic foundation for the minimization of the normalized noise gain by *jointly* optimizing the EF loops as well as state-space coordinate transformation. Formally, the problem being considered here can be stated as 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).

III. A JOINT OPTIMIZATION TECHNIQUE

A. High-order EF with diagonal matrices

In what follows, it is assumed that the high-order EF matrices are all diagonal in the objective function described by (20). Because the constraints in (19) 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 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

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

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 $J_2(\boldsymbol{D}, \boldsymbol{T})$ in (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 $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}})$ in (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, known as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, updates the most recent point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as [21]

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

where

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

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. The BFGS algorithm is a well-known descent algorithm, meaning that a twice continuously differentiable objective function at the iterates generated by the BFGS algorithm is monotonically decreasing. Although the objective function involved in the joint optimization is not convex in general, in theory it was shown [22] that if the starting point is sufficiently close to a local minimizer and the initial Hessian approximation is sufficiently close to the Hessian at that minimizer, then the BFGS iterates will converge to the minimizer. In practice, the above iterative algorithm was applied to quite a number of simulation examples, and fast convergence was observed in each and every instance where the BFGS algorithm converged to an identical solution point for a variety of initial state-space realizations. In the next section, the effectiveness of the proposed algorithm is illustrated using a representative example from our simulation studies.

The implementation efficiency and solution accuracy of the quasi-Newton algorithm greatly depends on how the gradient $\nabla J(\boldsymbol{x})$ is evaluated. With high-order diagonal EF matrices, we have derived closed-form formulas, as shown below, for the partial derivatives of $J(\boldsymbol{x})$ with respect to the elements of $\hat{\boldsymbol{T}}$ as well as those of $\boldsymbol{D} = [\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}]$. The derivation process is quite involved, details are omitted here.

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

$$\begin{aligned} \frac{\partial J(\boldsymbol{x})}{\partial t_{ij}} = &\, 2\boldsymbol{e}_j^T\Big[\hat{\boldsymbol{T}}\hat{\boldsymbol{W}}_4\hat{\boldsymbol{T}}^T \\ &- \sum_{p=0}^{N-1}\hat{\boldsymbol{T}}\left(\hat{\boldsymbol{\Psi}}_4(p+1,0) + \hat{\boldsymbol{\Psi}}_4^T(p+1,0)\right)\hat{\boldsymbol{T}}^T\boldsymbol{D}_p \\ &+ \frac{1}{2}\sum_{p=0}^{N-1}\sum_{q=0}^{N-1}\hat{\boldsymbol{T}}\left(\hat{\boldsymbol{\Psi}}_4(p,q) + \hat{\boldsymbol{\Psi}}_4^T(p,q)\right)\hat{\boldsymbol{T}}^T\boldsymbol{D}_p\boldsymbol{D}_q\Big]\hat{\boldsymbol{T}}\boldsymbol{g}_{ij} \\ &\qquad i, j = 1, 2, \cdots, n \end{aligned} \tag{27}$$

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

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

Let the high-order EF matrices assume the form

$$\boldsymbol{D}_p = \operatorname{diag}\{\beta_{p1}, \beta_{p2}, \cdots, \beta_{pn}\} \tag{28}$$

for $p = 0, 1, \cdots, N-1$. The gradient of $J(\boldsymbol{x})$ with respect to β_{pi} is given by

$$\begin{aligned} \frac{\partial J(\boldsymbol{x})}{\partial \beta_{pi}} = &\ 2\sum_{q=0}^{N-1}\beta_{qi}\left(\hat{\boldsymbol{T}}\hat{\boldsymbol{\Psi}}_4(p,q)\hat{\boldsymbol{T}}^T\right)_{ii} \\ &- 2\left(\hat{\boldsymbol{T}}\hat{\boldsymbol{\Psi}}_4(p+1,0)\hat{\boldsymbol{T}}^T\right)_{ii} \\ &p = 0, 1, \cdots, N-1 \text{ and } i = 1, 2, \cdots, n. \end{aligned} \tag{29}$$

B. Static ($N = 1$) EF with a general matrix

Consider static EF with a general matrix. Assuming that $N = 1$ and $\boldsymbol{D}_0 = \boldsymbol{D}$ in the model described by (5), the normalized noise gain in (11) can be written as

$$\begin{aligned} J_1(\boldsymbol{D}) = \operatorname{tr}\Big[&\boldsymbol{W}_4 - (\boldsymbol{k}_o^T\boldsymbol{b}_o^T\boldsymbol{W}_3^T + \boldsymbol{F}_o^T\boldsymbol{W}_4)\boldsymbol{D} \\ &- \boldsymbol{D}^T(\boldsymbol{W}_3\boldsymbol{b}_o\boldsymbol{k}_o + \boldsymbol{W}_4\boldsymbol{F}_o) + \boldsymbol{D}^T\boldsymbol{W}_4\boldsymbol{D}\Big] \end{aligned} \tag{30}$$

which is equivalent to

$$\begin{aligned} J_1(\boldsymbol{D}) = \operatorname{tr}\big[&(\boldsymbol{M} - \boldsymbol{D})^T\boldsymbol{W}_4(\boldsymbol{M} - \boldsymbol{D}) \\ &+ \boldsymbol{W}_4 - \boldsymbol{M}^T\boldsymbol{W}_4\boldsymbol{M}\big] \end{aligned} \tag{31}$$

where $\boldsymbol{M} = \boldsymbol{F}_o + \boldsymbol{L}\boldsymbol{b}_o\boldsymbol{k}_o$ and $\boldsymbol{L}$ is an $n \times n$ matrix satisfying $\boldsymbol{W}_4\boldsymbol{L} = \boldsymbol{W}_3$.

With an equivalent state-space realization as specified in (17), the normalized noise gain in (31) is changed to

$$J_2(\boldsymbol{D},\boldsymbol{T}) = \text{tr}\Big[(\boldsymbol{MT}-\boldsymbol{TD})^T\boldsymbol{W}_4(\boldsymbol{MT}-\boldsymbol{TD}) + \boldsymbol{T}^T(\boldsymbol{W}_4 - \boldsymbol{M}^T\boldsymbol{W}_4\boldsymbol{M})\boldsymbol{T}\Big]. \tag{32}$$

By virtue of (21), the normalized noise gain in (32) can be expressed as

$$J_3(\boldsymbol{D},\hat{\boldsymbol{T}}) = \text{tr}[(\hat{\boldsymbol{M}}\hat{\boldsymbol{T}}^T - \hat{\boldsymbol{T}}^T\boldsymbol{D})^T\hat{\boldsymbol{W}}_4(\hat{\boldsymbol{M}}\hat{\boldsymbol{T}}^T - \hat{\boldsymbol{T}}^T\boldsymbol{D})] + \text{tr}[\hat{\boldsymbol{T}}(\hat{\boldsymbol{W}}_4 - \hat{\boldsymbol{M}}^T\hat{\boldsymbol{W}}_4\hat{\boldsymbol{M}})\hat{\boldsymbol{T}}^T] \tag{33}$$

where

$$\hat{\boldsymbol{M}} = \boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{M}\boldsymbol{K}^{\frac{1}{2}}, \qquad \hat{\boldsymbol{W}}_4 = \boldsymbol{K}^{\frac{1}{2}}\boldsymbol{W}_4\boldsymbol{K}^{\frac{1}{2}}.$$

From (33) it is evident that the optimal choice of $\boldsymbol{D}$ is given by

$$\boldsymbol{D} = \hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{M}}\hat{\boldsymbol{T}}^T \tag{34}$$

which leads to

$$J_3(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{M}}\hat{\boldsymbol{T}}^T,\hat{\boldsymbol{T}}) = \text{tr}[\hat{\boldsymbol{T}}(\hat{\boldsymbol{W}}_4 - \hat{\boldsymbol{M}}^T\hat{\boldsymbol{W}}_4\hat{\boldsymbol{M}})\hat{\boldsymbol{T}}^T]. \tag{35}$$

In this case, the joint optimization of EF and realization is not needed because the optimal $\boldsymbol{D}$ has already been determined by the closed-form in (34), 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 (35).

In order to reduce solution sensitivity, the objective function in (35) is modified to

$$\begin{aligned} J_o(\boldsymbol{x}) &= \mu\,\text{tr}[\hat{\boldsymbol{T}}\hat{\boldsymbol{W}}_4\hat{\boldsymbol{T}}^T] + (1-\mu)\,\text{tr}[\hat{\boldsymbol{T}}(\hat{\boldsymbol{W}}_4 - \hat{\boldsymbol{M}}^T\hat{\boldsymbol{W}}_4\hat{\boldsymbol{M}})\hat{\boldsymbol{T}}^T] \\ &= \text{tr}[\hat{\boldsymbol{T}}\{\hat{\boldsymbol{W}}_4 - (1-\mu)\,\hat{\boldsymbol{M}}^T\hat{\boldsymbol{W}}_4\hat{\boldsymbol{M}}\}\hat{\boldsymbol{T}}^T] \end{aligned} \tag{36}$$

where $\boldsymbol{x}$ is defined from (23) as $\boldsymbol{x} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T, \cdots, \boldsymbol{t}_n^T)^T$ and $0 \le \mu \le 1$ is a scalar parameter that weights the importance of reducing $\text{tr}[\hat{\boldsymbol{T}}\hat{\boldsymbol{W}}_4\hat{\boldsymbol{T}}^T]$ relative to reducing (35).

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

$$\frac{\partial J_o(\boldsymbol{x})}{\partial t_{ij}} = 2\boldsymbol{e}_j^T\hat{\boldsymbol{T}}\{\hat{\boldsymbol{W}}_4 - (1-\mu)\,\hat{\boldsymbol{M}}^T\hat{\boldsymbol{W}}_4\hat{\boldsymbol{M}}\}\hat{\boldsymbol{T}}^T\hat{\boldsymbol{T}}\boldsymbol{g}_{ij} \tag{37}$$

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

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

This implies that the BFGS algorithm can also be applied to minimize (36) with respect to vector $\boldsymbol{x}$.

IV. A CASE STUDY

This section presents a case study for the purpose of demonstrating the numerical efficiency of the proposed joint optimization algorithm and effectiveness of using a high-order (i.e. dynamic) EF relative to a constant (i.e. static) EF for performance enhancement.

A. A system model and noise gain without EF

Consider a linear discrete-time system specified by

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

$$\boldsymbol{b}_o = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$$

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

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

$$\boldsymbol{g}_o = \begin{bmatrix} -0.006436 & 3.683651 & 5.083920 \end{bmatrix}^T$$

$$\boldsymbol{k}_o = \begin{bmatrix} 0.471552 & -0.367158 & 3.062267 \end{bmatrix}.$$

The controllability and observability Gramians $\boldsymbol{K}$ and $\boldsymbol{W}_i$ for $i = 1, 2, 3, 4$ relating to the closed-loop system in (6) were computed from (16) and (11) as

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

$$\boldsymbol{W}_1 = 10^{-1}\begin{bmatrix} 0.214594 & 0.722665 & 1.178558 \\ 0.722665 & 3.036154 & 4.569871 \\ 1.178558 & 4.569871 & 7.305712 \end{bmatrix}$$

$$\boldsymbol{W}_2 = \begin{bmatrix} -0.277086 & -0.644246 & -0.408932 \\ -1.314870 & -3.105240 & -1.934613 \\ -1.690779 & -4.034031 & -2.725265 \end{bmatrix}$$

$$\boldsymbol{W}_3 = \begin{bmatrix} -0.277086 & -1.314870 & -1.690779 \\ -0.644246 & -3.105240 & -4.034031 \\ -0.408932 & -1.934613 & -2.725265 \end{bmatrix}$$

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

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 with no EF was computed from (11) as

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

B. Noise gain when using static $(N = 1)$ and dynamic $(N \ge 2)$ EF

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 started from the initial noise gain $J_3(\boldsymbol{D},\hat{\boldsymbol{T}}) = 105.695229$ and 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 which the noise gain was reduced to

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

The profile of $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}})$ during the first 19 iterations of the algorithm is dipicted in Fig. 2.

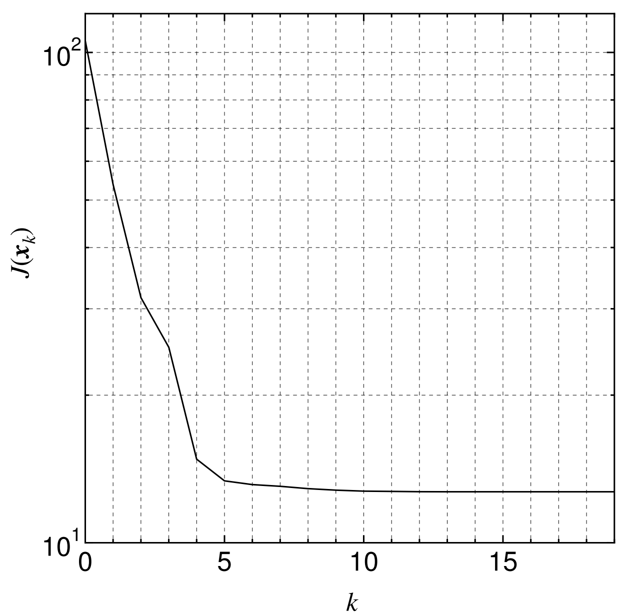

Fig. 2. Profile of $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}})$ during the first 19 iterations.

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 started from the initial noise gain $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}}) = 73.158734$ and 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 which the noise gain was reduced to

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

The profile of $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}})$ during the first 24 iterations of the algorithm is dipicted in Fig. 3.

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 started from the initial noise gain $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}}) = 105.588037$ and took the algorithm 105 iterations to converge to the solution

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

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

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

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

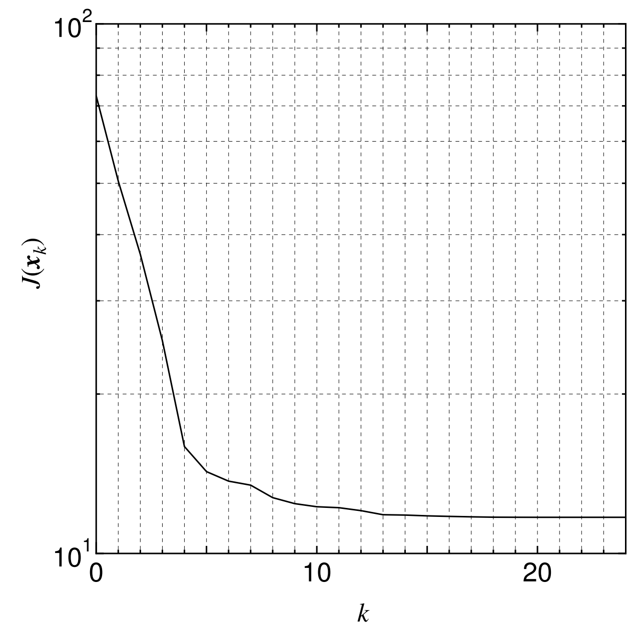

Fig. 3. Profile of $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}})$ during the first 24 iterations.

with which the noise gain was found to be

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

The profile of $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}})$ during the first 105 iterations of the algorithm is dipicted in Fig. 4.

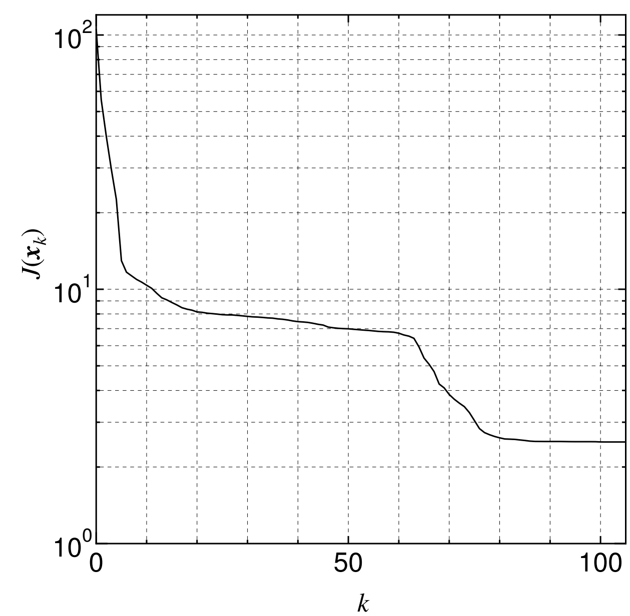

Fig. 4. Profile of $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}})$ during the first 105 iterations.

C. Joint versus separate optimization

Our case study also includes comparisons of the performance in terms of the normalized noise gain $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}})$ in (24) offered by jointly optimized EF (with $N = 1, 2, 3$) with those obtained by the EF of the same order that were optimized by using the *separate* optimization method [15] for three implementation settings. The main steps involved in the separate optimization method are included in the appendix for the reader's convenience. The numerical results of these nine cases are summarized in Table I where the column associated with "Infinite Precision" shows the values of $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}})$ derived from the optimal $\hat{\boldsymbol{T}}$ and $\boldsymbol{D}$, and the columns with "3-Bit Quantization (Integer Quantization)" are the values 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 so as to minimize $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}})$ with such a $\boldsymbol{D}$ fixed.

TABLE I
PERFORMANCE COMPARISON FOR DIAGONAL MATRICES EF

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

The results reported above have clearly demonstrated that the use of *high-order* (i.e. $N > 1$) EF, when jointly optimized with state-space realization, can improve the performance of roundoff noise reduction in a significant manner.

D. Noise gain when using static EF with a general matrix

In the case where $N = 1$ and matrix $\boldsymbol{D}_0 = \boldsymbol{D}$ is general, the quasi-Newton algorithm was applied to minimize (36) with $\mu = 0.01$ and tolerance $\varepsilon = 10^{-8}$ in (26). It started from the initial noise gain $J_o(\boldsymbol{x}) = 14.777910$ and took the algorithm 12 iterations to converge to the solution

$$\hat{\boldsymbol{T}} = \begin{bmatrix} -2.321835 & -0.568711 & 1.332833 \\ -1.361485 & -1.090543 & -0.541248 \\ 1.522368 & -1.059474 & -1.127862 \end{bmatrix}$$

$$\boldsymbol{D} = \begin{bmatrix} -0.887178 & 0.120591 & 1.053286 \\ 0.138885 & -0.742341 & -0.951235 \\ -0.376203 & -0.317240 & -0.069456 \end{bmatrix}$$

with which the noise gain was reduced to

$$J_o(\boldsymbol{x}) = 6.323423.$$

The profile of $J_o(\boldsymbol{x})$ during the first 12 iterations of the algorithm is dipicted in Fig. 5.

Using a way similar to Table I, the case study of static EF with a general matrix also includes comparisons of the performance in terms of the normalized noise gain $J_3(\boldsymbol{D}, \hat{\boldsymbol{T}})$ in (33) offered by iteratively optimized EF (with $N = 1$) with that obtained by the EF of the same order that was optimized by using the *separate* (analytic) optimization method [15] for three implementation settings. The main steps involved in the separate (analytic) optimization method are included in the appendix for the reader's convenience. The numerical results of these three cases are summarized in Table II.

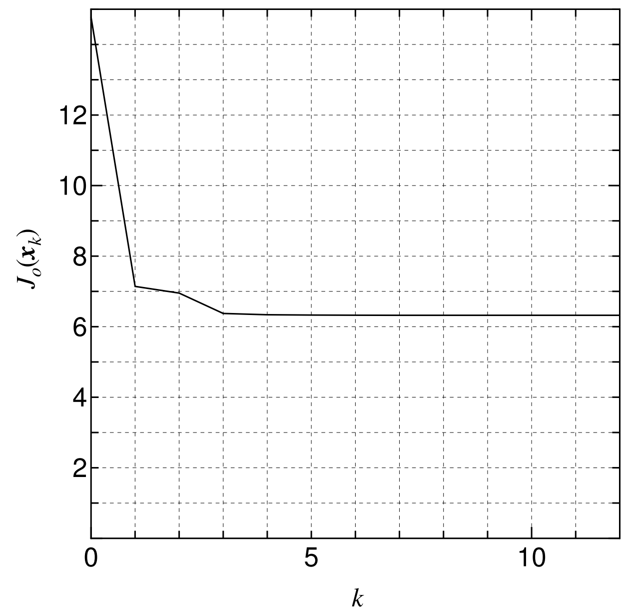

Fig. 5. Profile of $J_o(\boldsymbol{x})$ during the first 12 iterations.

The results reported above have clearly demonstrated that the use of *high-order* (i.e. $N = 3$) EF, when jointly optimized with state-space realization, can also improve the performance of roundoff noise reduction in a significant manner.

TABLE II
PERFORMANCE COMPARISON FOR GENERAL MATRIX EF

N	Optimization	Infinite Precision	3-Bit Quantization	Integer Quantization
1	Analytic	11.635118	11.808013	15.337683
	Iterative	5.541528	6.104080	11.366935

V. 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 regulator 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. Moreover, it has been revealed that the joint optimization of EF and realization can not be applied to the case where a static EF matrix is general, in addition to the case where high-order EF matrices are general. An iterative technique has also been developed for obtaining the optimal solution of static EF with a general matrix that minimizes the roundoff noise at the system's output subject to l_2-scaling constraints. The validity and effectiveness of the proposed technique has been demonstrated by a case study.

APPENDIX

ANALYTIC METHOD FOR SEPARATE OPTIMIZATION

The separate minimization is carried out in two major steps. First, we fix the EF matrices to $\boldsymbol{D}_i = \boldsymbol{0}$ for $i = 0, 1, \cdots, N-1$ in (20) so that the objective function is reduced to $J_2(\boldsymbol{0}, \boldsymbol{T}) = \mathrm{tr}[\boldsymbol{T}^T \boldsymbol{W}_4 \boldsymbol{T}]$ which is minimized with respect to matrix $\boldsymbol{T}$ subject to the l_2-scaling constraints in (19). Second, with $\boldsymbol{T}$ optimized in the first step, (20) is minimized under the fixed $\boldsymbol{T}$ with respect to matrices $\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}$. To perform the first step, we define the Lagrange function

$$J_o(\boldsymbol{P}, \lambda) = \mathrm{tr}[\boldsymbol{W}_4 \boldsymbol{P}] + \lambda \left(\mathrm{tr}[\boldsymbol{K}\boldsymbol{P}^{-1}] - n\right) \tag{A.1}$$

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

$$\begin{aligned} \frac{\partial J_o(\boldsymbol{P}, \lambda)}{\partial \boldsymbol{P}} &= \boldsymbol{W}_4 - \lambda \boldsymbol{P}^{-1}\boldsymbol{K}\boldsymbol{P}^{-1} = \boldsymbol{0} \\ \frac{\partial J_o(\boldsymbol{P}, \lambda)}{\partial \lambda} &= \mathrm{tr}[\boldsymbol{K}\boldsymbol{P}^{-1}] - n = 0 \end{aligned} \tag{A.2}$$

which lead to

$$\boldsymbol{P}\boldsymbol{W}_4\boldsymbol{P} = \lambda \boldsymbol{K}, \qquad \mathrm{tr}[\boldsymbol{K}\boldsymbol{P}^{-1}] = n. \tag{A.3}$$

From (A.3) it follows that

$$\begin{aligned} \boldsymbol{P} &= \sqrt{\lambda}\, \boldsymbol{W}_4^{-\frac{1}{2}} \left[\boldsymbol{W}_4^{\frac{1}{2}} \boldsymbol{K} \boldsymbol{W}_4^{\frac{1}{2}}\right]^{\frac{1}{2}} \boldsymbol{W}_4^{-\frac{1}{2}} \\ \frac{1}{\sqrt{\lambda}}\, \mathrm{tr}[\boldsymbol{K}\boldsymbol{W}_4]^{\frac{1}{2}} &= \frac{1}{\sqrt{\lambda}} \Big(\sum_{i=1}^{n} \theta_i\Big) = n \end{aligned} \tag{A.4}$$

where θ_i for $i = 1, 2, \cdots, n$ denote the eigenvalues of $\boldsymbol{K}\boldsymbol{W}_4$. Thus, we obtain

$$\boldsymbol{P} = \frac{1}{n} \Big(\sum_{i=1}^{n} \theta_i\Big) \boldsymbol{W}_4^{-\frac{1}{2}} \left[\boldsymbol{W}_4^{\frac{1}{2}} \boldsymbol{K} \boldsymbol{W}_4^{\frac{1}{2}}\right]^{\frac{1}{2}} \boldsymbol{W}_4^{-\frac{1}{2}}. \tag{A.5}$$

Substituting (A.5) into (A.1) yields the minimum value of $J_o(\boldsymbol{P}, \lambda)$ as

$$\min_{\boldsymbol{P}, \lambda} J_o(\boldsymbol{P}, \lambda) = \frac{1}{n} \big(\sum_{i=1}^{n} \theta_i\Big)^2. \tag{A.6}$$

Referring to (A.5), the optimal coordinate transformation matrix $\boldsymbol{T}$ that minimizes (A.1) can be obtained in a closed form as

$$\boldsymbol{T} = \frac{1}{\sqrt{n}} \Big(\sum_{i=1}^{n} \theta_i\Big)^{\frac{1}{2}} \boldsymbol{W}_4^{-\frac{1}{2}} \left[\boldsymbol{W}_4^{\frac{1}{2}} \boldsymbol{K} \boldsymbol{W}_4^{\frac{1}{2}}\right]^{\frac{1}{4}} \boldsymbol{U}_o. \tag{A.7}$$

where $\boldsymbol{U}_o$ is an arbitrary $n \times n$ orthogonal matrix. From (A.7) it follows that

$$\begin{aligned} \overline{\boldsymbol{K}} &= \boldsymbol{T}^{-1}\boldsymbol{K}\boldsymbol{T}^{-T} \\ &= n \Big(\sum_{i=1}^{n} \theta_i\Big)^{-1} \boldsymbol{U}_o^T \left[\boldsymbol{W}_4^{\frac{1}{2}} \boldsymbol{K} \boldsymbol{W}_4^{\frac{1}{2}}\right]^{\frac{1}{2}} \boldsymbol{U}_o \end{aligned} \tag{A.8}$$

Next, we choose the $n \times n$ orthogonal matrix $\boldsymbol{U}_o$ such that matrix $\overline{\boldsymbol{K}}$ in (A.8) satisfies the l_2-scaling constraints in (19). To this end, we perform the eigenvalue-eigenvector decomposition

$$\left[\boldsymbol{W}_4^{\frac{1}{2}} \boldsymbol{K} \boldsymbol{W}_4^{\frac{1}{2}}\right]^{\frac{1}{2}} = \boldsymbol{Q}\, \mathrm{diag}\{\theta_1, \theta_2, \cdots, \theta_n\} \boldsymbol{Q}^T \tag{A.9}$$

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

$$n \Big(\sum_{i=1}^{n} \theta_i\Big)^{-1} \left[\boldsymbol{W}_4^{\frac{1}{2}} \boldsymbol{K} \boldsymbol{W}_4^{\frac{1}{2}}\right]^{\frac{1}{2}} = \boldsymbol{Q}\boldsymbol{\Lambda}^{-2}\boldsymbol{Q}^T \tag{A.10}$$

where

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

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

$$\boldsymbol{Z}\boldsymbol{\Lambda}^{-2}\boldsymbol{Z}^T = \begin{bmatrix} 1 & * & \cdots & * \\ * & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ * & \cdots & * & 1 \end{bmatrix} \tag{A.11}$$

can be obtained by numerical manipulations [13, p.278]. By choosing $\boldsymbol{U}_o = \boldsymbol{Q}\boldsymbol{Z}^T$ in (A.7), the optimal coordinate transformation matrix $\boldsymbol{T}$ satisfying (19) and (A.6) can be constructed as

$$\boldsymbol{T} = \frac{1}{\sqrt{n}} \Big(\sum_{i=1}^{n} \theta_i\Big)^{\frac{1}{2}} \boldsymbol{W}_4^{-\frac{1}{2}} \left[\boldsymbol{W}_4^{\frac{1}{2}} \boldsymbol{K} \boldsymbol{W}_4^{\frac{1}{2}}\right]^{\frac{1}{4}} \boldsymbol{Q}\boldsymbol{Z}^T. \tag{A.12}$$

Suppose $\boldsymbol{D}_p$ is a diagonal matrix for $p = 0, 1, \cdots, N-1$. In this case, matrix $\boldsymbol{D}_p$ assumes the form

$$\boldsymbol{D}_p = \mathrm{diag}\{\beta_{p1}, \beta_{p2}, \cdots, \beta_{pn}\} \quad \text{for} \quad p = 0, 1, \ldots, N-1. \tag{A.13}$$

It follows that

$$\begin{aligned} \frac{\partial J_2(\boldsymbol{D}, \boldsymbol{T})}{\partial \beta_{pi}} &= -2\left(\boldsymbol{T}^T \boldsymbol{\Psi}_4(p+1, 0)\boldsymbol{T}\right)_{ii} \\ &\quad + 2\sum_{q=0}^{N-1} \beta_{qi} \left(\boldsymbol{T}^T \boldsymbol{\Psi}_4(p, q)\boldsymbol{T}\right)_{ii} = 0 \end{aligned} \tag{A.14}$$

for $i = 1, 2, \cdots, n$. As a result, matrix $\boldsymbol{D}_p$ can be derived from

$$\begin{bmatrix} \beta_{0i} \\ \beta_{1i} \\ \vdots \\ \beta_{N-1,i} \end{bmatrix} = \boldsymbol{V}_i^{-1} \begin{bmatrix} v_i(1,0) \\ v_i(2,0) \\ \vdots \\ v_i(N,0) \end{bmatrix} \tag{A.15}$$

where

$$\boldsymbol{V}_i = \begin{bmatrix} v_i(0,0) & v_i(0,1) & \cdots & v_i(0,N-1) \\ v_i(1,0) & v_i(1,1) & \cdots & v_i(1,N-1) \\ \vdots & \vdots & \ddots & \vdots \\ v_i(N-1,0) & v_i(N-1,1) & \cdots & v_i(N-1,N-1) \end{bmatrix}$$

with $v_i(p,q) = (\boldsymbol{T}^T \boldsymbol{\Psi}_4(p,q)\boldsymbol{T})_{ii}$ for $i = 1, 2, \cdots, n$.

Finally, consider the case where static ($N = 1$) EF has a general matrix. Assuming that $N = 1$ and $\boldsymbol{D}_0 = \boldsymbol{D}$ in (5), the normalized noise gain in (30) is written as

$$\begin{aligned} J_1(\boldsymbol{D}) &= \mathrm{tr}\left[\boldsymbol{W}_4\right] + \mathrm{tr}\left[\boldsymbol{D}^T \boldsymbol{W}_4 \boldsymbol{D}\right] \\ &\quad - 2\,\mathrm{tr}\left[(\boldsymbol{W}_3 \boldsymbol{b}_o \boldsymbol{k}_o + \boldsymbol{W}_4 \boldsymbol{F}_o)^T \boldsymbol{D}\right]. \end{aligned} \tag{A.16}$$

With an equivalent state-space realization as specified in (17), the normalized noise gain in (A.16) is changed to

$$\begin{aligned} J_2(\boldsymbol{D},\boldsymbol{T}) = \operatorname{tr}\left[\boldsymbol{T}^T\boldsymbol{W}_4\boldsymbol{T}\right] + \operatorname{tr}\left[\boldsymbol{D}^T\boldsymbol{T}^T\boldsymbol{W}_4\boldsymbol{T}\boldsymbol{D}\right] \\ -2\operatorname{tr}\left[\boldsymbol{T}^T(\boldsymbol{W}_3\boldsymbol{b}_o\boldsymbol{k}_o+\boldsymbol{W}_4\boldsymbol{F}_o)^T\boldsymbol{T}\boldsymbol{D}\right]. \end{aligned} \tag{A.17}$$

From (A.17) it follows that

$$\begin{aligned} \frac{\partial J_2(\boldsymbol{D},\boldsymbol{T})}{\partial \boldsymbol{D}} &= 2\boldsymbol{T}^T\left[\boldsymbol{W}_4\boldsymbol{T}\boldsymbol{D} - (\boldsymbol{W}_3\boldsymbol{b}_o\boldsymbol{k}_o+\boldsymbol{W}_4\boldsymbol{F}_o)\boldsymbol{T}\right] \\ &= \boldsymbol{0} \end{aligned} \tag{A.18}$$

which leads to

$$\boldsymbol{D} = \boldsymbol{T}^{-1}\boldsymbol{W}_4^{-1}(\boldsymbol{W}_3\boldsymbol{b}_o\boldsymbol{k}_o+\boldsymbol{W}_4\boldsymbol{F}_o)\boldsymbol{T}. \tag{A.19}$$

This provides an analytic method in order for static ($N=1$) EF with a general matrix to reduce the roundoff noise at the system's output subject to l_2-scaling constraints.

REFERENCES

[1] H. A. Spang, III and P. M. Shultheiss, "Reduction of quantizing noise by use of feedback," *IRE Trans. Commun. Syst.*, vol. CS-10, pp. 373-380, Dec. 1962.

[2] T. Thong and B. Liu, "Error spectrum shaping in narrowband recursive digital filters," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 25, pp. 200-203, Apr. 1977.

[3] D. C. Munson and D. Liu, "Narrow-band recursive filters with error spectrum shaping," *IEEE Trans. Circuits Syst.*, vol. 28, pp. 160-163, Feb. 1981.

[4] T.-L. Chang and S. A. White, "An error cancellation digital-filter structure and its distributed-arithmetic implementation," *IEEE Trans. Circuits Syst.*, vol. 28, pp. 339-342, Apr. 1981.

[5] 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.

[6] M. Renfors, "Roundoff noise in error-feedback state-space filters," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.* (ICASSP'83), Apr. 1983, pp. 619-622.

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

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

[9] 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.

[10] 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.

[11] S. Y. Hwang, "Roundoff noise in state-space digital filtering: A general analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, pp. 256-262, June 1976.

[12] 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.

[13] 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.

[14] L. B. Jackson, A. G. Lindgren and Y. Kim, "Optimal synthesis of second-order state-space structures for digital filters," *IEEE Trans. Circuits Syst.*, vol. 26, pp. 149-153, Mar. 1979.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] T. Hinamoto and T. Kawagoe, "Optimal synthesis of state-estimate feedback controllers with minimum l_2-sensitivity," *IEEE Trans. Circuits Syst. I*, vol. 55, pp. 2402-2410, Sept. 2008.

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

[22] J. E. Dennis and J. J. More, "Quasi-Newton methods, motivation and theory," *SIAM Review*, vol. 19, pp. 46-89, 1977.

Takao Hinamoto (M'77-SM'84-F'01-LF'11) received the B.E. degree from Okayama University, Okayama, Japan, in 1969, the M.E. degree from Kobe University, Kobe, Japan, in 1971, and the Dr. Eng. degree from Osaka University, Osaka, Japan, in 1977, all in electrical engineering.

From 1972 to 1988, he was with the Faculty of Engineering, Kobe University, Kobe, Japan. From 1979 to 1981, he was a Visiting Member of Staff in the Department of Electrical Engineering, Queen's University, Kingston, ON, Canada, on leave from Kobe University. During 1988-1991, he was Professor of electronic circuits in the Faculty of Engineering, Tottori University, Tottori, Japan. During 1992-2009, he was Professor of Electronic Control in the Department of Electrical Engineering, Hiroshima University, Hiroshima, Japan. Since 2009, he has been a Professor Emeritus of Hiroshima University and a Teaching Member of Staff at Hiroshima Institute of Technology, Hiroshima, Japan. His research interests include digital signal processing, system theory, and control engineering. He has published over 400 papers in these areas. He is co-editor and co-author of Two-Dimensional Signal and Image Processing (Tokyo, Japan: SICE, 1996).

Dr. Hinamoto received the IEEE Third Millennium Medal in January 2000. From 1993 to 1995 he was an Associate Editor of the IEEE Trans. on Circuits and Systems II. During 2002-2003 and 2006-2007 he was an Associate Editor of the IEEE Trans. on Circuits and Systems I. He also served as General Chair of the 47th IEEE Int. Midwest Symposium on Circuits and Systems held in Hiroshima, Japan, in July 2004. Since 1995, he has been a Steering Committee Member of the IEEE Int. Midwest Symposium on Circuits and Systems. Since 1998, he has been a Digital Signal Processing Technical Committee Member in the IEEE Circuits and Systems Society. He played a leading role in establishing the Hiroshima Section of IEEE and served as Interim Chair of the Section.

Akimitsu Doi received the B.E. degree in Electronic Engineering from Tottori University in 1992, and M.E. and Ph.D. degrees in System Engineering from Hiroshima University in 1994 and 1997, respectively. From 1997 to 2002, he was a Research Associate with the Faculty of Engineering, Hiroshima University, Hiroshima, Japan. From 2002 to 2005, he was a Lecturer in the Faculty of Engineering, Hiroshima Institute of Technology, Hiroshima, Japan. Since 2005, he has been an Associate Professor at the same Institute. His research interests are in the area of digital signal and image processing.

Wu-Sheng Lu (S'81-M'85-SM'90-F'99-LF'12) received the B.Sc. degree in Mathematics from Fudan University, Shanghai, China, in 1964, the M.S. degree in electrical engineering, and the Ph.D. degree in control science from the University of Minnesota, Minneapolis, USA, in 1983 and 1984, respectively. He was a post-doctoral fellow at the University of Victoria, Victoria, B.C., Canada, in 1985 and held a visiting assistant professorship with the University of Minnesota in 1986. Since 1987, he has been with the University of Victoria where he is currently a Professor. His current teaching and research interests are in the general areas of digital signal processing and application of optimization methods. He is the co-author with A. Antoniou of Two-Dimensional Digital Filters (Marcel Dekker, 1992). and Practical Optimization: Algorithms and Engineering Applications (Springer, 2007).

Dr. Lu served as an Associate Editor of the Canadian Journal of Electrical and Computer Engineering in 1989, and was Editor of the same journal from 1990 to 1992. He was an Associate Editor of the IEEE Trans. on Circuits and Systems II from 1993 to 1995, for IEEE Trans. on Circuits and Systems I from 1999 to 2001 and from 2004 to 2005, for International Journal of Multidimensional Systems and Signal Processing from 1997 to 2006, and for Circuits, Systems, and Signal Processing from 2009 to 2011.