Optimal Error Feedback and Realization for Roundoff Noise Minimization in Linear Discrete-Time Systems with Full-Order State Observer Feedback

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*— The optimization problem of error-feedback and realization to minimize the roundoff noise in a closed-loop system with full-order state observer feedback subject to l_2-scaling constraints is investigated. It is shown that joint optimization of error-feedback and realization is not needed when utilizing a general-matrix error-feedback loop because the optimal error-feedback matrix, in closed-form, can be determined in advance. An analytic technique is developed for obtaining the optimal realization that minimizes the roundoff noise subject to l_2-scaling constraints. A numerical example is presented to demonstrate the validity and effectiveness of the proposed technique.**

1. INTRODUCTION

Due 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 realization that minimizes the roundoff noise at the filter output subject to l_2-scaling constraints have been explored by appropriately choosing a linear transformation for state-space coordinates [6],[7]. As a natural extension of the aforementioned methods, efforts have been made to develop new methods that combine EF and realization for achieving better performance [8]-[10]. Another relevant development was initiated in [11] 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].

In this paper, the roundoff noise in a closed-loop system equipped with full-order state obsever feedback is analyzed in connection with EF, and the optimization problem of EF and realization for minimizing the roundoff noise in the closed-loop system subject to l_2-scaling constraints is investigated.

2. SYSTEM MODEL AND PROBLEM FORMULATION

Consider a stable, controllable and observable linear discrete-time system $(\boldsymbol{A}_o, \boldsymbol{b}_o, \boldsymbol{c}_o)_n$ 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 full-order state observer and associated state feedback law 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 full-order state observer and associated state feedback system 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 $\boldsymbol{R}_o = \boldsymbol{F}_o - \boldsymbol{b}_o\boldsymbol{k}_o$.

Next, we consider the FWL implementation of the system due to the quantizations that are performed before matrix-vector multiplication, and integrate an EF into the system for noise reduction purposes. Then, the model under these circumstances becomes

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

where $\boldsymbol{e}(k) = \tilde{\boldsymbol{x}}(k) - \boldsymbol{Q}[\tilde{\boldsymbol{x}}(k)]$ and $\boldsymbol{D}$ is an $n \times n$ EF matrix. 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 B_c-bit. The FWL state-variable vector $\tilde{\boldsymbol{x}}(k)$ and each output $u(k)$ in (5) has B-bit fractional representations, while the reference input $r(k)$ and the signal $y(k)$ are $(B - B_c)$-bit fractions. The quantizer $\boldsymbol{Q}[\cdot]$ in (5) rounds the B-bit fraction $\tilde{\boldsymbol{x}}(k)$ to $(B - B_c)$-bit after the multiplications and additions, where the sign bit is not counted. It is assumed that the quantization error vector $\boldsymbol{e}(k)$ can be modeled as a Gaussian random process with zero mean and covariance $\sigma^2\boldsymbol{I}_n$. A block diagram of a full-order state observer feedback system with EF is illustrated in Fig. 1.

978-1-4799-8391-9/15/$31.00 ©2015 IEEE

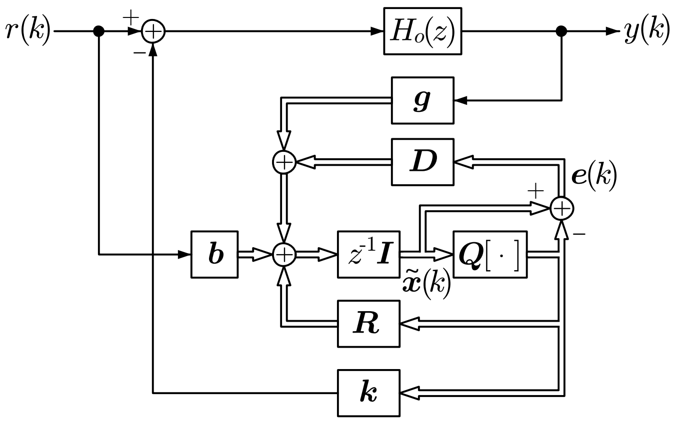

Fig. 1. Block diagram of a full-order state observer feedback system with EF.

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

$$\begin{bmatrix} \boldsymbol{x}(k+1) \\ \tilde{\boldsymbol{x}}(k+1) \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_o & -\boldsymbol{b}_o\boldsymbol{k} \\ \boldsymbol{g}\boldsymbol{c}_o & \boldsymbol{R} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(k) \\ \tilde{\boldsymbol{x}}(k) \end{bmatrix} + \begin{bmatrix} \boldsymbol{b}_o \\ \boldsymbol{b} \end{bmatrix} r(k) + \begin{bmatrix} \boldsymbol{b}_o\boldsymbol{k} \\ \boldsymbol{D}-\boldsymbol{R} \end{bmatrix} \boldsymbol{e}(k) \quad (6)$$

$$y(k) = \begin{bmatrix} \boldsymbol{c}_o & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(k) \\ \tilde{\boldsymbol{x}}(k) \end{bmatrix}.$$

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

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

where

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

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

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

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

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

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

$$\begin{aligned} \boldsymbol{G}_D(z) &\simeq \tilde{\boldsymbol{c}}\boldsymbol{S}(z\boldsymbol{I}_{2n} - \boldsymbol{S}\tilde{\boldsymbol{A}}\boldsymbol{S})^{-1}\boldsymbol{S}\tilde{\boldsymbol{B}} \\ &= \tilde{\boldsymbol{c}}(z\boldsymbol{I}_{2n} - \boldsymbol{\Phi})^{-1}\boldsymbol{U}(z\boldsymbol{I}_n - \boldsymbol{D}) \end{aligned} \quad (10)$$

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

$$\begin{aligned} J_1(\boldsymbol{D}) = \mathrm{tr}\Big[&\boldsymbol{D}^T\boldsymbol{W}_4\boldsymbol{D} - (\boldsymbol{k}_o^T\boldsymbol{b}_o^T\boldsymbol{W}_3^T + \boldsymbol{F}_o^T\boldsymbol{W}_4)\boldsymbol{D} \\ &-\boldsymbol{D}^T(\boldsymbol{W}_3\boldsymbol{b}_o\boldsymbol{k}_o + \boldsymbol{W}_4\boldsymbol{F}_o) + \boldsymbol{W}_4\Big] \end{aligned} \quad (11)$$

where

$$\begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{W}_2 \\ \boldsymbol{W}_3 & \boldsymbol{W}_4 \end{bmatrix} = \boldsymbol{\Phi}^T \begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{W}_2 \\ \boldsymbol{W}_3 & \boldsymbol{W}_4 \end{bmatrix} \boldsymbol{\Phi} + \begin{bmatrix} \boldsymbol{c}_o^T \\ \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{c}_o & \boldsymbol{0} \end{bmatrix}.$$

To reach an adequate formulation for the problem at hand, we derive an explicit expression for the controllability Gramian $\boldsymbol{K}$ for state-variable vector $\tilde{\boldsymbol{x}}(k)$ in system (6), because $\boldsymbol{K}$ is a critical quantity involved in the l_2-scaling constraints on $\tilde{\boldsymbol{x}}(k)$. To this end, the transfer function from the input $r(k)$ to the state 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.

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

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

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

Hence

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

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

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

which can be obtained by solving the Lyapunov equation

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

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

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

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

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

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

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

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

$$J_2(\boldsymbol{D},\boldsymbol{T}) = \mathrm{tr}\Big[\boldsymbol{T}^T\boldsymbol{W}_4\boldsymbol{T}\boldsymbol{D}\boldsymbol{D}^T\Big] + \mathrm{tr}\Big[\boldsymbol{T}^T\boldsymbol{W}_4\boldsymbol{T}\Big] - 2\,\mathrm{tr}\Big[\boldsymbol{T}^T(\boldsymbol{W}_3\boldsymbol{b}_o\boldsymbol{k}_o + \boldsymbol{W}_4\boldsymbol{F}_o)^T\boldsymbol{T}\boldsymbol{D}\Big]. \tag{20}$$

The problem considered here is to design the general EF matrix $\boldsymbol{D}$ 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. ANALYTIC OPTIMIZATION TECHNIQUE

Notice that the normalized noise gain in (11) is equivalent to

$$J_1(\boldsymbol{D}) = \mathrm{tr}\,\big[(\boldsymbol{D}-\boldsymbol{M})^T\boldsymbol{W}_4(\boldsymbol{D}-\boldsymbol{M}) + \boldsymbol{W}_4 - \boldsymbol{M}^T\boldsymbol{W}_4\boldsymbol{M}\big] \tag{21}$$

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 (21) can be expressed as

$$J_2(\boldsymbol{D},\boldsymbol{T}) = \mathrm{tr}\,\big[(\boldsymbol{T}\boldsymbol{D}-\boldsymbol{M}\boldsymbol{T})^T\boldsymbol{W}_4(\boldsymbol{T}\boldsymbol{D}-\boldsymbol{M}\boldsymbol{T}) + \boldsymbol{T}^T(\boldsymbol{W}_4 - \boldsymbol{M}^T\boldsymbol{W}_4\boldsymbol{M})\boldsymbol{T}\big]. \tag{22}$$

From (22), it is evident that the optimal choice of $\boldsymbol{D}$ is given by

$$\boldsymbol{D} = \boldsymbol{T}^{-1}\boldsymbol{M}\boldsymbol{T} \tag{23}$$

which reduces the normalized noise gain $J_2(\boldsymbol{D},\boldsymbol{T})$ in (22) to

$$J_3(\boldsymbol{T}^{-1}\boldsymbol{M}\boldsymbol{T},\boldsymbol{T}) = \mathrm{tr}\,\Big[\boldsymbol{T}^T(\boldsymbol{W}_4 - \boldsymbol{M}^T\boldsymbol{W}_4\boldsymbol{M})\boldsymbol{T}\Big]. \tag{24}$$

The above derivation clearly indicates that joint optimization of EF and realization is not needed because the optimal $\boldsymbol{D}$ has already been determined by the closed-form expression in (23), hence it is sufficient to optimize matrix $\boldsymbol{T}$ by minimizing $J_3(\boldsymbol{T}^{-1}\boldsymbol{M}\boldsymbol{T},\boldsymbol{T})$ in (24).

The solution of minimizing $J_3(\boldsymbol{T}^{-1}\boldsymbol{M}\boldsymbol{T},\boldsymbol{T})$ in (24) turns out to be fairly sensitive to how the EF matrix $\boldsymbol{D}$ is quantized in the system implementation. As a remedy for this sensitivity issue, we consider a regularized minimization problem with the objective function given by

$$\begin{aligned} J_4(\boldsymbol{T}) &= \mu\,\mathrm{tr}[\boldsymbol{T}^T\boldsymbol{W}_4\boldsymbol{T}] + (1-\mu)\,\mathrm{tr}[\boldsymbol{T}^T(\boldsymbol{W}_4 - \boldsymbol{M}^T\boldsymbol{W}_4\boldsymbol{M})\boldsymbol{T}] \\ &= \mathrm{tr}[\boldsymbol{T}^T\{\boldsymbol{W}_4 - (1-\mu)\,\boldsymbol{M}^T\boldsymbol{W}_4\boldsymbol{M}\}\boldsymbol{T}] \end{aligned} \tag{25}$$

where $0 < \mu \le 1$ is a regularization parameter that can be selected to make an appropriate trade-off between minimizing (24) and reducing $\mathrm{tr}[\boldsymbol{T}^T\boldsymbol{W}_4\boldsymbol{T}]$.

To minimize (25) with respect to matrix $\boldsymbol{T}$ subject to the l_2-scaling constraints in (19), we define the Lagrange function

$$J_o(\boldsymbol{P},\lambda) = \mathrm{tr}[\boldsymbol{V}_\mu\boldsymbol{P}] + \lambda\,(\mathrm{tr}[\boldsymbol{K}\boldsymbol{P}^{-1}] - n) \tag{26}$$

where $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$ and $\boldsymbol{V}_\mu = \boldsymbol{W}_4 - (1-\mu)\,\boldsymbol{M}^T\boldsymbol{W}_4\boldsymbol{M}$. The optimal coordinate transformation matrix $\boldsymbol{T}$ can be determined by solving the equations

$$\begin{aligned} \frac{\partial J_o(\boldsymbol{P},\lambda)}{\partial \boldsymbol{P}} &= \boldsymbol{V}_\mu - \lambda\boldsymbol{P}^{-1}\boldsymbol{K}\boldsymbol{P}^{-1} = \boldsymbol{0} \\ \frac{\partial J_o(\boldsymbol{P},\lambda)}{\partial \lambda} &= \mathrm{tr}[\boldsymbol{K}\boldsymbol{P}^{-1}] - n = 0 \end{aligned} \tag{27}$$

which lead to

$$\boldsymbol{P}\boldsymbol{V}_\mu\boldsymbol{P} = \lambda\boldsymbol{K}, \qquad \mathrm{tr}[\boldsymbol{K}\boldsymbol{P}^{-1}] = n. \tag{28}$$

It follows that

$$\begin{aligned} \boldsymbol{P} &= \sqrt{\lambda}\,\boldsymbol{V}_\mu^{-\frac{1}{2}}\Big[\boldsymbol{V}_\mu^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{V}_\mu^{\frac{1}{2}}\Big]^{\frac{1}{2}}\boldsymbol{V}_\mu^{-\frac{1}{2}} \\ \frac{1}{\sqrt{\lambda}}\,\mathrm{tr}[\boldsymbol{K}\boldsymbol{V}_\mu]^{\frac{1}{2}} &= \frac{1}{\sqrt{\lambda}}\Big(\sum_{i=1}^{n}\theta_i\Big) = n \end{aligned} \tag{29}$$

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

$$\boldsymbol{P} = \frac{1}{n}\Big(\sum_{i=1}^{n}\theta_i\Big)\boldsymbol{V}_\mu^{-\frac{1}{2}}\Big[\boldsymbol{V}_\mu^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{V}_\mu^{\frac{1}{2}}\Big]^{\frac{1}{2}}\boldsymbol{V}_\mu^{-\frac{1}{2}}. \tag{30}$$

Substituting (30) into (26) yields the minimum value of $J_o(\boldsymbol{P},\lambda)$ as

$$\min_{\boldsymbol{P},\lambda} J_o(\boldsymbol{P},\lambda) = \frac{1}{n}\Big(\sum_{i=1}^{n}\theta_i\Big)^2. \tag{31}$$

Referring to (30), the optimal coordinate transformation matrix $\boldsymbol{T}$ that minimizes (26) can be obtained in closed form as

$$\boldsymbol{T} = \frac{1}{\sqrt{n}}\Big(\sum_{i=1}^{n}\theta_i\Big)^{\frac{1}{2}}\boldsymbol{V}_\mu^{-\frac{1}{2}}\Big[\boldsymbol{V}_\mu^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{V}_\mu^{\frac{1}{2}}\Big]^{\frac{1}{4}}\boldsymbol{U}_o. \tag{32}$$

where $\boldsymbol{U}_o$ is an arbitrary $n \times n$ orthogonal matrix. From (18) and (32), we obtain

$$\overline{\boldsymbol{K}} = n\Big(\sum_{i=1}^{n}\theta_i\Big)^{-1}\boldsymbol{U}_o^T\Big[\boldsymbol{V}_\mu^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{V}_\mu^{\frac{1}{2}}\Big]^{\frac{1}{2}}\boldsymbol{U}_o \tag{33}$$

Next, we choose the $n \times n$ orthogonal matrix $\boldsymbol{U}_o$ such that matrix $\overline{\boldsymbol{K}}$ in (33) satisfies the l_2-scaling constraints in (19). To this end, we perform the eigenvalue-eigenvector decomposition

$$\Big[\boldsymbol{V}_\mu^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{V}_\mu^{\frac{1}{2}}\Big]^{\frac{1}{2}} = \boldsymbol{Q}\,\mathrm{diag}\{\theta_1, \theta_2, \cdots, \theta_n\}\boldsymbol{Q}^T \tag{34}$$

where $\boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{I}_n$. As a result, we have

$$n\Big(\sum_{i=1}^{n}\theta_i\Big)^{-1}\Big[\boldsymbol{V}_\mu^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{V}_\mu^{\frac{1}{2}}\Big]^{\frac{1}{2}} = \boldsymbol{Q}\boldsymbol{\Lambda}^{-2}\boldsymbol{Q}^T \tag{35}$$

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

$$(\boldsymbol{Z}\boldsymbol{\Lambda}^{-2}\boldsymbol{Z}^T)_{ii} = 1 \quad \text{for} \quad i = 1, 2, \cdots, n. \tag{36}$$

can be obtained by numerical manipulations [7, p.278]. By choosing $\boldsymbol{U}_o = \boldsymbol{Q}\boldsymbol{Z}^T$ in (32), the optimal coordinate transformation matrix $\boldsymbol{T}$ satisfying (19) and (31) is constructed as

$$\boldsymbol{T} = \frac{1}{\sqrt{n}}\Big(\sum_{i=1}^{n}\theta_i\Big)^{\frac{1}{2}}\boldsymbol{V}_\mu^{-\frac{1}{2}}\Big[\boldsymbol{V}_\mu^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{V}_\mu^{\frac{1}{2}}\Big]^{\frac{1}{4}}\boldsymbol{Q}\boldsymbol{Z}^T. \tag{37}$$

Therefore, the optimal EF matrix $\boldsymbol{D}$ can easily be derived from (23) and (37).

4. A NUMERICAL 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 state 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. Under the circumstances, the coefficient vector of the full-order state observer and the gain vector of the state feedback were found to be

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

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

$$\boldsymbol{K} = \begin{bmatrix} 10.140736 & -9.324397 & 7.646117 \\ -9.324397 & 10.140736 & -9.324397 \\ 7.646117 & -9.324397 & 10.140736 \end{bmatrix}$$
$$\boldsymbol{W}_4 = 10 \begin{bmatrix} 0.946720 & 2.004689 & 1.053033 \\ 2.004689 & 4.432667 & 2.355278 \\ 1.053033 & 2.355278 & 1.389340 \end{bmatrix}.$$

By performing l_2-scaling to the system with

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

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

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

By setting $\mu = 0.01$ in (25), the optimal coordinate transformation matrix $\boldsymbol{T}$ and the optimal EF matrix $\boldsymbol{D}$ were computed from (37) and (23) as

$$\boldsymbol{T} = \begin{bmatrix} 3.956006 & -3.836672 & 2.965780 \\ -2.863405 & 0.513265 & -0.998894 \\ 0.404642 & 2.102568 & 0.813402 \end{bmatrix}$$
$$\boldsymbol{D} = \begin{bmatrix} -0.140440 & -0.394082 & 0.397784 \\ 1.071446 & -0.962399 & -0.025115 \\ 0.903949 & -0.229442 & -0.596136 \end{bmatrix}$$

respectively, with $J_4(\boldsymbol{T}) = 6.323423$. We observe that the normalized noise gain has been considerably reduced through the use of an optimal general EF matrix $\boldsymbol{D}$ and a separately optimized coordinate transformation matrix $\boldsymbol{T}$.

The other results regarding the normalized noise gain $J_2(\boldsymbol{D}, \boldsymbol{T})$ in (22) were summarized with those achieved by the analytic optimization method [12] in Table I where "Infinite Precision" shows the value of $J_2(\boldsymbol{D}, \boldsymbol{T})$ derived from the optimal $\boldsymbol{T}$ and $\boldsymbol{D}$, and "3-Bit Quantization (Integer Quantization)" means the value of $J_2(\boldsymbol{D}, \boldsymbol{T})$ 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).

TABLE I
PERFORMANCE COMPARISON FOR GENERAL MATRIX EF

Methods	Infinite Precision	3-Bit Quantization	Integer Quantization
Analytic [12]	11.635118	11.808013	15.337683
Proposed Analytic	5.541528	5.749012	12.711196

5. CONCLUSION

The optimization problem of EF and realization to minimize the roundoff noise in a closed-loop system with full-order state observer feedback subject to l_2-scaling constraints has been investigated. It has been shown that joint optimization of EF and realization is not needed when utilizing a general-matrix EF loop, because the optimal EF matrix, in closed-form, can be determined in advance. An efficient analytic technique has been developed for obtaining the optimal realization that minimizes the roundoff noise. A case study has demonstrated the validity and effectiveness of the proposed technique.

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, A. Doi and W.-S. Lu, "Roundoff noise minimization in state-space discrete-time systems using joint optimization of high-order error feedback and realization," *IEEE Trans. Circuits Syst. I*, vol. 61, pp. 3460-3468, Dec. 2014.